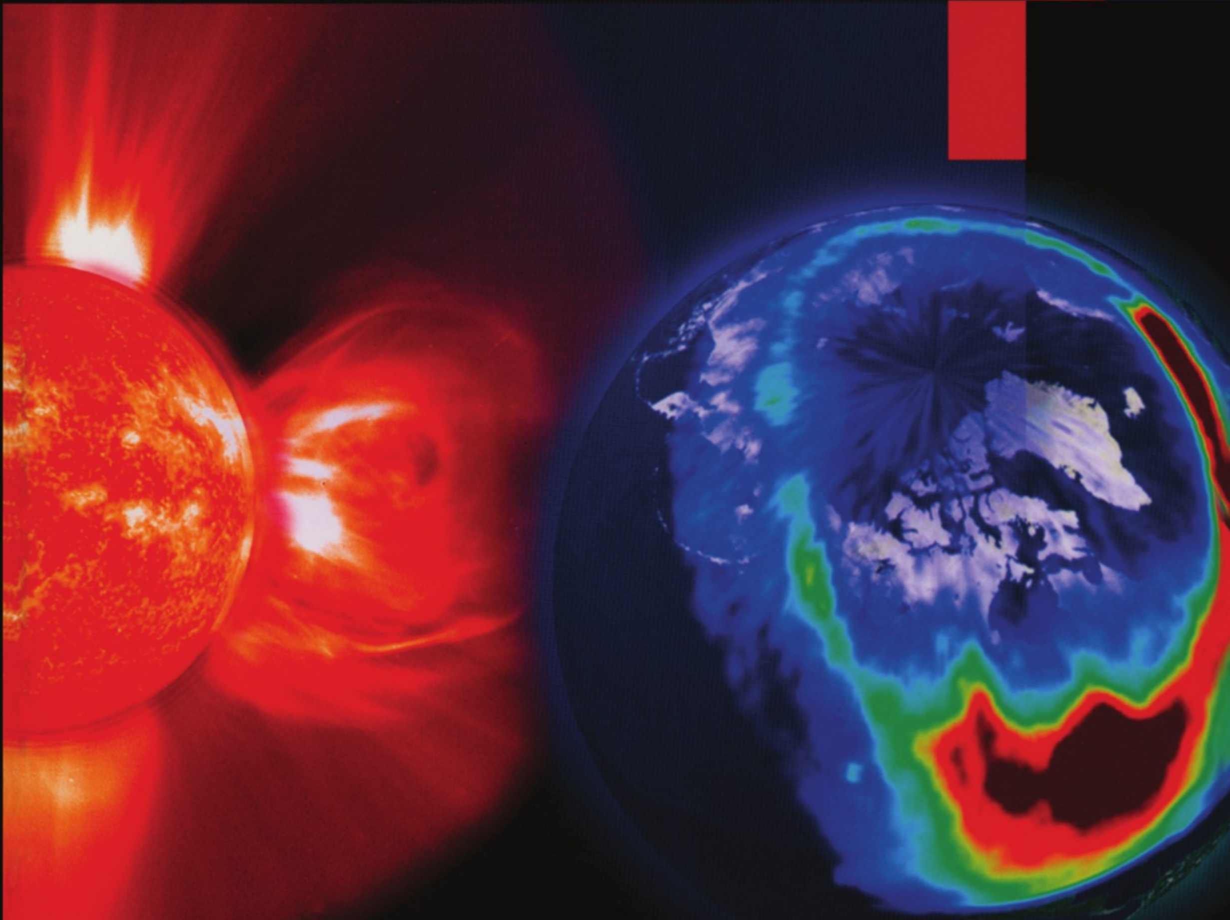
SPACE WEATHER

Physics and Effects

Volker Bothmer and Ioannis A. Daglis



Springer

PRAXIS

Space Weather – Physics and Effects

Volker Bothmer and Ioannis A. Daglis

Space Weather – Physics and Effects

Springer

Published in association with
Praxis Publishing
Chichester, UK

PRAXIS

Dr Volker Bothmer
Institute for Astrophysics
University of Göttingen
Göttingen
Germany

Dr Ioannis A. Daglis
National Observatory of Athens
Athens
Greece

SPRINGER–PRAXIS BOOKS IN ENVIRONMENTAL SCIENCES
SUBJECT *ADVISORY EDITOR*: John Mason B.Sc., M.Sc., Ph.D.

ISBN 10: 3-540-23907-3 Springer-Verlag Berlin Heidelberg New York
ISBN 13: 978-3-540-23907-9 Springer-Verlag Berlin Heidelberg New York

Springer is part of Springer-Science + Business Media (springer.com)

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available from the Internet at http://dnb.ddb.de

Library of Congress Control Number: 2006921904

Printed in Germany

Cover design: Jim Wilkie
Project management: Originator Publishing Services, Gt Yarmouth, Norfolk, UK

Printed on acid-free paper

Contents

Preface

The compilation of this book on the Physics of Space Weather was stimulated by the exciting new results of the joint ESA/NASA SoHO mission. SoHO has certainly set a breathtaking milestone in terms of its unprecedented remote sensing and *in situ* observations, ranging from the Sun's interior to the outer solar atmosphere and heliosphere, including geospace. It helped scientists to view the Sun–Earth connections with new eyes and to prove or reject existing theoretical models and concepts.

Paal Brekke as the SoHO Deputy Project Manager of the Project Scientist Bernhard Fleck at the NASA Goddard Space Flight Center in Greenbelt, Maryland, was contacted by Clive Horwood from Praxis Publishing Ltd, but being so busy with the new scientific achievements from SoHO he asked me to take over this challenging project – both of us sharing the fascination to explore the physics of the Sun–Earth connections. It then took a while to outline the content of the book because of the strong interdisciplinary science that comprises the subject of Space Weather. This was addressed using a team of experts. Ioannis (Yannis) Daglis, with his outstanding contributions to the field of Space Weather, appeared to me as the appropriate European colleague, at the forefront of knowledge, who could help to fulfill the scope of the project successfully. I would like to express my great pleasure to Ioannis for his acceptance to coordinate this project with me.

During the following months it turned out, without surprise, that the book would attain a strong international character, with leading experts from both Europe and the United States making contributions. However, it also turned out that despite the great excitement expressed by all the authors for the project, they all were extremely busy in their daily tasks. The endeavour to compile the material from the various fields of research, to present a coherent book as comprehensive as possible, claimed a remarkable amount of time. This would not have been possible without the unbelievable patience and great support of Clive Horwood and his team – Neil Shuttlewood and Jim Wilkie.

I want to acknowledge that this book would not have become a reality without the financial funding of my research project Stereo/Corona through the German Space Agency DLR (Deutsches Zentrum für Luft- und Raumfahrt) as a science and hardware contribution to the SECCHI optical imaging suite for the NASA STEREO mission, through funding for two joint EU-ESA/INTAS projects and a research collaboration with the SECCHI principal investigator Dr. Russell Howard at the US Naval Research Laboratory, Washington, D.C., the excitement of Thomas Kraupe as director of the Planetarium Hamburg and my colleague Wolfgang Keil from EADS/Astrium at Friedrichshafen. I am grateful to Manfred Siebert, Rainer Schwenn and Kristian Schlegel who introduced me to the field of solar terrestrial physica. Finally, I would like to express my gratitude to my scientific colleagues in the scientific consortia of the Yohkoh, Ulysses, SoHO, TRACE, and STEREO missions for their scientific support and all international colleagues who helped supply important results and materials. Last, but not least, I am grateful for the overall excitement of the members of our scientific community, my PhD and diploma students at the University of Göttingen and the outstanding patience of my family. Thank you all very much!

Volker Bothmer
Göttingen, July 2006

From the editors to their parents and to their families:
Gudrun, Hannes, Tobias and Anna,
Alexandros, Thanasis, Dimitris

Figures

Tables

Abbreviations and acronyms

ACE	Advanced Composition Explorer
ACR	Anomalous Cosmic Ray
AE	Auroral Electrojet (geomagnetic index)
AIR	Atmospheric Ionizing Radiation
AMIE	Assimilative Mapping of Ionospheric Electrodynamics
AMPTE	Active Magnetospheric Particle Tracer Explorers
AR-8	Aerospace Electron Radiation Belt Model 8
ARMA	AutoRegressive Moving Average (model)
AU	Astronomical Unit
BDC	Brewer–Dobson Circulation
BFO	Blood-Forming Organs
CAM	Computerized Anatomical Man
CCE	Charge Composition Explorer (part of Active Magnetospheric Particle Tracer Explorers)
CERN	*Centre Européenne pour la Recherche Nucléaire*
CH	Coronal Hole
CHAMP	CHAllenging Microsatellite Project
CID	Cold Ion Detector
CIM	Complex Image Method
CIR	Co-rotating Interaction Regions
CISM	Integrated Space Weather Modeling
CME	Coronal Mass Ejection
COSTEP	COmprehensive SupraThermal and Energetic Particle analyzer
COTS	Commercial-Off-The-Shelf
CRAND	Cosmic Ray Albedo Neutron Decay
CRCM	Comprehensive Ring Current Model
CRRES	Combined Release and Radiation Effects Satellite

DMSP	Defense Meteorological Satellite Program
Dst	Disturbance storm-time index
ECMWF	European Centre for Medium Range Weather Forecasting
EETES	EGNOS End-To-End Simulator
EGNOS	European Geostationary Navigation Overlay System
EISCAT	European Incoherent SCATter Association
EIT	Exact Image Theory
EKF	Extended Kalman Filter
EM	ElectroMagnetic
EPS	Energetic Particle Sensor
ESA	European Space Agency
ESP	Energetic Storm Particle
EUV	extreme ultraviolet
FAC	Field-Aligned Current
FH	Front-side Halo
FIP	First Ionization Potential
FIR	Finite Impulse Response (model)
FL	Flight Level
FMI	Finnish Meteorological Institute
FOV	Field Of View
GAGAN	GPS-Aided Geo Augmented Navigation
GAIM	Global Assimilation of Ionospheric Measurements
GCM	Ceneral Circulation Model
GCR	Galactic Cosmic Ray
GCS	Graduated Cylindrical Shell
GEOTAIL	Satellite
GIC	Geomagnetically Induced Current
GISM	Global Ionospheric Scintillation Model
GLE	Ground Level Enhancement
GLONASS	GLObal Navigation Satellite System
GOES	Geostationary Operational Environmental Satellites
GONG	Global Oscillation Network Group
GPS	Global Positioning System
GPS/MET	METeorological application of GPS
GSEQ	Geocentric Solar EQuatorial
GSM	Geocentric Solar Magnetospheric Coordinates system
GZK	Greisen–Zatsepin–Kuz'min
HC	Hadley Circulation
HEPAD	High-Energy Proton and Alpha Detector
HILDCAA	High-Intensity Long-Duration Continuous AE Activity
HST	Hubble Space Telescope
IBC	International Brightness Coefficient
ICME	Interplanetary Coronal Mass Ejection
ICRP	International Commission on Radiological Protection
IGY	International Geophysical Year

IMF	Interplanetary Magnetic Field
IRE	Ionospheric Range Error
IRF	Impulse Response Function
ISES	International Space Environment Service
ISS	International Space Station
Kp	*Kennziffer Planetarisch*
KPNSO	Kitt Peak National Observatory
L	Magnetic particle drift parameter
LAAS	Local Area Augmentation System
LASCO	Large Angle Spectrometric COronagraph
LDEF	Long Duration Exposure Facility
LEO	Low Earth orbit
LET	Linear Energy Transfer
LNT	Linear No Threshold
MA	Moving Average (model)
MAS	Magnetohydrodynamics Around a Sphere
MDI	Michelson Doppler Imager
MHD	MagnetoHydroDynamic
MHD	MagnetoHydroDynamics
MMC	Mean Meridional Circulation
MOS	Model Output Statistics
MSAS	MT-Sat Augmentation System
MSFM	Magnetospheric Specification and Forecast Model
NASA	National Aeronautics and Space Administration
NCEP/NCAR	U.S. National Center for Environmental Prediction/ National Center for Atmospheric Research
NCRP	National Council on Radiation Protection and Measurements
NN	Neural Network
NSWP	US National Space Weather Program
OLR	Outgoing Longwave Radiation
OSO	Orbiting Solar Observatory
OWL	Orbiting Wide-angle Light-collectors
PCA	Polar Cap Absorption event
PFSS	Potential Field Source Surface
Polar	NASA satellite; part of the international Solar Terrestrial Physics Program
QBO	Quasi-Biennial Oscillation
RBE	Relative Biological Effectiveness
RCRU	Radio Communications Research Unit
REMSIM	Radiation Exposure and Mission Strategies for Interplanetary Manned Missions (ESA project)
RTK	Real-Time Kinematic
RWC	Regional Warning Center
SA	Selective Availability; Shock-Associated

SAC-C	*Satélite de Aplicaciones Cientifícas*-*C*
SAE	Super GIC-inducing Auroral Electrojet
SAIC	Science International Corporation
SAMPLEX	Solar, Anomalous, and Magnetispheric ParticLE Explorer
SBAS	Satellite-Based Augmentation System
SCIP	Sun-Centered Imaging Package
SEC	Space Environment Center
SEE	Single Event Effects
SEM	Space Environment Monitor
SEP	Solar Energetic Particle
SESAMe	SECCHI Experiment Sun Aperture Mechanism
SEU	Single-Event Upset
SFHCME	Superfast Front-side Halo CME
SFU	Solar Flux Unit
SMM	Solar Maximum Missions
SO	Southern Oscillation
SoL	Safety-of-Life
SPE	Solar Proton Event
SSC	SunSpot Cycle
STEREO	Solar TErrestrial RElations Observatory
STRV-1a	Space Technology Research Vehicle
SXT	Soft X-ray Telescope
Sym-H	Symmetric component of horizontal magnetic field vector near the Earth's surface
TEC	Total Electron Content
TEPC	Tissue Equivalent Proportional Counter
TRACE	Transition Region And Coronal Explorer
UHECR	UltraHigh-Energy Cosmic Rays
ULF	Ultra Low Frequency
UV	UltraViolet
VLF	Very Low Frequency
WSA	Wang–Sheeley–Arge (model)
WTS	Westward Travelling Surge

About the authors

Dr. Volker Bothmer
Project Lead Stereo/Corona, Institute for Astrophysics, University of Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany
email: bothmer@astro.physik.uni-goettingen.de

Volker Bothmer uses modern spacecraft observations for his research in solar and heliospheric physics. He was involved in the planning and payload design of missions like STEREO and Solar Probe where he also was a member of NASA's Science Definition Teams. He currently leads the German project Stereo/Corona, a science and hardware contribution to STEREO's optical imaging package SECCHI, at the University of Göttingen and is a co-investigator for the *in situ* instrument suite IMPACT. He is leader of the EU-ESA/INTAS space weather project 03-51-6206. His research focuses on coronal mass ejections and their effects on geospace. His university lectures include solar physics, ground- and space-instrumentation, and data analysis techniques. Amongst his more than 100 publications as author and co-author there are science firsts, invited reviews, public outreach articles, and instrumentation and mission proposals. He has organized international symposia and workshops and convened and chaired various conference sessions at major international meetings. He currently serves as secretary of the EGU, Co-chair of COSPAR sub-commission D2/E3 as member of ESA's SWEN team and of the IAU.

Dr. Ioannis A. Daglis
Director of the Institute for Space Applications & Remote Sensing, National Observatory of Athens, Metaxa and Vas. Pavlou St., Penteli 15236 Athens, Greece
email: daglis@space.noa.gr

Ioannis (Yannis) A. Daglis is a space physicist who has been active mainly in the field of solar–terrestrial physics. His research has focused on magnetospheric substorms, geospace magnetic storms, and magnetosphere–ionosphere coupling. He has published more than 60 scientific papers and more than 100 public outreach articles, and has edited 2 monographs – *Space Storms and Space Weather Hazards* and *Effects of Space Weather on Technology Infrastructure* – and 3 special journal issues on space weather. He has organized several international conferences and symposia on solar–terrestrial physics and has given invited review talks at more than 20 conferences. He is currently serving as the Editor for Magnetospheric Physics of Annales Geophysicae, as member of the ESA Solar System Working Group, and as national delegate of Greece at the ESA/EC High-level Space Policy Group.

List of contributors

Chapter 1
Dr. Volker Bothmer and Dr. Ioannis A. Daglis (see About the authors for address details)

Chapter 2
Dr. George Siscoe, 68 Dutton Rd, Sudbury, MA 01776, USA
email: siscoe@bu.edu

Chapter 3
Dr. Volker Bothmer (see About the authors for address details)

Dr. Andrei Zhukov, Royal Observatory of Belgium, Avenue Circulaire 3, B-1180 Brussels, Belgium
email: andreiz@oma.be

Chapter 4
Prof. C. T. Russell, Institute of Geophysics and Planetary Physics, University of California Los Angeles, 3845 Slicter Hall, Los Angeles, CA 90095-1567, USA
email: ctrussell@igpp.ucla.edu

Chapter 5
Dr. Norma B. Crosby, Belgian Institute for Space Aeronomy, Ringlaan-3-Avenue Circulaire, B-1180 Brussels, Belgium
email: norma.crosby@oma.be

Chapter 6

Prof. Daniel N. Baker, Director, Laboratory for Atmospheric and Space Physics, Professor, Astrophysical and Planetary Sciences, Campus Box 590, University of Colorado, Boulder, CO 80309, USA
email: daniel.baker@lasp.colorado.edu

Dr. Ioannis A. Daglis (see About the authors for address details)

Chapter 7

Prof. Kristian Schlegel, Kapellenweg 24, 37191 Katlenburg-Lindau, Germany
email: schlegel@linmpi.mpg.de

Chapter 8

Prof. Dr. Karin Labitzke, Meteorologisches Institut der Freien Universität Berlin, Carl-Heinrich-Becker-Weg 6-10, D-12165 Berlin, Germany
email: karin.labitzke@mail.met.fu-berlin.de

Prof. Harry van Loon, National Center for Atmospheric Research, CGD/NCAR, P.O. Box 3000, Boulder, CO 80307, USA
email: vanloon@ucar.edu

Chapter 9

Dr. Louis J. Lanzerotti, Center for Solar–Terrestrial Research, Department of Physics, Tiernan 101, New Jersey Institute of Technology, Newark, New Jersey 07102, USA
email: ljl@adm.njit.edu

Chapter 10

Dr. Risto Pirjola, Finnish Meteorological Institute, Space Research Unit, P.O. Box 503, FIN-00101 Helsinki, Finland
email: risto.pirjola@fmi.fi

Chapter 11

Dr. Günther Reitz, Head of Radiation Biology Department, German Aerospace Center, Aerospace Medicine, Linder Höhe, 51147 Köln, Germany
email: guenther.reitz@dlr.de

Dr. Rainer Facius, German Aerospace Center, Institute of Aerospace Medicine, Linder Höhe, 51147 Köln, Germany
email: rainer.facius@dlr.de

Chapter 12

Dr. Eamon Daly (Dr. Alexi Glover), TEC-EES/ Rm Dk214, ESTEC, Keplerlaan 1, P.O. Box 299, 2200AG Noordwijk, The Netherlands
email: eamonn.daly@esa.int

Dr. Alain Hilgers, ESA-EUI-SI, 8–10 rue Mario Nikis, 75015 Paris, France
e-mail: alain.hilgers@esa.int

Chapter 13

Dr. Bertram Arbesser-Rastburg, ESA-ESTEC, EM&SE Division, TEC-EEP, Keplerlaan 1, PB 299, NL-2200 AG Noordwijk, The Netherlands
email: bertram@tec-ee.esa.int

Dr. Norbert Jakowski, DLR, Institute of Communications and Navigation, Kalkhorstweg 53, D-17235 Neustrelitz, Germany
email: norbert.jakowski@dlr.de

Chapter 14

Dr. Dimitris Vassiliadis, ST at NASA/Goddard Space Flight Center, Building 21, Room 265B, Mailstop 612.2, Greenbelt, MD 20771, USA
email: vassi@electra.gsfc.nasa.gov

1

Introduction

Volker Bothmer and Ioannis A. Daglis

The term ‘space weather’ has achieved great international scientific and public importance today. Space weather can be understood as a field of research that will provide new insights into the complex influences and effects of the Sun and other cosmic sources on interplanetary space, the Earth’s magnetosphere, ionosphere, and thermosphere, on space- and ground-based technological systems, and beyond that, on their endangering affects to life and health. Space weather is hence a highly relevant subject of research for our modern society. It has evolved from indications of solar effects on interplanetary space and Earth since the middle of the 18th century in a dramatic fashion, based on the new scientific results that have been obtained through the successful development and operation of space missions (e.g., AMPTE, Yohkoh, CRRES, SoHO, Ulysses, TRACE, WIND, ACE). The sophisticated instruments flown on spacecraft have provided us with unprecedented views of the Sun’s surface and outer atmosphere, out to distances far beyond our home planet.

At the same time, various articles in scientific journals, magazines, and newspapers reported on the many effects of solar activity on Earth and in space, such as satellite damage, radiation hazards to astronauts and airline passengers, telecommunication problems, outages of power and electronic systems, and even the relevance of solar and cosmic influences for the evolution of the Earth’s climate. Nowadays the media, students, teachers, and the general public are eager to learn about the state-of-the-art knowledge in the field of space weather. In particular, there is an interest in the achievement of the first real-time space weather forecasts similar to the daily weather forecasts. About 20 years ago we would probably not have dared to dream of having made such great steps forward in our understanding of the physics of the Sun–Earth system, which will certainly continue at a fast pace in the next decades. At the same time our modern technologies and space developments will continue to be vulnerable to the effects of space weather. This is especially relevant for micro-satellite technologies, future manned missions to the Moon and

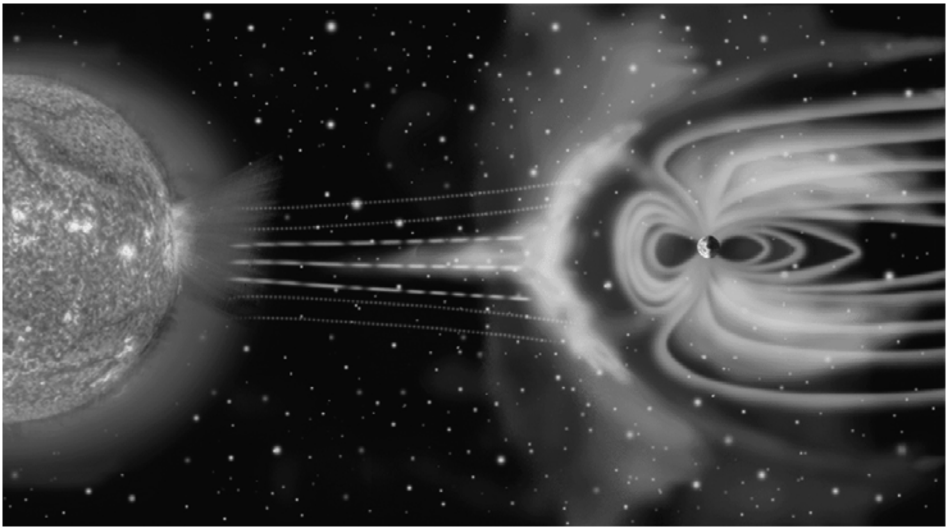


Figure 1.1. Schematic view of the complex Sun–Earth system (courtesy: NASA/JPL/Caltech).

Mars, GPS navigation systems under development, such as GALILEO, communication satellites, and electronic systems operating in near-Earth space.

The Sun–Earth system – schematically shown in Figure 1.1 – is a very complex system with a large variety of physical processes, ranging from magnetic field reconnection and plasma acceleration processes, to impacts of charged particles on electronic and biological systems. Furthermore, these processes cover multiple spatial and time scales, requiring interdisciplinary scientific collaborations at the highest levels of excellence.

This book aims, in an interdisciplinary manner, to provide insight into the different major fields of research in space weather. The editors started this project with the ambition to produce a state-of-the-art compendium on the importance and understanding of the many different aspects and effects of space weather. The individual chapters of this book have been written by scientists at the forefront of research in various fields. Each chapter is meant to be understandable relatively independently from the other chapters. In this sense we did not remove redundancies of topics amongst the individual chapters. We aimed to provide each author as much flexibility as possible and believe the resulting individual chapters have benefited from this approach.

Following this introduction, the Chapter 2 starts with an outline of the similarities and differences between the fields of meteorology and space weather viewed in a historical perspective by George Siscoe. Our new view of the dynamic solar atmosphere, as the prime source of space weather, is presented by Volker Bothmer and Andrei Zhukov in Chapter 3 together with an introduction into the interplanetary consequences and outline of the major parameters that play a key role for triggering space weather hazards. In Chapter 4 Chris Russell reviews how the continuous

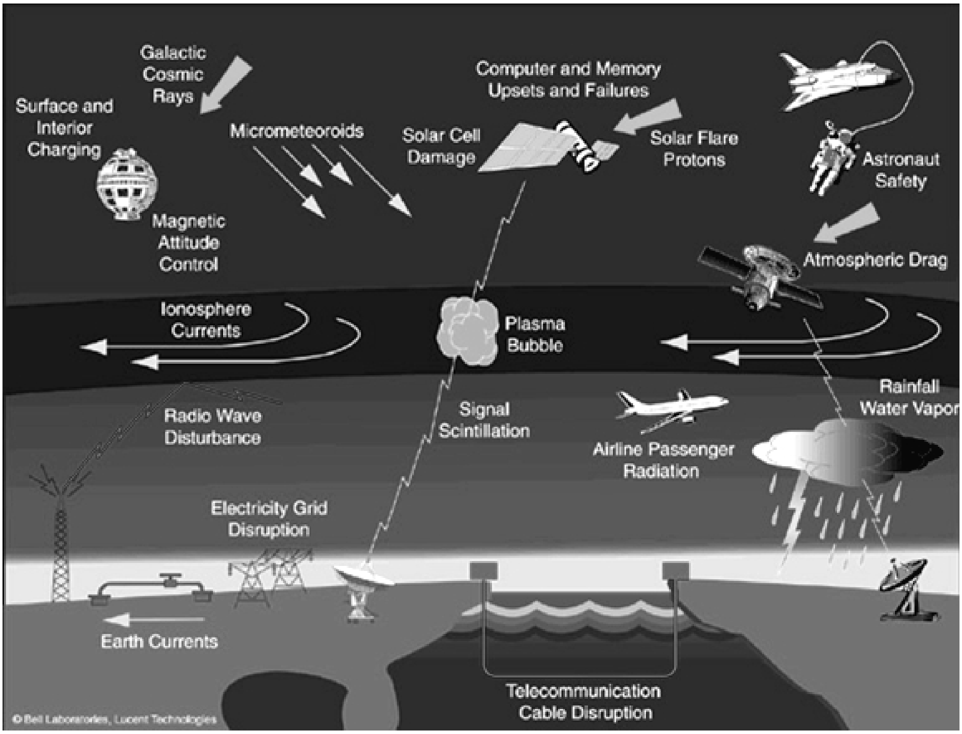


Figure 1.2. Summary of the known space weather effects (from Lanzerotti, 1997).

stream of supersonic charged particles, the solar wind, interacts with the Earth's magnetic field and summarizes the complex physical processes inside the Earth's magnetosphere. The magnetosphere is one of the many particle environments in our Solar System, with a variety of particle sources. An overview on the different particle environments in the heliosphere is provided by Norma Crosby in Chapter 5, together with an outline of their implications in terms of radiation hazards during future interplanetary travel to the Moon or Mars. The radiation environment, particles, and electric currents in the inner magnetosphere itself, which impact near-Earth orbiting and geo-synchronous satellites, are introduced by Daniel Baker and Ioannis Daglis in Chapter 6. Precipitation of accelerated charged particles from the magnetosphere into the ionosphere during geomagnetic storms causes changes in the electric current systems and aurorae. The physical background of these processes, together with the effects of cosmic rays and solar flares, are reviewed by Kristian Schlegel in Chapter 7. The impact of increased EUV radiation on the upper stratosphere leading to ozone increase and subsequent warming is explained by Karin Labitzke and Harry van Loon in Chapter 8. In Chapter 9 Lou Lanzerotti presents an overview of how space weather affects telecommunication systems operating on the ground and in space, beginning with the earliest electric telegraph systems and continuing to today's wireless communications using satellite and land links. On the Earth's surface space weather effects occur in the form of induced electric currents which cause technical problems to power grids, oil pipelines, or telecommunication equipment as Risto Pirjola explains in Chapter 10. How space weather impacts on space radiation protection, a subject of prime importance to avoiding life threatening doses to humans in space or air crew passengers, is explained in depth by Rainer Facius and Günther Reitz in Chapter 11. A description of the causes and consequences of the variability of the space environment affecting spacecraft operation and hardware, viewed in terms of the operational side of space weather, is summarized by Alain Hilgers, Alexi Glover, and Eamonn Daly in Chapter 12. The physical proceses that cause disturbances to global navigation satellite systems are introduced by Bertram Arbesser-Rastburg and Norbert Jakowski in Chapter 13. The book closes with an outline by Dimitris Vassiliadis (Chapter 14) on the methods and models that help space weather centers to establish realistic forecasts.

At a time marked by strong world-wide initiatives in support of new satellite missions and planning to help establish the required necessary infrastructure for space weather research and services, as for example in form of space weather competence centers, which will form the backbone of forecast systems in the near future, we hope that this book will provide an invaluable resource on the physics of space weather for students, teachers, and researchers.

2

Space weather forecasting historically viewed through the lens of meteorology

George Siscoe

The history of progress in the effectiveness of meteorological forecasting can be divided into ten stages: (1) recognition of societal need; (2) development of rules for forecasts based on visual observations; (3) quantification of storm parameters through instrument observations; (4) development of retrospective synoptic weather maps; (5) institution of forecast centers after the technological means of forecasting (the telegraph and instrument-based weather maps) came into being; (6) development of models of storm structure; (7) subjective analysis based on weather chart analysis; (8) objective analysis based on empirical formulas; (9) numerical predictions based on integrating the equations of atmospheric motion; and (10) storm tracking by radar and satellites. A parallel division of the history of space weather forecasting is here recounted. Whereas the effectiveness of meteorological forecasting dramatically increased with the advent of the numerical forecasting (stage 9), space weather forecasting is presently making progress through massively expanding its repertoire of objective forecast algorithms (stage 8). The advent of physics-based numerical space weather predictions (the stage of dramatic improvement in forecast effectiveness in meteorology, stage 9) is still in the future for space weather, although codes to achieve such predictions are under development. The crucial role that teaching forecasting in core meteorology courses has played in producing researchers motivated to improve forecasting effectiveness (and its absence in space weather curricula) is emphasized.

2.1 SIBLING SCIENCES

The fields of meteorology and space weather, as producers of forecasting services, are siblings, but the former is much older. Thinking that experience of the older might guide the younger, we survey milestones in the development of forecast meteorology where analogues to space weather forecasting exist.

The exercise should work to a certain extent because the basic dynamics of both enterprises is continuum mechanics (magnetohydrodynamics, MHD, in one case)

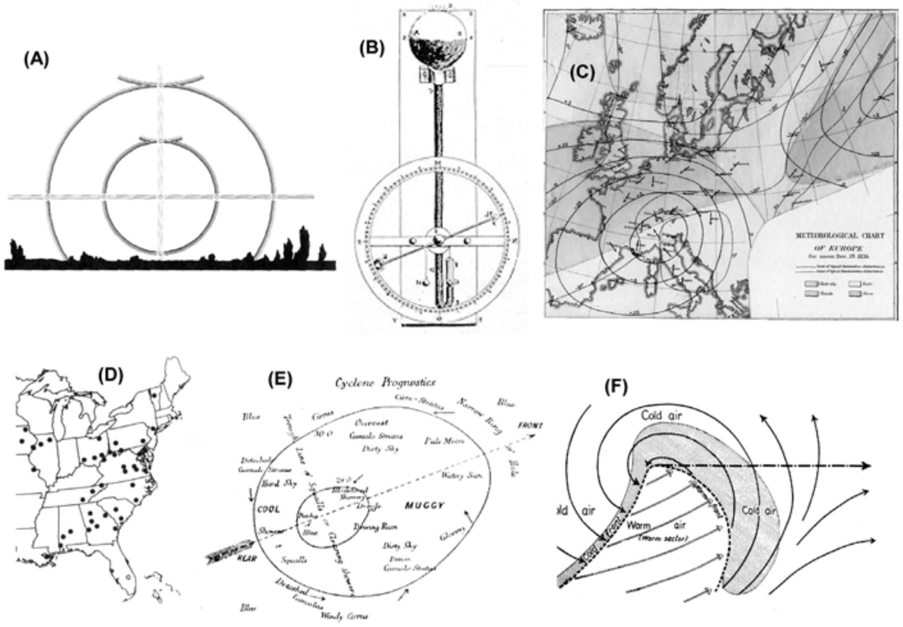


Figure 2.1. Illustrating six stages in the development of weather forecasting from prehistory to mid-20th century. (A) Weather signs illustrated by a double halo with tangent arcs and parhelia (from Menzel, 1953, p. 190); (B) instrument-based observations represented by the wheel barometer of Robert Hooke as pictured in 1665 in *Micrographia* (reprinted 1987); (C) synoptic imaging of storms and their motions exemplified by the European storm of December 1836 as analyzed by Elias Loomis (Loomis, 1859); (D) network of US telegraph stations in 1860 sending weather reports to the Smithsonian in Washington DC (Fleming, 1990, p. 145); (E) forecasts based on the internal structure of cyclonic depressions as depicted in an 1885 chart by Abercromby (from Shaw, 1911, p. 87); (F) forecasts based on the cyclone seen as a moving, breaking wave along the polar front (Bjerknes and Solberg, 1923).

applied to gaseous media subject to instabilities that produce the disturbances we recognize as weather. Nonetheless, analogies between the two fields run deeper than a shared paradigm might suggest; that is, in both, travelling storm systems – extratropical cyclones and corotating interaction regions, hurricanes, and coronal mass ejections – dominate prediction efforts.

That said, we cannot push the analogy too far, as algorithms to forecast snow in Duluth or Stockholm, for instance, differ in kind from those to forecast local ground induced currents. We should therefore look for commonalities not at the local level but at the level of the discipline as a whole. What are the discipline-wide stages with parallels to space weather through which forecast meteorology has progressed?

2.2 STEPS IN THE ADVANCE OF ENVIRONMENTAL FORECASTING: THE METEOROLOGICAL EXPERIENCE

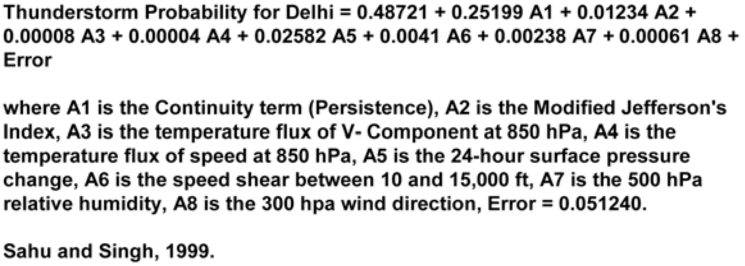
Visual observations came first. Solar halos and cirrus anvils told sailors and farmers to brace for storm (Figure 2.1A). Reading sky signs was already ancient lore by

300 BC when Theophrastus – a student of Aristotle – codified the practice in *Concerning Weather Signs*. Here, for example, we read: 'If two mock-suns appear, one to the south, the other to the north, and there is at the same time a halo, these indicate that it will shortly rain' (p. 405). Reading sky signs remained central to good weather-telling well into the age of instrument-based forecasts, as Robert FitzRoy, who pioneered the posting of storm warnings, here insists: 'To know the state of the atmosphere, not only barometer and thermometers should be watched, but *appearances* of the sky should be vigilantly noticed, invariably' (FitzRoy, 1863, p. 15). The distance into the future of forecasts based on sky signs is set by the curvature of the Earth, the height of the visual phenomena, and the speed of storms, and is typically half a day.

After visual observations came instrument-based observations, beginning in the seventeenth century with the inventions of the thermometer, barometer, and hygrometer (Figure 2.1B). Instrument-based observations at a single location increased the range of weather forecasting little beyond that achieved by visual observations, but by combining contemporaneous observations from many locations it was possible to construct synoptic weather charts that show, by means of iso-contours of temperature and pressure, a picture of a storm system and its movement. For synoptic charts to be most useful a network of observers was required, wide enough to encompass the storm-affected area. Such networks, comprising military forts, universities, and amateurs, gradually grew (through negotiation and solicitation by visionaries) sufficient enough to enable Elias Loomis – one of the first great synopticians – to publish synoptic charts of two 1836 storms – one in North America based on barometric data from 32 observers, and one in Eurasia based on barometric data from 47 observers (Figure 2.1C). The first weather charts were used to document the structure and dynamics of storm systems, which by the 1850s led to the realization that mid-latitude storms are big – typically more than 1,500 km across – and as a rule travel from west to east.

Realization that storms are big and move systematically led to the next major advance in weather forecasting: the idea of using the telegraph – which though barely a decade old already laced America and Eurasia with more than 100,000 miles of wire – to warn of approaching storms (Figure 2.1D). Between 1860 and 1870 the idea was implemented in various countries, and government weather bureaus were created or co-opted to administer the operation.

There then followed a period of maturing of the science of forecasting, during which a storm pictured as a moving, amorphous mass of wind and precipitation was instead increasingly pictured as having distinguishable parts in fixed relation to the storm as a whole. In 1885 the British Meteorological Office published a set of forecast rules based largely on the internal structure of cyclonic depressions as then understood (Figure 2.1E). Such templates depicting a cyclone as a static distribution of clouds and weather around a moving depression of varying intensity served the weather forecaster well into the twentieth century. In the 1920s the static picture began to be replaced by one in which the cyclone evolved as it moved. It was seen as an instability that taps and dissipates free energy stored in the form of a warm air mass to the south pressing against a cold air mass to the north along an



Thunderstorm Probability for Delhi = 0.48721 + 0.25199 A1 + 0.01234 A2 + 0.00008 A3 + 0.00004 A4 + 0.02582 A5 + 0.0041 A6 + 0.00238 A7 + 0.00061 A8 + Error

where A1 is the Continuity term (Persistence), A2 is the Modified Jefferson's Index, A3 is the temperature flux of V- Component at 850 hPa, A4 is the temperature flux of speed at 850 hPa, A5 is the 24-hour surface pressure change, A6 is the speed shear between 10 and 15,000 ft, A7 is the 500 hPa relative humidity, A8 is the 300 hpa wind direction, Error = 0.051240.

Sahu and Singh, 1999.

Figure 2.2. Example of a MOS forecast algorithm for a single-station giving, in this case, the probability of thunderstorms in the region of Delhi, India. Various input parameters, A3, A4, etc., are obtained from local radiosonde measurements.

east–west line called the polar front (Figure 2.1F). The polar front theory of extratropical cyclones, with its choreographed waltz of pivoting cold and warm fronts gliding eastward, dominated weather forecasting until after the mid-twentieth century. But it was still basically a qualitative or subjective method of analysis with no real potential for improvement. In 1954, meteorologists Ludlam and Scorer could complain that 'during the last twenty or thirty years there has been no noticeable increase in the accuracy of general forecasts for a day ahead' (p. 102). Matters began to improve around mid-century when quantitative methods of analysis were introduced. These took two forms: objective analysis and numerical predictions.

Objective analysis applies statistical techniques such as multivariate regression analysis to variables that can be measured at a single station to construct formulae that produce forecast probabilities of various weather elements such as temperature or precipitation (Figure 2.2). In effect, because of its localization to a single station this is a modern version of weather-sign forecasting. With the advent of numerical models (touched on next) objective analysis formulae have been based on statistics generated from model outputs, referred to as model output statistics (MOS).

In the 1950s, weather forecasting was revolutionized by the introduction of operational numerical weather prediction codes (Cressman, 1996). These codes and their research counterparts, general circulation models (GCMs), are based on the hydrodynamics and thermodynamics that govern the global motion of the atmosphere and on the physics of clouds and precipitation. It was, of course, the advent of powerful computers that enabled the revolution that transformed weather forecasting in the 1950s. Once forecasting became based on codes that integrate the applicable dynamical equations of motion, forecast range and accuracy could steadily improve because the codes are numerical simulations of nature itself. Therefore, improving the simulations necessarily increases forecast skill, and the simulations can be improved *in perpetuum* through the always-increasing number-crunching power of computers, better coding of the equations of motion, better parametric representation of sub-grid-scale process, and better techniques for reinitializing the code periodically with weighted combinations of observations and predicted values (data assimilation). The impact of numerical weather prediction on weather forecasting cannot be overstated, as illustrated in Figure 2.3, which shows the increase in

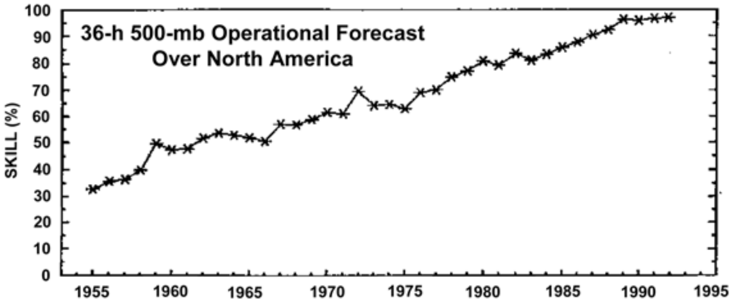


Figure 2.3. Forecast skill from 1955 to 1992 in the 36-hour prediction of the height of the 500-mb surface over the United States (from McPherson, 1994).

forecast skill in the 36-hour prediction of the height of the 500-mb surface over the United States. It shows that when numerical weather prediction was introduced into the weather service in 1955, the skill as quantified here was 33%, which matched that of subjective forecasts. But whereas the skill of subjective forecasts remained basically unchanged over the ensuing 37 years covered by the chart, numerical forecast skill increased to 98%.

The most recent major advance in weather information service is weather tracking by satellite imagery and radar. These are of great value in tracking hurricanes and the most damaging manifestations, such as tornadoes, of severe mid-latitude storms.

The final transformational steps in forecast meteorology from sky signs to satellite tracking were facilitated by a critical supplementary step: the teaching of weather forecasting in universities. University forecasting courses expanded greatly during the Second World War, to supply forecasters for the war effort, but they have continued to the present (Koelsch, 1996; Houghton, 1996). In the United States more than a hundred degree-granting colleges and universities now have departments of meteorology, or offer programmes in meteorology which include courses in weather forecasting methods. To illustrate the seriousness of this training, we note that students at various campuses compete against each other in weekly forecast. This training in real-life forecasting provides students with hands-on experience, and has had the positive result that many of the best minds in the field, after graduation, have directed their research toward improving forecast range and accuracy.

In summary, weather forecasting has advanced through a series of stages, most of which, if not all, can be seen in retrospect to be necessary steps for any environmental service to ascend. These stages are as follows: 1. There must first be some impact on society that calls for a prediction capability, which is a prerequisite stage; 2, predictions based on visual observations; 3, predictions based on instrument observations; 4, synoptic imagery using multi-station data (a necessary prerequisite for further advances); 5, real-time data transfer to a central office; 6, predictions based on advection of static storm structure; 7, predictions based on advection of an evolving storm structure (subjective analysis); 8, objective analysis; 9, numerical predictions; and 10, storm tracking. Numbers 8, 9 and 10 characterize the most advanced stage of forecast and nowcast capability. Regarding number 9, the advent in an environmental service of physics-based numerical prediction codes marks the stage at which forecast skill can be assured of continual improvements. Even so, improvement takes place incrementally over a long time, as Figure 2.3 shows. There is neither a discontinuity in forecast capability nor a final correct answer that must first be achieved before numerical predictions can begin. The advice from the meteorologists is that the sooner an environmental science starts numerical predictions, the sooner it can begin to move up the ladder of increasing skill scores. Supplementing these stages toward forecast improvement is the critical step of teaching forecasting methods to future researchers.

This section started by discussing storms, since these were the first meteorological systems to have an impact strong enough to generate scientific interest and to motivate the creation of weather prediction services. But in the quotidian forecasts of tomorrow's weather, be it fair or stormy, prediction services, once created, do good by facilitating everyday planning by those whose activities expose them to the weather. The same is true in space weather, as explained below.

2.3 RELEVANT ANALOGIES BETWEEN TERRESTRIAL WEATHER AND SPACE WEATHER

Before proceeding to apply to space weather the chronology of advances in forecasting as derived from the example of meteorology, we look at space weather analogues of terrestrial weather. First regarding storms, there is a useful division of storms into types based on relative size. The sizes of storms in the terrestrial case are three, the largest of which is the extratropical cyclone (the main subject of the previous section), which is a mid-latitude low-pressure centre that migrates from west to east (the direction of Earth's rotation, note) with cold and warm fronts attached like swinging arms, often spanning a whole continent. Next smaller in size is the hurricane – a mostly tropical storm, which though smaller than the extratropical cyclone is more intense. The origins and dynamics of these two storm systems are basically different. The smallest of the three storm systems is the tornado, which though locally the most intense of the three, owes its existence to the operation of the other two.

Space weather storms can also be divided into three types by size, the largest of which is the M-region storm typically comprising a fast solar wind stream from a coronal hole and the corotating interaction region (CIR) at its leading edge. (The designation 'M-region' storm is not new. It derives from the old designation M-region – M signifying 'magnetically effective' – coined by Bartels in 1932 to refer to long-lived regions on the Sun that produce recurring magnetic disturbances. An appellation such as 'M-region storm' is required here because the disturbance under discussion often results not just from a CIR but from a combination of a CIR and a fast stream (Crooker and Cliver, 1994).) The M-region storm is the space weather analogue of the extratropical cyclone, with the CIR being the attendant cold front. The storm it produces follows its arrival, in analogy with a terrestrial cold front and in contrast to a terrestrial warm front whose storm precedes it. One difference between a CIR and a cold front is that the former is a high-pressure feature, whereas the latter is a low-pressure feature. Next smaller in size, yet greater in intensity, is the coronal mass ejection (CME). It is the obvious space weather analogue of the terrestrial hurricane, and is often referred to as such. M-region storms and CMEs, like extratropical cyclones and hurricanes, differ basically in their origins and dynamics. Smallest in size of the three space-weather storm types, but greatest in intensity as measured by magnetic perturbation, is the auroral electrojet whose rapid fluctuations of the magnetic field it generates at ground level, dB/dt, induce ground currents big enough to disturb system operations in the industry that delivers electrical power or cable communication. Both positive and negative electrojets have been known to do this (Boteler and van Beek, 1999; Bulduc *et al.*, 2000). Such super geomagnetically induced current (GIC)-inducing auroral electrojets (SAEs) we take to be the space weather analogue of terrestrial tornadoes. Like tornadoes, SAEs owe their existence to the operation of the two larger space weather storm systems.

Analogies between extratropical cyclones, hurricanes, and tornadoes on one hand and M-region storms, CMEs, and SAEs on the other go beyond their relative sizes and intensities to their modes of prediction and tracking, as the following section points out. But we need to recognize one important way in which terrestrial and space storms differ. Terrestrial storms – all three types – inflict their damage directly; that is, the parameters of the storm, for example, wind speed and precipitation, are the very same parameters that inflict the damage. M-region storms and CMEs, on the other hand, except when considered in connection with extraterrestrial space missions, inflict their damage indirectly through the intermediary of the magnetosphere. The parameters of these storms induce a secondary storm within the magnetosphere, and it is the parameters of the secondary storm within the magnetosphere that inflict the damage. One of these secondary-storm parameters, for instance, is dB/dt of the SAE. In the case of space weather, therefore, one must consider, besides parameters of the M-region storms and CMEs, parameters of the secondary magnetospheric storms that they induce.

Beyond warning and alerting customers of impending space storms, space weather forecasts, in their day-to-day role, serve radio communication engineers who plan best frequencies to transmit via ionospheric reflection – which is affected by solar and magnetospheric activity – from hours to months in advance. For this purpose they use empirical equations that relate the ionospheric critical frequencies, f0E, f0F1, and f0F2, to geomagnetic activity or the sunspot number (Mitra, 1947), which must be predicted (or solar F10.7 radio flux, which is a proxy for solar EUV).

2.4 STEPS IN THE ADVANCE OF SPACE WEATHER FORECASTING

2.4.1 Stage 1: social impacts

The first stage in our recounting of the advance of an environmental science toward forecast capability is the emergence of social impacts that create a need for forecasts. Whereas a need for meteorological forecasting can be traced at least to the advent of farming in prehistory, a need for space weather forecasting did not arise until the advent of electricity-based technology in the mid-nineteenth century, for it is this technology that space weather affects. The initial technology so impacted was the electrical telegraph, and the earliest recorded such incident occurred on Friday, 19 March 1847, merely eight years after the first telegraph line was set up in England, when, as the operator of the Derby telegraph station reported, ‘a brilliant aurora was seen, and during the whole time of its remaining visible, strong alternating deflections occurred on all the (telegraph) instruments’ (Barlow, 1849). This disturbance was relatively minor compared to a great magnetic storm that occurred one solar cycle later, in 1859, and disrupted telegraph communication globally. Telegraph service was affected in at least 52 cities worldwide according to a thorough tally made at the time (Loomis, 1860a, b, c, and 1861). Worldwide disruptions of telegraph service during major magnetic storms continued into the twentieth century as, for example, on 14 May 1921, a magnetic storm damaged telegraph equipment at several locations and set fire to a telegraph station in Brewster, New York. (A collection of reports on the disturbances caused by this storm is contained in *Mon. Weather. Rev.*, **49**, 406, 1921.) More recently, disturbances on transatlantic submarine cables owing to magnetic storms have occurred in 1957, 1958, and 1960 (Winkler *et al.*, 1959; Axe, 1968) and on a long line ground cable in 1972 (Anderson *et al.*, 1974).

The advent of wireless telegraphy (radio) gave rise to the next social sector affected by space weather. Retrospectively from 1930, transatlantic radio reception could already be seen to rise and fall with magnetic activity in 1924, five years after its inauguration (Austin, 1930). For example, during a magnetic storm on 8 July 1928, transatlantic radio signal strength dropped about 30 db while the horizontal magnetic field strength (a rough proxy for Dst) dropped about 500 nT (Anderson, 1929). Magnetic storm disturbances of the ionosphere continue to disrupt HF radio communications in the airline business and in military operations. For example, the US Federal Aviation Administration requires commercial flight dispatchers to consider HF communication degradation for each polar flight and to reroute if communication is threatened.

After radio communication, next to be affected by space weather was the electric power industry. By 1940, electric power lines had extended far enough from generating plants to suffer disturbances from GICs. The first such incident in North America occurred on 24 March 1940, during a storm which also disrupted transatlantic cable and radio communications; so for the first time all three of the services then vulnerable to space weather – cable telegraphy, radio, and electric power – were impacted (McNish, 1940).

The last of the major societal sectors to be affected by space weather that we wish to mention is the industry that dispenses satellite services such as entertainment and communication relays and navigation. In the satellite group in general, one might name as the original victim of the space environment that very first mission in 1958 that discovered Earth’s radiation belts, Explorer III, seeing that its Geiger tubes saturated during the first orbit. Although they were thus rendered useless for quantitative measurements, they nonetheless allowed Van Allen and his team to recognize the radioactive hazard that space presents. Of course, this hazard still continues to seriously impact the industry that provides services by means of satellites (Odenwald, 2001, Chapter 6).

To conclude this précis of the prerequisite stage for the advance of environmental forecasting as it applies to space weather (the societal-impacts stage) we note that the US Space Environment Center, which is the government office authorized to provide real-time monitoring and forecasting of solar–terrestrial events, now lists twenty-five user groups as customers. This number increased, keeping pace with the growth of space-weather-vulnerable technologies, from one in the early 1940s, when the United States and other countries began to forecast ionospheric disturbances to support HF radio communication during the war, to the present number of twenty-five, the latest of which stands for cellular telephones.

2.4.2 Stage 2: visual observations

After the prerequisite, social-impacts stage there follows the stage of forecasting by means of visual identification of conditions likely to cause disturbances. In meteorology these are weather signs, which solar haloes epitomize. In space weather, sighting aurorae above a telegraph line was known from the beginning (almost) of electric telegraphy to signify disturbances to the line, as shown by the 1849 quote by Barlow (given earlier). Sighting aurorae is not forecasting, however, but nowcasting, since if an aurora is seen, the storm is already in progress. As for forecasting, an obvious space weather analogue of solar halos is sunspots. In 1855 Edward Sabine, a British geophysicist, discovered a correlation between magnetic disturbances and sunspots, with a correspondence between the quasi-decadal cycle of sunspot frequency (discovered not long before by Heinrich Schwabe and now called the 11-year cycle) and a similar cycle in the list of magnetic disturbances recorded by observatories under his supervision (see Cliver, 1994a for details).

Over the next fifty-or-so years, Sabine's statistical correlation between sunspots and magnetic disturbances was supported by instances of major storms following the central meridian passage of large sunspot groups. Such instances led, in 1941, to the first reported civilian prediction of a magnetic storm several days in advance of its occurrence. The story is based on visual observations by H. W. Wells of the Department of Terrestrial Magnetism of the Carnegie Institution of Washington, who alerted radio transmission engineers of probable loss of signals. The storm came to pass, and transatlantic HF radio communication indeed suffered, as predicted. The forecast was 'made through close study of daily reports of areas, numbers and locations of sunspots supplied by the United States Naval Observatory' (Fleming, 1943, p. 203). Figure 2.4 shows the responsible spot group a day after having crossed the central meridian and a day before the storm struck.

In his warnings to the radio engineers, Wells must have recognized that he was taking a risk, for the correspondence between magnetic storms and even big sunspots was known to be imperfect (Cliver, 1994a, b). He might have trusted his forecast more had he taken into account another 'visual' sign besides the menacing active region drifting across the central meridian; namely, the eruption of a major solar

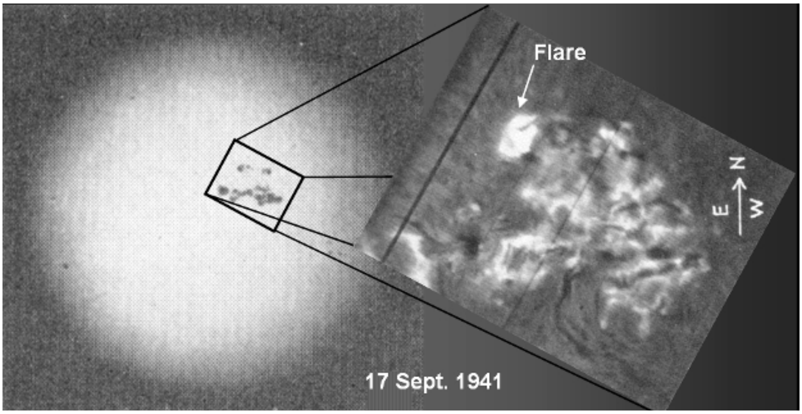


Figure 2.4. Sunspot group that enabled the first formal forecast of radio disruptions. The blowup shows the associated flare on 17 September 1941 (constructed from Fleming, 1943).

flare very near the central meridian associated with the same active region (also shown in Figure 2.4). Already in 1908, George Ellery Hale had suggested that, rather than sunspots, the origin of magnetic storms 'may be sought with more hope of success in the eruptions shown on spectroheliograph plates (solar flares) in the regions surrounding spots' (Hale, 1908, p. 341). And indeed, by about the time that Wells made his forecast in 1941 it had been established that: 'There is a high probability that the more intense solar flares (class +3) occurring within the central half of the Sun's disk will be followed about 24 hours later by a great geomagnetic storm' (Newton, 1958, p. 134).

Sunspots and solar flares have been, and remain, key visual signatures on which space weather forecasters base predictions of magnetic storms. Of course, by now space-based data sources have added other diagnostic predictors (such as halo CMEs) to these ground-based visual records. One of the new space-based predictors should be mentioned here because of its purely visual nature. This is the sigmoid feature, imaged in soft X-rays, that solar active regions often develop before they erupt (Canfield *et al.*, 1999), and because of its predictive value has been developed into an online service for forecasting CME eruptions (*http://sd-www.jhuapl.edu/UPOS/CME/index.html*).

2.4.3 Stages 3 and 4: instrument observations and synoptic images

In the advance of environmental predictions, after forecasts based on visual observations (stage 2) come forecasts based on instrument observations (stage 3). In meteorology the barometer was the enabling instrument, the magnetometer in space weather. As with meteorology, instrumented space weather observations increased the forecast range little over that of visual observations, but it increased the accuracy of nowcasts. Like the barometer, the magnetometer's chief role in advancing forecasting capability was to inspire people to combine data from multiple stations to construct synoptic images of magnetic disturbances, which is stage 4 in our sequence.

In meteorology the history of the synoptic-image stage is straightforward. When the locations of barometer readings grew wide and dense enough, their combination produced a low-resolution image of a moving area of low pressure, which defined the storm. The corresponding history in space weather is not so straightforward, in part because, as previously mentioned, storms from the Sun (M-region storms and CMEs) drive secondary storms within the magnetosphere, and in part because there are two kinds of storms within the magnetosphere: magnetic storms and substorms (SAEs being perhaps manifestations of extreme forms of the latter). Therefore, in space weather the history of the synoptic-image stage has two parts – one concerning M-region storms and CMEs, and the other concerning magnetic storms and substorms, which chronologically came first.

The key incidents in the magnetometer investigation of storms within the magnetosphere are the following (condensed from Chapman and Bartels, 1962, Chapter 26). First, in 1722, George Graham noticed that the magnetic needle of sensitive instruments he built to measure the secular variation in the declination of the geomagnetic field at London moved from day to day and sometimes from hour to hour. Graham thereby discovered geomagnetic disturbance. Intrigued by this, in 1741 Anders Celsius, in Uppsala, exchanged notes with Graham to determine whether magnetic disturbances in England and Sweden coincided, and found that they did. Magnetic disturbances therefore had considerable areal extent. Five years later, Olof Hiorter, Celsius' brother-in-law, observed that a strong magnetic disturbance coincided with a bright auroral display. A connection between magnetic disturbances and auroral activity was thereby implied and, through repeated observations, established.

These findings opened a new field of geomagnetic observations which by 1841 had built up an international network of magnetic observatories carrying out simultaneous observations according to a prearranged schedule. The leading spirit behind this effort was Alexander von Humboldt, who in 1806/07 in Berlin, observed magnetic disturbances around midnight on subsequent nights on several occasions and wondered whether the midnight, temporal localization of the occurrences implied spatial localization as well (Humboldt, 1864, p. 185). Humboldt called these disturbances 'magnetic storms' (ironically, since they occurred around midnight they would now be called substorms). The idea that, like terrestrial storms, magnetic storms might have a specific areal footprint offered a promising subject for investigation. Thus at the same time that meteorologists were expanding their network of instrumented observations, Humboldt called for a network of magnetic observatories to carry out synoptic observations of magnetic storms.

The global magnetometer network that resulted, comprising mostly mid-latitude stations, revealed that instead of being localized like terrestrial storms, magnetic storms occurred simultaneously worldwide. Humboldt's project of constructing a synoptic image of a magnetic storm was therefore replaced by one of discerning regularities between variations seen simultaneously around the globe. The first regularity discovered concerned the horizontal component of the magnetic field (Broun, 1861) which quickly came to be understood to entail an abrupt increase followed by a deep decrease and slow recovery of the horizontal component that characterizes

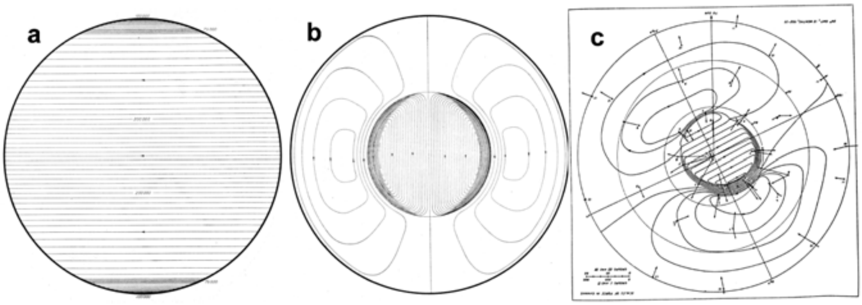


Figure 2.5. (a) Electric current system of the symmetrical part of magnetic storms (Chapman, 1935). (b) Eastward and westward auroral electrojet equivalent current systems that represent the non-symmetrical part of the magnetic storm disturbance field (Chapman, 1935). (c) Average equivalent current system for magnetic bays (Silsbee and Vestine, 1942).

the magnetic storm, as it is now called. A method of representing a synoptic image of this storm was devised by Chapman (1935) in terms of equivalent overhead currents consisting of circles parallel to lines of latitude with a concentration at the auroral zones to acknowledge the tendency of the storm-time disturbance to maximize there (Figure 2.5a). Chapman also described a non-symmetric storm-time equivalent current system associated with the (now-called) auroral electrojets (Figure 2.5b), which brings us closer to a synoptic representation that might be useful for forecasting SAEs.

The disturbance that is now called a substorm, which initially motivated the installation of magnetometer networks and which (as Humboldt surmised) has a true synoptic magnetic footprint, did not emerge as a defined storm system with its own unique character distinct from the globe-spanning magnetic storm until 1908, when Kristian Birkeland identified it as such and named it the elementary polar storm (Birkeland, 1908). Subsequently, the magnetic signature of a substorm was referred to as a magnetic bay (Figure 2.5c) and relegated to a class of minor disturbances, which included magnetic pulsations (Chapman and Bartels, 1962, Chapter 10). From a space weather perspective, the isolated polar elementary storm-aka magnetic bay-aka substorm can cause radio blackouts in the auroral regions (Wells, 1947) and high-altitude spacecraft charging (Spence *et al.*, 1993). It can also be a major player as part of a full-up magnetic storm when it compounds the effects of auroral electrojets. In summary, Figure 2.5 displays the basic synoptic characteristics of storms in the magnetosphere analogous to early synoptic maps of terrestrial storms (Figure 2.1c).

Turning from secondary storms in the magnetosphere to primary storms from the Sun – M-region storms and CMEs – the quote by G. E. Hale, given earlier, shows that by 1908 there was already a hint that two classes of solar storm existed – one associated with drifting sunspots, which often recurred after a solar rotation, and one associated with solar flares. The distinction became explicit by 1932 when, as previously mentioned, Julius Bartels designated the solar sources of recurring storms as M regions and emphasized that they often corresponded to no visible solar

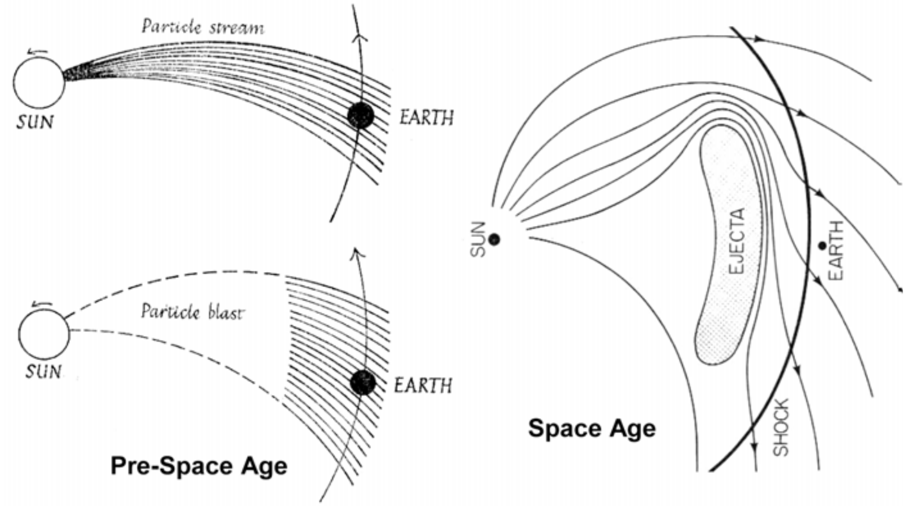


Figure 2.6. Pre-space-age synoptic images of M-region and CME storms (modified from Newton, 1958) and a space-age synoptic image of the CME storm (from Cane, 1988).

feature. In 1939 the two types were referred to as the 'beam- and flare-theories of solar outbursts' (Hulburt, 1939, p. 562). The geometry of the long-lived beam of solar particles responsible for M-region storms was easy to represent, since the duration of the storms was known (giving the angular width of the beam as seen from the Sun) and the speed at which particles in the beam travelled outward was known from the delay between sunspots crossing the central meridian and storm onset (giving the curvature of the spiral of the beam owing to the rotation of the Sun). Figure 2.6 shows typical synoptic images of the M-region and CME storms as drawn before and after the advent of space data. The space-age image of the M-region storm is essentially the same as the pre-space-age image (Hundhausen, 1977). The synoptic image of the CME changed, however, by addition of features. In words, the pre-space-age image of the CME storm was something like a 'frontal shield' preceding a cone of solar material (Bartels, 1940, p. 343) or a 'corpuscular cloud ... blown out bodily', forming at the orbit of Earth a shell like that 'blown off from ordinary nova' (Kiepenheuer, 1953, p. 449). The space-age synoptic image adds a bow shock and the interplanetary magnetic field. It was, however, the pre-space-age images that forecasters had in mind in the early days of space weather forecasting.

2.4.4 Stages 5 and 6: real-time predictions based on advection of static structures

In meteorology, three circumstances united to bring real-time weather predictions into being. A scientific basis for prediction had been established: west-to-east advection of an areally confined storm system. A technology came into being to transmit information about the existence and motion of a storm in advance of its arrival: the telegraph network. And a compelling benefit of such predictions could be envisioned: mainly safeguarding life and commercial assets; for example, ships at sea and on the Great Lakes. These motivated and enabled governments in the 1860s and 1870s to establish weather-forecast centres. In the case of space weather, governments established forecast centres in the 1940s, during the Second World War, to support military shortwave radio operations. The technological means for transmitting real-time forecasts, of course, already existed. Centres to forecast radio-transmission quality based on solar observations were set up in Australia, Canada, Great Britain, Germany, the Soviet Union, and the United States (Hufbauer, 1991, pp. 120–129). Because the Second World War played out during the declining phase of Solar Cycle 17, when recurring streams and associated M-region storms predominated over CME storms (with a notable exception in September 1941, highlighted earlier), the principle on which forecasts were first made resembled that initially used in meteorology: advection of static structures. In this case the static structure was either a recurring M region – used to forecast magnetic disturbance – or a recurring sunspot group – used to forecast sunspot number. The advent of the coronagraph allowed east-limb observations of the coronal green line – the brightness of which correlates with geomagnetic activity – to be used to update predictions based on 27-day recurrence (Shapley and Roberts, 1946). Figure 2.7 reproduces these correlations as depicted in the first (unclassified) report in the United States on weekly, operational, 4–10-day forecasts of geomagnetic activity.

Regarding forecast method, Shapley (1946) states: 'Short term forecasts are based primarily upon two considerations: (1) The geomagnetic activity 27 days before the forecast-period and (2) an estimate of disturbance based on the location and degree of activity of solar regions during the period covered by the forecast' (p. 249). Method (2) rules when the 27-day recurrence tendency is weak. Then 'the most important criteria for forecasting disturbances from solar data' are

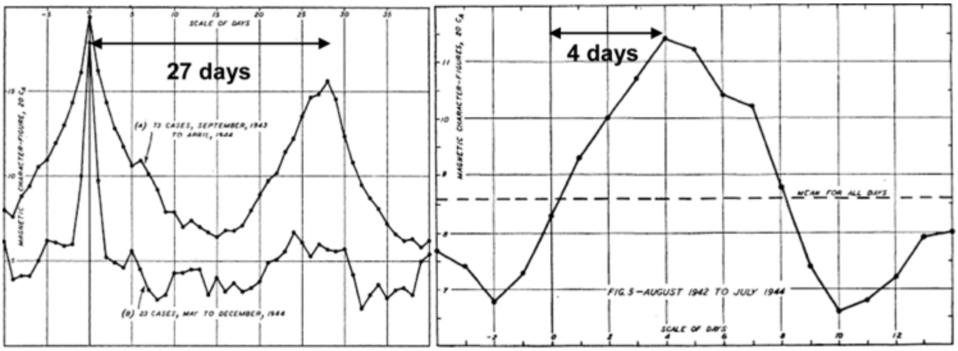


Figure 2.7. Illustrating the 27-day recurrence of magnetic activity, 1944 (left) and the correlation of the coronal green line with geomagnetic activity, which peaks about four days after east-limb brightening, 1942 to 1944 (right) (Shapley and Roberts, 1946). Both were used at the onset of space weather services to make short-term forecasts of ionospheric conditions pertinent to radio transmission quality (Shapley, 1946).

'the position on the solar disk and amount of activity in (regions that may be associated with geomagnetic disturbance)' (pp. 250/251). After all pertinent factors are evaluated, the 'estimates of time of disturbance based on solar data are reconciled with the 27-day recurrence-data to form the final forecast. Subjective weighting is applied when the two disagree' (p. 251). Subjectivity enters the forecast, therefore, in the weighting of the observations. In effect the forecast is a form of subjective abstract pattern recognition. Subjectivity remains a significant factor in space weather forecasting, as the next section describes.

Since the early war days, space weather forecasting, while continuing to be an activity conducted within various military agencies, also became institutionalized in offices associated with civilian environmental services. In the United States this was the Central Radio Propagation Lab (a predecessor of the Space Environment Center, SEC), which made its first official forecast in 1965. Worldwide, about a dozen Regional Warning Centers (including SEC in Boulder, which also plays the role of World Warning Agency) now constitute the operational arm of the International Space Environment Service (ISES). These RWCs distribute warnings and alerts as conditions warrant, and daily reports and forecast of standard products. Clearly, the requirements of stage 6 have been met.

2.4.5 Stage 7: subjective analysis

In meteorology, subjective analysis refers to that method of prediction in which a weather forecaster – with long practical experience applying concepts of air masses and fronts, and who recognizes patterns that storm systems take as they move, mature, and dissipate – after examining weather charts, makes a judgment about the ensuing day's weather. The judgment is subjective because given the same information two forecasters can produce different forecasts. Subjective analysis describes the state of forecast capability that preceded objective and numerical forecasts, and concerning which, as noted earlier, Ludlam and Scorer ascribed a 20–30-year stasis in forecast accuracy.

The concept of subjective analysis, as Shapley indicated, also applies to space weather forecasting. In the case of recurrent storms (the beam picture), one can update the forecast with the coronal green line and use probabilities of geomagnetic disturbance based on previous cases. Such procedure is an early form of objective analysis. But to forecast a solar eruption (flare or CME – the Holy Grail of space weather forecasting) in lieu of an established, proven eruption mechanism, pattern recognition (including how the pattern changes in time) was, and still is, mainly all that the forecaster can utilise, and this procedure is inherently subjective. (Nonetheless, we note that objective algorithms for flare predictions have been proposed (Bartkowiak and Jakimiec, 1986; Neidig, 1986), but apparently these have not been incorporated into routine forecast procedures.) Joseph Hirman, a forecaster at the Space Environment Center in Boulder, stated explicitly: 'Experience plays a big part in the SESC (a forerunner of SEC) forecasting process' (Hirman, 1986, p. 384). In the subjective stage of meteorology, forecasters had an established mechanism that explained the birth and development of the extratropical cyclone (the polar front theory); consequently, subjective forecasting developed into a highly complex procedure entailing much quantitative analysis to evolve a current weather pattern into a future weather pattern using, for example, the wind fields to advect fronts and low-pressure centres forward. The procedure, called synoptic air mass analysis, was explicit, and could be taught in textbooks such as Petterssen's *Weather Analysis and Forecasting* (1940). The stage of teachable procedures has not been attained in space weather forecasting of solar eruptions. Instead, according to Hirman: 'The SESC forecasters (that he) surveyed had no universal approach to evaluation (of information leading to a flare forecast)' (Hirman, 1986, p. 387). But certain parameters – for example, the location, complexity, and form of the inversion line (aka neutral line) and field strength gradients – appear to be important in most flare forecasts (Simon and McIntosh, 1972). The Space Environment Center in Boulder experimented with an automated expert system, which encoded pattern-recognition procedures that the best forecasters used for forecasting flares, but abandoned it in favour of hands-on predictions (Kunches and Carpenter, 1990; Patrick McIntosh, private communication, 2005).

Excepting solar eruptions, teachable subjective analysis procedures have been developed in space weather for improving forecasts of M-region storms such as coronal hole images (Sheeley *et al.*, 1976), and IMF polarity and the Russell–McPherron effect (Crooker and Cliver, 1994), and of the geomagnetic response to CMEs such as Bz polarity in the associated magnetic cloud (Bothmer and Rust, 1997; Mulligan *et al.*, 1998). These are illustrated in Figure 2.8.

Subjective analysis procedures dominated forecast meteorology up to the advent of numerical weather prediction around 1960, although objective procedures also grew in importance. At this point in our narrative the history of forecast meteorology continues while that of space weather forecasting stops because we are at the current stage of the field. By analogy to the advance of forecast meteorology, space weather is in the 1950s before numerical weather prediction was inaugurated, and at the end of a period of subjective forecasting characterized (in forecast meteorology) by 'twenty to thirty years' of no progress. Figure 2.9 indicates a similar situation for

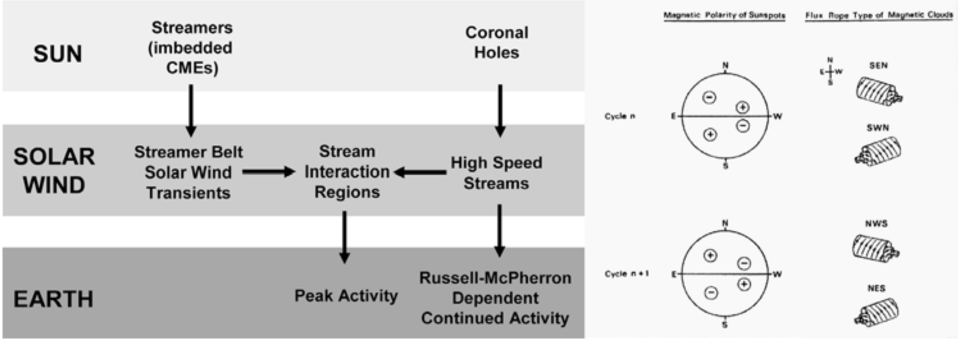


Figure 2.8. Illustrating subjective analysis guidelines for predicting M-region storms (left, modified from Crooker and Cliver, 1994) and CME storms (right, modified from Bothmer and Rust, 1997).

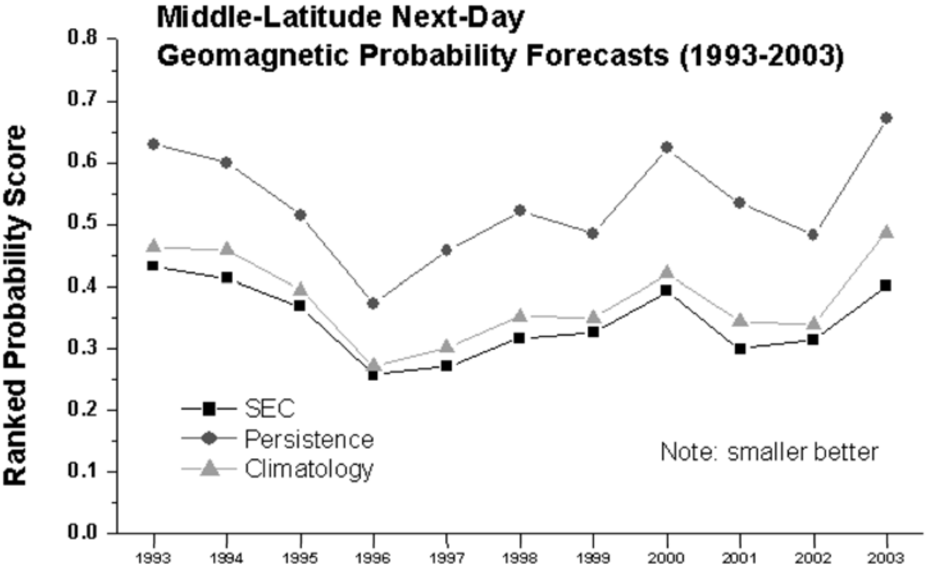


Figure 2.9. Showing relative stasis from 1993 to 2003 in the skill of a standard space weather forecast product. The quantity plotted relates to the highest daily value of the Fredericksburg K index and is the mean square error of forecast probability. Therefore, smaller numbers correspond to greater skill (from SEC web site).

space weather, which, though on the threshold of implementing numerical forecast algorithms, has experienced a period of subjective forecasting characterized by at least ten years of relative stasis in predicting the day-to-day value of geomagnetic activity. On the other hand, space weather forecasts of CME storms have improved remarkably in recent years, owing in part to the real-time availability of observations of halo CMEs and solar wind data from L1 (good examples of visual- and instrument-observation forecasting). Nonetheless, the contrast between Figures 2.3 and 2.9 highlights the 'before and after' difference that numerical weather prediction can make.

Since, as mentioned, space weather has yet to enter the stage of operational Sun-to-Earth numerical predictions that can provide general, multi-purpose forecasts, forecasting with objective algorithms is expanding to fill the need. This is in contrast to meteorology, where objective algorithms are tuned mainly to provide local forecasts. As we discuss next, in space weather, objective algorithms are usurping the role of providing global and regional forecasts that in meteorology is the purview of numerical weather predictions.

2.4.6 Stage 8: objective space weather forecasting

Objective weather forecasting has been defined as 'any method of deriving a forecast which does not depend for its accuracy upon the forecasting experience or the subjective judgment of the meteorologist using it' (Allen and Vernon, 1951). Creating an objective forecast algorithm typically entails multiple linear regression, where the predictors are past-time or real-time observables and the predictand is a weather element of interest. In space weather there are many examples of objective algorithms for predicting virtually all geomagnetic indices from L1 data. These, of course, give less than a one-hour warning of oncoming disturbance, and so can be used in operational forecasting only for sending alerts. (We note the irony that the objective algorithm with the highest skill (Temerin and Li, 2002) predicts Dst from L1 data, but outside of the scientific community there is virtually no call for Dst forecasts.)

Objective algorithms with greatest operational utility predict space weather parameters more than one hour ahead. (For economy of words we leave out of discussion medium- and long-term predictions of sunspot numbers which are based almost exclusively on objective forecast algorithms; see, for example, reports in Simon *et al.*, 1986.) Examples of more-than-one-hour objective forecast algorithms being used to generate operational space weather products available through the RWCs include a neural net algorithm that generates the global magnetic disturbance index called Kp three hours ahead (Costello, 1997); the Wang–Sheeley semi-empirical model that generates solar wind speed and IMF polarity profiles for about three days ahead (Arge and Pizzo, 2000); a linear prediction filter code (based on Baker *et al.*, 1990) that predicts relativistic electron flux at 6.6 Re several days ahead; and a code developed by R. L. McPherron that applies extended auto-regressive techniques to prior Ap values to predict the following day's Ap index, soon to be a test product at SEC (Onsager, private communication, 2005).

In addition, many space weather products offered by individual servers using locally generated objective algorithms are available on the web (for example, *http://www.ava.fmi.fi/spee/links_forecast.html* and *http://sd-www.jhuapl.edu/UPOS/spaceweather.html*). Of particular interest are services offered to nowcast and forecast GICs (the tornadoes of space weather), since this service is not explicitly addressed by standard products of official space weather centers (for example, *http://www.metatechcorp.com/aps/PowerCastFrame.html*, a commercial service, *http://sd-www.jhuapl.edu/UPOS/FAC/index.html*, and *http://www.lund.irf.se/gicpilot/gicforecastprototype/*).

To illustrate the point that the role of objective forecast algorithms is expanding in scope and importance in lieu (presently) of an operational capability to perform numerical space weather predictions, we recount the experience of a multi-institutional research-and-development project in the United States to produce an end-to-end – Sun-to-ionosphere/thermosphere – numerical space weather prediction code. The project, carried out under the rubric Center for Integrated Space Weather Modeling (CISM) (Hughes and Hudson, 2004), entails interactively coupling together like links in a chain separate physics-based numerical codes. Starting from the Sun, the separate links correspond to the solar surface, the extended corona, the solar wind out to Earth, the magnetosphere, and the ionosphere/thermosphere. The time scale to achieve a tested and validated prototype of the coupled code is ten years.

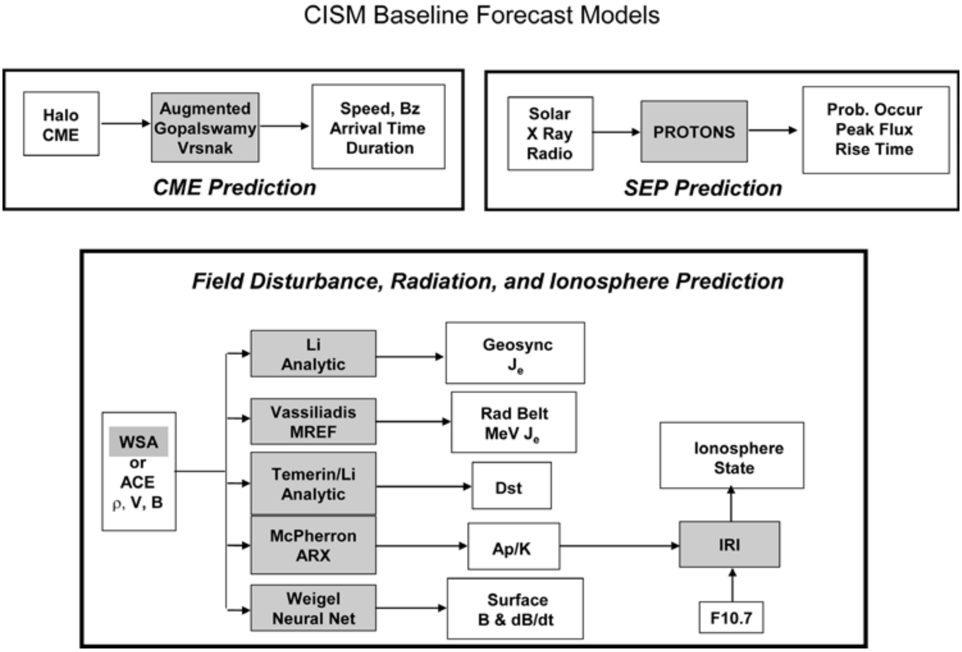


Figure 2.10. A comprehensive suite of objective forecast algorithms being assembled within the CISM project to supplement the development of a sun-to-ionosphere/thermosphere numerical space weather prediction code (see Siscoe *et al.*, 2004, for descriptions of the named boxes).

Meanwhile, to attempt to produce something of value on a shorter time scale, CISM is assembling a suite of new or existing objective algorithms that cover many of the parameters of interest to operational space weather forecasting (Figure 2.10). Versions of these algorithms already exist, but the problem in implementing them is the pragmatic one of producing operational quality codes (fully tested and validated, easy to use, and operate reliably under ‘battle conditions’), which takes time and resources. The point of the example for the present discussion, however, is simply that the space weather research community has achieved the technical capability to generate a broad suite of objective forecast algorithms that, though far from fully implemented, can collectively address global, regional, and local forecast needs in space weather; and this defines the present, inherent operational capability of the field.

2.4.7 Stage 9: numerical space weather prediction

As we have emphasized, probably the most significant lesson that space weather forecasting has learned from forecast meteorology is the importance of achieving stage 9 – numerical weather prediction. Once this stage is reached, improvements in machine performance, coding efficiency, data acquisition, data assimilation, and coding of the physics, combine into a self-propelled escalator that carries forecast accuracy steadily up the skill curve. Indeed, the escalator metaphor lies behind the oft-repeated imperative to begin numerical weather prediction as soon as possible, regardless of how poorly the initial effort performs. For example: 'Need to approach the problem (of numerical space weather prediction) incrementally – cannot wait for 100% solution – get something on-line fast and then incrementally grow and improve' (Tascione and Preble, 1996, p. 25); and 'The point is that we will not get better unless we start (numerical space weather prediction), even if the early models are crude and give less accurate predictions than experienced solar observers' (Hildner, 1996, p. 624).

So powerful is the draw to achieve numerical prediction of space weather that the first indication in the early 1990s that it could realistically happen – the advent of global magnetospheric MHD codes and the implementation of the Magnetospheric Specification and Forecast Model (MSFM, a physics-based code that uses data from L1 to predict the energetic particle environment within the magnetosphere, described in Freeman, 2001) at the Air Force Space Weather Forecast Center – provided strong incentive for the creation of the US National Space Weather Program (NSWP). The implementation plan of the NWSP states that its 'ultimate goal is to develop an operational model that incorporates basic physical understanding to enable specification and forecasting of the space environment by following the flow of energy from the Sun to Earth. This coupled system of models is to be constructed by merging parallel models for the solar/solar wind, the magnetosphere, and the ionosphere/thermosphere (Section 3.2).' The CISM project, described above, aims at producing a prototype of such a model, and has succeeded in following an event continuously through each of its coupled codes from the Sun to the ionosphere (Luhmann *et al.*, 2004).

Other Sun-to-ionosphere numerical models besides CISM are being developed; for example, the Space Weather Modeling Framework code at the University of Michigan. There is an expectation that some form of operational numerical space weather prediction will start in the near future. Whichever programme is achieved first will be rewarded by being at the front in riding up the escalator of improving skill scores.

2.4.8 Stage 10: storm tracking

Analogies exist in space weather to tracking hurricanes and extratropical cyclones by satellite imagery and tornadoes by radar. The space weather analogue of hurricane tracking is tracking CMEs with space-based coronagraph images provided by SOHO and SMEI (Plunkett *et al.*, 1998). These give initial speeds of Earth-directed CMEs (projected onto the viewing plane), which can be converted into arrival times by means of empirical models (objective forecasting) (Gopalswamy *et al.*, 2001). Presumably, real-time CME storm tracking will become more robust after images from the STEREO mission (scheduled for 2006) become available. Regarding M-region storms, interplanetary scintillation data are being assimilated into real-time tomographic images of the solar wind capable of tracking the interaction region where the disturbance is greatest (Hick and Jackson, 2004). When made operational, this capability would be analogous to tracking extratropical cyclones with satellite imagery. Regarding tracking super GIC-inducing auroral electrojets (SAEs – space weather tornadoes), the equatorward boundary of the auroral oval is probably as close as current capability allows. (A real-time auroral oval boundary tracker exists at *http://sd-www.jhuapl.edu/UPOS/AOVAL/index.html.*)

2.4.9 Critical supplementary step: university teaching of space weather forecasting

One lesson from forecast meteorology that seems to have gone unlearned in space weather forecasting is the benefit to the profession to be had by exposing higher-education space physics students to real-life space weather forecasting; that is, to forecasting the same parameters that constitute the daily task of operational space weather forecasters. In this way students would be frustrated by the same limitation on the skill they can achieve with existing subjective forecast rules and objective forecast algorithms. If operational space weather forecasting were taught as part of the space physics curriculum, the result would surely be a steady stream of space physicists strongly motivated to improve existing forecast rules and algorithms and to advance the application of numerical space weather prediction.

2.5 IMPORTANT COMPARATIVE TOPICS NOT COVERED

Important topics in the comparison of forecast meteorology and space weather forecasting not covered in this article include the role of private sector forecasting services, which through newspapers, radio and television have made meteorology the most publicly viewed of all the sciences, but has yet to develop to its potential in space weather forecasting (though several entrepreneurs are working hard to change this).

Another missing topic is the different roles that data assimilation plays in the two forecast arenas. In numerical weather prediction, data assimilation is absolutely essential because the equations that govern the general circulation of the atmosphere manifest deterministic chaos, and so numerical integrations must constantly be nudged back to the values that the real atmosphere displays. Within the magnetosphere, however, weather is to a large extent externally driven, so that variations of the external drivers causally determine most of what happens within. This accounts for the success, for example, of the Temerin and Li objective forecast algorithm for Dst.

Yet another topic omitted is the inherent immeasurability of one of the key external drivers of magnetospheric weather; namely, the z-component of the IMF. For any forecast longer than the L1 to Earth transit time ($\sim$1 hour), and excepting large-scale orderings imposed by the IMF sector structure (such as the Russell–McPherron effect) and the internal order of CMEs, IMF Bz is a stochastic quantity and must be forecast probabilistically (McPherron and Siscoe, 2004).

We also have left out of the discussion (except for storm tracking) the important topic of nowcasting, which is a big part of the operational weather services provided in both arenas.

2.6 SUMMARY

Generic analogies exist between forecast meteorology and space weather forecasting. Besides sharing a common dynamical paradigm in the form of continuum mechanics, it also happens that both disciplines are concerned with three types of storm systems that have parallel aspects in relative sizes, energies, and difficulty of predictability. Arranged hierarchically by size, these are extratropical cyclones and M-region storms, hurricanes and CMEs, tornadoes and super GIC-inducing auroral electrojets. Forecast meteorology, being older than space weather forecasting, has progressed further along the path of improving forecast services. We have identified ten general stages through which forecast meteorology has progressed: 1, societal need; 2, visual observations; 3, instrument observations; 4, synoptic imagery; 5, official forecast centres; 6, storm-structure definition; 7, subjective analysis; 8, objective analysis; 9, numerical predictions; and 10, storm tracking. Space weather forecasting has progressed through, or is progressing into, nine of these stages, missing only stage 9, numerical weather prediction.

Forecast meteorology has learned useful lessons, perhaps the most important of which is the ‘escalator effect’ (constantly improving skill scores) which happens once forecasts enter the stage of numerical predictions. Space weather forecasting is moving rapidly in the direction of numerical predictions. Meanwhile, objective forecast algorithms have assumed or are assuming a relatively larger role in space weather forecasting. Another important lesson from forecast meteorology is the benefit that would accrue to the profession were it to expose space physics students to actualities of operational space weather forecasting.

2.7 ACKNOWLEDGEMENTS

I owe information on the history and current status of SEC to communications with Patrick McIntosh, Barbara Poppe, Terry Onsager, and Howard Singer. This work is supported by the CISM project which is funded by the STC Program of the National Science Foundation under Agreement Number ATM-0120950.

2.8 REFERENCES

Allen, R. A., and E. M. Vernon, Objective Weather Forecasting, in *Compendium of Meteorology*, T. F. Malone, editor, American Meteorological Society, Boston, pp. 796–801, 1951.

Anderson, C. N., Solar disturbances in transatlantic radio transmission, *Proc. Inst. Radio Eng.*, **17**, 1528, 1929.

Anderson, C.W., L.J. Lanzerotti, and C.G. Maclennan, Outage of the L-4 system and the geomagnetic disturbances of August 4, 1972, *Bell Syst. Tech. J.*, **53**, 1817–1837, 1974.

Arge, C.N. and V.J. Pizzo, Improvements in the prediction of solar wind in the prediction of solar wind conditions using near-real time solar magnetic field updates, *J. Geophys. Res.*, **105**, 10 465–10 479, 2000.

Austin, L. W., Comparison between sun-spot numbers—intensity of Earth's magnetic field and strength of radio telegraphic signals, *J. Wash. Acad. Sci.*, **20**, 73, 1930.

Axe, G.A., The effects of the earth's magnetism on submarine cable, *Post Office Electrical Engineers Journal*, **61**(1), 37–43, 1968.

Barlow, W. H., On the spontaneous electrical currents observed in the wires of the electric telegraph, *Phil. Trans. Roy. Soc.*, **139**, 61–72, 1849.

Bartels, J., Terrestrial-magnetic activity and its relations to solar phenomena, *Terr. Mag. Atmos. Elec.*, **37**, 1–52, 1932.

Bartels, J., Solar radiation and geomagnetism, *Terr. Mag. Atmos. Elec.*, **45**, 339–343, 1940.

Birkeland, K., *Norwegian Aurora Polaris Expedition 1902–1903*, Vol 1, On the causes of magnetic storms and the origin of terrestrial magnetism, Christiania, H. Aschehoug & Co., 1908.

Bjerknes, J., and H. Solberg, Life cycle of cyclones and the polar front theory of atmospheric circulation, *Geofysiske Publikasjoner*, **3**(1), 3–18, 1923.

Bolduc, L., P. Langlois, D. Boteler, and R. Pirjola, A study of geomagnetic disturbances in Québec, 2. Detailed analysis of a large event, *IEEE Trans. Power Delivery*, **15**(1), 272–278, 2000.

Boteler, D. H., and G. J. van Beek, August 4, 1972 revisited: A new look at the geomagnetic disturbance that caused the L4 cable system outage, *Geophys. Res. Lett.*, **26**(5), 577–580, 1999.

Bothmer, V., and D. M. Rust, The field configuration of magnetic clouds and the solar cycle, in *Coronal Mass Ejections, Geophys. Monogr. Ser.*, **99**, edited by N. U. Crooker, J. A. Joselyn, and J. Feynman, pp. 137–146, AGU, Washington, D. C., 1997.

Broun, J. A., On the horizontal force of the earth's magnetism, *Proc. Roy. Soc. Edinburgh*, **22**, 511, 1861.

Cane, H. V., The large-scale structure of flare-associated interplanetary shocks, *J. Geophys. Res.*, **93**, 1–6, 1988.

Canfield, R. C., H. S. Hudson, D. E. McKenzie, Sigmoid morphology and eruptive solar activity, *Geophys. Res. Lett.*, **26**(6), 627–630, 1999.

Chapman, S., An outline of a theory of magnetic storms, *Proc. Roy. Soc. London*, **95**, 61–83, 1918.

Chapman, S., The electric current-systems of magnetic storms, *Terr. Mag. Atmos. Elec.*, **40**, 349–370, 1935.

Chapman, S., and J. Bartels, *Geomagnetism*, Oxford University Press, 1962.

Cliver, E. W., Solar activity and geomagnetic storms: The first 40 years, *EOS, Trans. AGU*, **75**, pp. 569, 574–575, 1994a.

Cliver, E. W., Solar activity and geomagnetic storms: The corpuscular hypothesis, *EOS, Trans. AGU*, **75**, pp. 609, 612–613, 1994b.

Cliver, E. W., Solar activity and geomagnetic storms: From M regions and flares to coronal holes and CMEs, *EOS, Trans. AGU*, **76**, pp. 75, 83, 1995.

Costello, Kirt A., Moving the Rice MSFM into a Real-Time Forecast Mode Using Solar Wind Driven Forecast Models, Ph.D. dissertation, Rice University, Houston, TX, June 1997.

Cressman, G. P., The origin and rise of numerical weather prediction, in *Historical Essays on Meteorology: 1919–1995*, J. R. Fleming, editor, American Meteorological Society, Boston, pp. 21–42, 1996.

Crooker, N. U., and E. W. Cliver, Postmodern view of M-regions, *J. Geophys. Res.*, **99**, 23383–23390, 1994.

FitzRoy, R., *The Weather Book*, Longman, Green, Longman, Roberts, & Green, London, second edition, 1863.

Fleming, J. A., The sun and the earth's magnetic field, *Annual Report of the Smithsonian Institution, 1942*, United States Government Printing Office, 173–208, 1943.

Fleming, J. R., *Meteorology in America, 1800–1870*, The Johns Hopkins University Press, Baltimore, 1990.

Freeman, J. W., *Storms in Space*, Cambridge University Press, Cambridge, 2001.

Garriott, E. B., *Weather Folk-Lore and Local Weather Signs, Bulletin No. 33.—W. B. No. 294*, U. S. Department of Agriculture, Weather Bureau, Government Printing Office, Washington, D. C., 1903.

Gopalswamy, N., A. Lara, S. Yashiro, M. L. Kaiser, and R. A. Howard, Predicting the 1-AU arrival times of coronal mass ejections, *J. Geophys. Res.*, **106**(A12), 29207–29218, 2001.

Hale, G. E., On the probable existence of a magnetic field in sun-spots, *Astrophys. J.*, **28**, 315–343, 1908.

Hale, G. E., Solar eruptions and their apparent terrestrial effects, *Astrophys. J.*, **73**, 379–412, 1931.

Hick, P. P., and B. V. Jackson, Heliospheric tomography: an algorithm for the reconstruction of the 3D solar wind from remote sensing observations, in *Telescopes and Instruments for Solar Astrophysics*, S. Fineschi and M. A. Gummin, editors, *Proceedings of the SPIE*, **5171**, pp. 287–297, 2004.

Hildner, E., Solar Physics and Space Weather: A Personal View, in *Solar Drivers of Interplanetary and Terrestrial Disturbances*, K. S. Balasubramaniam *et al.*, editors, Astronomical Society of the Pacific, San Francisco, pp. 619–626, 1996.

Hirman, J. W., SESC methods for short-term (3-day) flare forecasts, in *Solar–Terrestrial Predictions: Proceedings of a Workshop at Meudon, France, June 18–22, 1984*, edited by P. A. Simon *et al.*, published jointly by NOAA, Boulder, CO, and AFGL, Bedford, MA, pp. 384–389, 1986.

Hooke, R., *Micrographia*, 1665, reprinted by Science Heritage Limited, Lincolnwood, Illinois, 1986.

Horghton, D. M., Meteorology Education in the United States after 1945, in *Historical Essays on Meteorology*, J. R. Fleming, editor, American Meteorological Society, Boston, pp. 511–540, 1996.

Hufbauer K., *Exploring the Sun: Solar Science since Galileo*, The Johns Hopkins University Press, Baltimore, 1941.

Hughes, W. J., and M. K. Hudson, Editorial, *J. Atmos. Sol. Terr. Phys.*, **66**, 1241–1242, 2004.

Hulburt, E. O., in *Terrestrial Magnetism and Electricity*, J. A. Fleming, editor, McGraw-Hill Book Company, Inc., New York, 492–572, 1939.

Humboldt, Alexander von, *Cosmos, Vol. 1*, E. C. Otté translator, Henry G. Bohn, London, 1864.

Hundhausen, A. J., An Interplanetary View of Coronal Holes, in *Coronal Holes and High Speed Wind Streams*, J. B. Zirker, editor, Colorado Associated University Press, Boulder, 225–329, 1977.

Kiepenheuer, K. O., Solar Activity, in *The Sun*, G. P. Kuiper, editor, The University of Chicago Press, Chicago, 322–465, 1953.

Koelsch, W. A., From Geo- to Physical Science: Meteorology and the American University, in *Historical Essays on Meteorology*, J. R. Fleming, editor, American Meteorological Society, Boston, pp. 511–540, 1996.

Kunches, J., and C. Carpenter, A forecasters experience using Theo, an expert system, in *Solar-Terrestrial Predictions: Proceedings of a Workshop at Leura, Australia, 1989*, R. J. Thompson, *et al.*, editors, NOAA/AFGL, USA, pp. 482–486, 1990.

Loomis, E., On Certain Storms in Europe and America, December, 1836, Smithsonian Contributions to Knowledge, 1859.

Loomis, E., The great auroral exhibition of August 28th to September 4th, 1859 (2nd article), *Am. J. Sci.*, **79**, 92, 1860a.

Loomis, E., The great auroral exhibition of August 28th to September 4th, 1859 (4th article), *Am. J. Sci.*, **79**, 386, 1860b.

Loomis, E., The great auroral exhibition of August 28th to September 4th, 1859 (6th article), *Am. J. Sci.*, **80**, 339, 1860c.

Loomis, E., The great auroral exhibition of August 28th to September 4th, 1859 (7th article), *Am. J. Sci.*, **82**, 71, 1861.

Ludlam, F. H., and R. S. Scorer, *Further Outlook*, Allan Wingate, London, 1954.

Luhmann, J. G., *et al.*, Coupled model simulation of a Sun-to-Earth space weather event, *J. Atmos. Sol. Terr. Phys.*, **66**, 1243–1256, 2004.

McNish, A. G., The magnetic storm of March 24, 1940, *Terr. Mag. Atmos. Elec.*, **45**, 359–364, 1940.

McPherson, R. D., The National Centers for Environmental Prediction: Operational climate, ocean, and weather prediction for the 21st Century, *Bull Am. Met. Soc.*, **75**, 363, 1994.

Menzel, D. H., *Flying Saucers*, Harvard University Press, 1953.

Mitra, S. K., *The Upper Atmosphere*, The Royal Asiatic Society of Bengal, Calcutta, 1947.

Mulligan, T., C. T. Russell, and J. G. Luhmann, Solar cycle evolution of the structure of magnetic clouds in the inner heliosphere, *Geophys. Res. Lett.*, **25**, 2959–2963, 1998.

Newton, H. W., *The Face of the Sun*, Penguin Books, 1958.

Odenwald, S. F., *The 23rd Cycle: Learning to Live with a Stormy Star*, Columbia University Press, New York, 2001.

Petterssen, S., *Weather Analysis and Forecasting. A Textbook on Synoptic Meteorology*, McGraw-Hill, New York, 1940.

Plunkett, S. P., *et al.*, LASCO observations of an Earth-directed coronal mass ejection on May 12, 1997, *Geophys. Res. Lett.*, **25**(14), 2477–2480, 1998.

Sahu, J. K., and S. V. Singe, Single station forecasting of thunderstorm from large scale features, in *Proceedings of the Symposium on Advanced Technologies in Meteorology*, R. K. Gupta and S. J. Reddy editors, Indian Meteorological Society, M/s Tata McGraw-Hill Publishing Company Limited, New Delhi, 1999.

Shapley, A. H., The application of solar and geomagnetic data to short-term forecasts of ionospheric conditions, *Terr. Mag. Atmos. Elec.*, **51**, 247–266, 1946.

Shapley, A. H., and W. O. Roberts, the correlation of magnetic disturbances with intense emission regions of the solar corona, *Astrophys. J.*, **103**, 257–274, 1946.

Shaw, W. N., *Forecasting Weather*, D. Van Nostrand Company, New York, 1911.

Sheeley, N. R., Jr., J. W. Harvey, and W. C. Feldman, Coronal holes, solar wind streams, and recurrent geomagnetic disturbances, 1973–1976, *Sol. Phys.*, **49**, 271–278, 1976.

Silsbee, H. C., and E. H. Vestine, Geomagnetic bays, their frequency and current systems, *Terr. Mag. Atmos. Elec.*, **47**, 195–208, 1942.

Simon, P., and P. S. McIntosh, Survey of Current Solar Forecast Centers, in *Solar Activity Observations and Predictions*, The MIT Press, Cambridge, Massachusetts, 343–358, 1972.

Simon, P. A., G. Heckman, and M. A. Shea, editors of *Solar–Terrestrial Predictions: Proceedings of a Workshop at Meudon France*, 1984, published jointly by NOAA, Boulder, CO, and AFGL, Bedford, MA, 1986.

Siscoe, G., D. Baker, R. Weigel, J. Hughes, and H. Spence, Roles of Empirical Modeling within CISM, *J. Atmos. Sol. Terr. Phys.*, **66**, 1481–1489, 2004.

Spence, H. E., J. B. Blake, and J. F. Fennell, Surface charging analysis of high-inclination, high-altitude spacecraft: Identification and physics of the plasma source region, *IEEE Trans. on Nucl. Sci.*, **40**, 1521, 1993.

Spiers, I. H. B., A. G. H. Spiers, and F. Barry, *The Physical Treatises of Pascal*, Columbia University Press, New York, 1937.

Tascione, T. F., and A. J. Preble, Need for Space Weather Models, in *Solar Drivers of Interplanetary and Terrestrial Disturbances*, K. S. Balasubramaniam *et al.*, editors, Astronomical Society of the Pacific, San Francisco, pp. 16–25, 1996.

Temerin, M., and X. Li, A new model for the prediction of Dst on the basis of the solar wind, *J. Geophys. Res.*, **107**(A12), pp. SMP 31–1, CiteID 1472, DOI 10.1029/2001JA007532, 2002.

Theophrastus, Concerning Weather Signs, in *Enquiry into Plants II*, Loeb Classic Library, Harvard University Press, 1980.

Wells, H. W., Polar radio disturbances during magnetic bays, *Terr. Mag. Atmos. Elec.*, **52**, 315–320, 1947.

Winckler, J.R., L. Peterson, R. Hoffman, and R. Arnoldy, Auroral X-rays, cosmic rays, and related phenomena during the storm of Feb 10–11, 1958, *J. Geophys. Res.*, **64**, 597, 1959.

3

The Sun as the prime source of space weather

Volker Bothmer and Andrei Zhukov

The Sun is the prime source of energy in our solar system and it is the prime source of space weather. This chapter provides an overview on the main forms of solar energy output – fast and slow solar wind streams, co-rotating interaction regions, flares, coronal mass ejections and their interplanetary counterparts, solar energetic particle events – that determine space weather conditions in the interplanetary medium and in geospace and their variation with the solar activity cycle. The chapter also addresses the processes through which the energy transfer is modulated by solar, interplanetary and terrestrial conditions. The outlook of the chapter aims at defining the required observations that are crucial to help establishing real-time space weather forecasts.

3.1 INTRODUCTION – THE SUN'S ENERGY OUTPUT AND VARIABILITY

The Sun is the most powerful source of energy in our solar system and sustains life on Earth. It primarily emits energy in the form of electromagnetic (EM) radiation. The solar irradiance spectrum is shown in Figure 3.1. Since the full spectrum is made up of several components, it varies from an idealized black-body spectrum. The Sun's atmospheric layers overlying the visible disk – the photosphere, the chromosphere and corona (see Figure 3.4 in color section and Figure 3.7) – contribute to the spectr5m at EUV and X-ray wavelengths. These layers consist of a tenuous fully ionized plasma and hence cannot be treated as a black-body. Since the magnetic structure of the Sun's photosphere varies continuously (see Section 3.3.1) transient processes – such as micro- and nano-flares, shock waves, erupting prominences, flares and coronal mass ejections (CMEs) – cause short-time increases of EUV-, X-ray, gamma-ray and radio-wave emissions superimposed on the Sun's continuous spectrum (Figure 3.2). For a more detailed introduction to the physics of the solar spectrum the reader is referred to the introductions in the books by Aschwanden (2004) and Stix (2004).

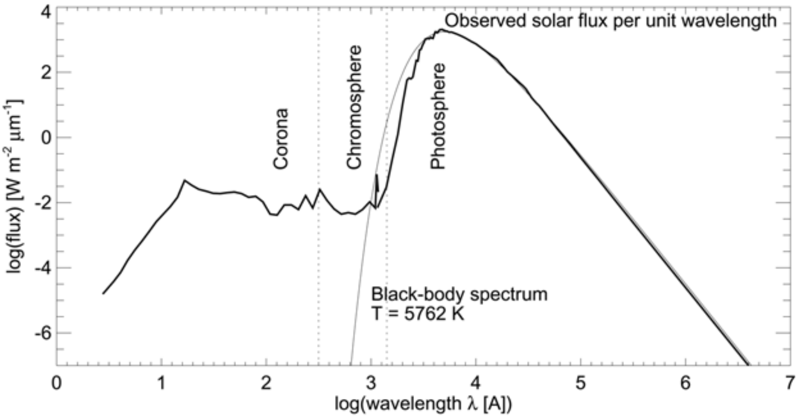


Figure 3.1. Measured spectrum of the solar flux and that of a black-body with $T = 5762\,\mathrm{K}$. From Aschwanden (2004).

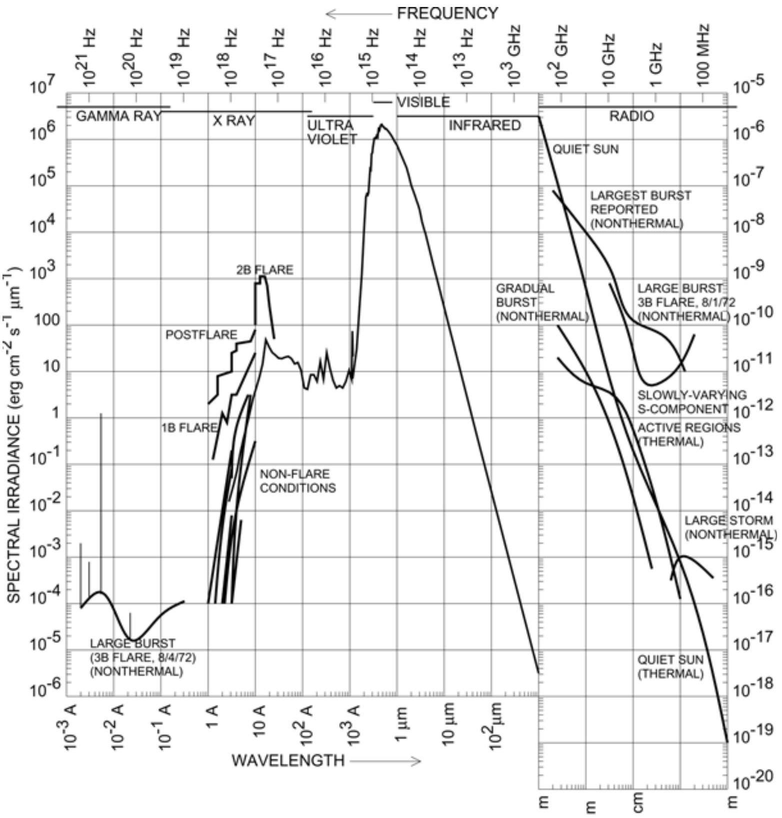


Figure 3.2. Full solar irradiance spectrum. Note the shift in the *y*-axis scaling by 12 orders of magnitude at a wavelength of 1 mm. From Aschwanden (2004).

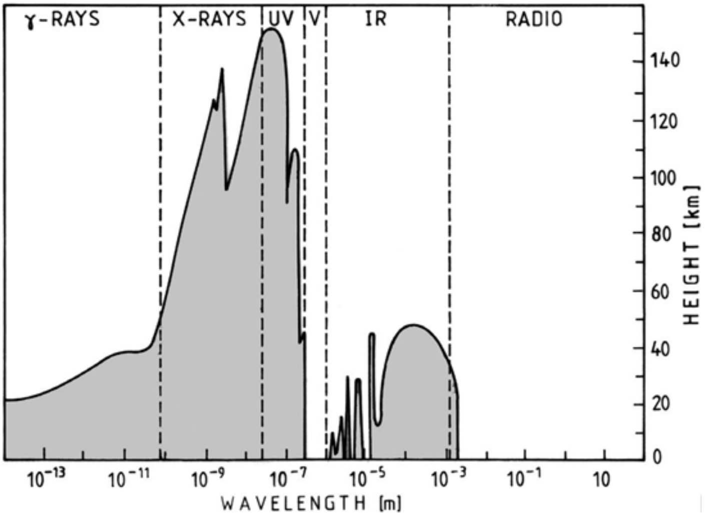


Figure 3.3. Absorption of solar radiation by the Earth's atmosphere. The shaded areas provide the height above ground where the incoming intensity is reduced to 50% of its original strength. After Nicolson (1982), adapted by Stix (2004).

On Earth – only at the visible wavelengths and part of the radio wavelength regime – the atmosphere is fully transparent to the Sun's radiation (see Figure 3.3) so that solar observations at X-ray and EUV wavelengths have to be achieved through space missions. The Yohkoh mission, launched on August 31, 1991 can be regarded as a milestone in terms of continuous high spatial resolution, full disk solar remote-sensing observations at X-ray wavelengths. Until the end of its mission life-time – on December 14, 2001 – it provided stunning new views of the Sun's X-ray corona (Figure 3.4, color section).

Although the Sun's total irradiance – the 'Solar Constant', being roughly 1367 W/m^2 as measured at the distance of the Earth – varies only at the order of 0.1% in the course of the solar cycle (e.g., Froehlich, 2003), the variation at specific wavelength intervals can be much larger (see Chapter 8). Since intensity variations at UV- and EUV-wavelengths may have important effects on the Earth's atmosphere, this subject is one of the hot current research topics. Figure 3.5 (color section) shows the variation of the longitudinal component of the photospheric magnetic field from solar activity maximum around 1992 until the next one around 2000, as measured by the US Kitt Peak National Solar Observatory (KPNSO), Tucson, Arizona, together with the measurements of the Sun's coronal soft X-ray emission, as measured by the Soft X-ray Telescope (SXT) onboard the Japanese/US Yohkoh satellite. The total variation of irradiance – in the range 2–30 Å – was roughly at the order of 10^2 between 1992 and 1996, but can in principle be much larger in case the solar photospheric magnetic fields are more frequent and intense, as can be expected, for example, from the highly varying number of sunspots in the different c. 11-yr long solar activity cycles (Figure 3.6). Benevolenskaya *et al.* (2002) found that the soft X-ray intensity shows

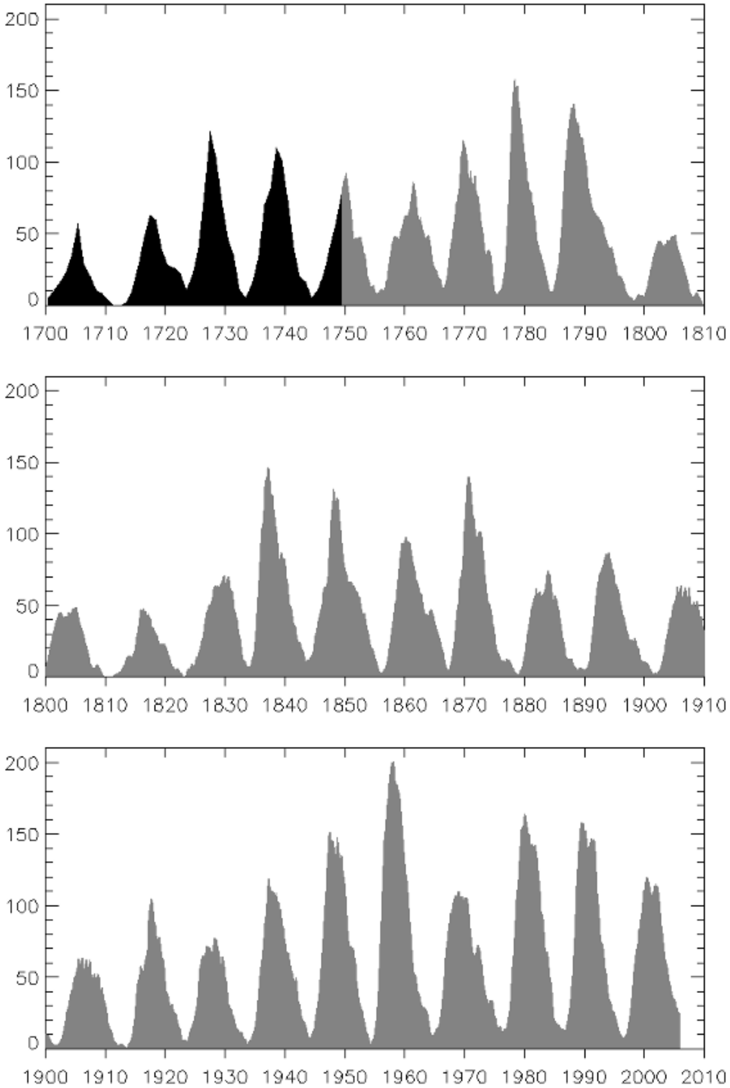


Figure 3.6. The yearly (black, up to 1750) and monthly (gray, from 1750 on) smoothed sunspot numbers. Courtesy: Solar Influences Data Analysis Center (SIDC), Brussels (*http://www.sidc.be*, July 1, 2006).

the following dependence on the longitudinal component of the photospheric field:

$$I_{SXR} \propto \langle |B_{\parallel}| \rangle^{n} \left\{ \begin{array}{l} n = 1.6\text{–}1.8, \text{ solar maximum} \\ n = 2.0\text{–}2.2, \text{ solar minimum} \end{array} \right\}$$

Solar cycles are counted from one solar minimum to the next – for example, the current cycle has the number 23 and runs from 1996 until the next minimum expected after 2006. At low solar activity the Sun's internal magnetic field may be treated to first order as a magnetic dipole. From about 1991 until 2000 the polarity of the Sun's magnetic field was positive – that is, the direction of the magnetic field lines was predominantly directed away from the Sun's photosphere – at its northern heliographic pole and negative – that is, the direction of the magnetic field lines was predominantly directed towards the photosphere – at the Sun's southern heliographic pole (compare with Figure 3.14, color section). A complete magnetic polarity reversal of the Sun's magnetic field hence takes about 22 years, commonly referred to as the 'Hale Cycle'.

The long-term behavior of the 11-yr solar cycle variations and that of even higher periodicities (e.g., the Gleissberg cycle of $\sim$90 years) is highly important in terms of space climate (e.g., Eddy, 1977), whereas for space weather, similarly to terrestrial weather, the momentary conditions at the Sun, in the interplanetary medium and at geospace are of prime importance – that is, forecasts require day by day services. Information derived from the long-term characteristics of individual cycles – like the obviously rapid rise and slow decay phases typically seen in individual cycles, as well as prediction of the strength of the next cycle – appear additionally as helpful means. Recently, Dikpati *et al.* (2006) have developed a flux transport model for the variation /f the Sun's magnetic flux in different cycles based on data from the Michelson Doppler Imager (MDI) onboard the ESA/NASA SoHO spacecraft, being in a halo orbit around the L1 point, 1.5 million km ahead of Earth in the sunward direction. According to the results of this model, solar cycle 24 will be 30–50% stronger in terms of activity, being somehow proportional to the evolution of the magnetic flux in the Sun's photosphere, compared with the current cycle and that activity will start rising at the end of 2007 or early 2008. So far, the prediction of solar cycle strength has unfortunately been relatively uncertain and the estimated peak sunspot number, R, for the coming cycle ranges from less than 50 to over 180 in the various model calculations (Badalyan *et al.*, 2001; Dikpati *et al.*, 2006).

The solar magnetic field structures the overlying solar corona and hence the global shape of the Sun's corona varies with respect to the phase of the solar activity cycle (Figure 3.7). The faint white light of the solar corona, being 10^6 times less in intensity compared with normal sunlight, is due to scattering of originally photospheric radiation by free electrons in the fully ionized hot outer atmosphere. The existence of the Sun's hot corona with temperatures of some million K above the cooler photosphere, $T = 5762\,\mathrm{K}$, is one of the unsolved astrophysical problems to date. During solar minimum the corona is roughly symmetric with respect to the solar equator and often reveals the presence of large coronal streamers that can be identified from their helmet-like appearances.

Besides EM radiation, the Sun continuously emits a flow of charged particles (plasma) from the corona, the solar wind. With speeds of several hundred km/s, it fills a region in interstellar space, the heliosphere, that occupies distances greatly exceeding 100 AU – that is, it extends out far beyond the distances of the furthest planets of our solar system. The properties of the solar wind and its evolution in the heliosphere control, together with flows of energetic particles in the eV (suprathermal) to MeV energy range, interplanetary space weather conditions. The supersonic solar wind flow

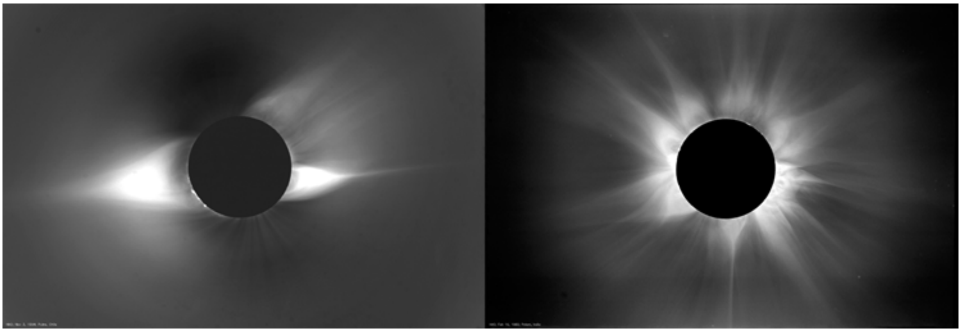


Figure 3.7. Left: the solar corona near solar minimum, as seen during the total solar eclipse on November 3, 1994. Right: the solar corona near solar maximum, observed during the total eclipse on February 16, 1980. Courtesy: High Altitude Observatory, Boulder, CO.

continuously impinges on the Earth's magnetic field. Superimposed on this quasi-steady flow of the solar wind and energetic particles are transient solar wind and particle flows in which the solar wind can blow against the Earth's magnetosphere with speeds of more than 2000 km/s and particle energies up to the GeV range as well as large solar flares, sporadic emissions of EM radiation, that can lead to short-term disturbances of the Earth's atmospheric conditions (see also Chapters 7 and 13). The different forms of energy released by the Sun are summarized in Tables 3.1 and 3.2. The following sections will provide a detailed overview of the solar/interplanetary origins of space weather, including slow and fast solar wind streams, stream interactions, coronal mass ejection events, flares and solar energetic particles.

Table 3.1. Power figures for different forms of solar %nergy output and mass flux estimates.

Total power input	
Solar radiation	$\sim 4 \times 10^{26}$ W
Solar wind from coronal hole	$\sim 4 \times 10^{20}$ W
Large coronal mass ejection	$\sim 1 \times 10^{23}$ W
Large solar flare	$\sim 1 \times 10^{23}$ W
Power at Earth	
Solar radiation	$0.137\ \mathrm{W/cm^2}$
Total on Earth	1.73×10^{17} W
Solar wind on disk with 1 R_E	$\sim 1.0 \times 10^{13}$ W
Solar mass loss	
Radiation	4.24×10^{12} g/s
Solar wind	$\sim 1.4 \times 10^{12}$ g/s
Coronal mass ejection	$\sim 1.0 \times 10^{12}$ g/s

Adapted from Schwenn (1988).

Table 3.2. The different forms of solar energy output.

1. Radiation

Spectral range	Source(s)	Characteristics
Radio mm IR	Quiet and disturbed corona, chromosphere	Electromagnetic radiation from moving charged particles (thermal radiation), transient radio emission caused by coronal shocks
White light	Photosphere Chromosphere K-Corona F-Corona	Continuum, thermal radiation Line emission and absorption Spectral lines from 6arious ions Continuum, photospheric light reflected from dust particles
UV	Chromosphere Transition region Corona	Spectral lines from various ions at various ionization stages
EUV	Corona Flares	See UV See X-rays
X-Rays	Upper corona 'Hot' corona, flares, etc.	Spectral lines as for UV, Bremsstrahlung Bremsstrahlung
γ-Rays	Strong flares	Bremsstrahlung and line emission from nuclear processes

2. Particles

Type	Source(s)	Characteristics
Solar wind (magnetized plasma)	Corona	H^+ up to 2 keV, electrons up to 1 keV
'Low energy' particles	Magnetic reconnection, flares, shocks	H, He, C, N, O, up to $\sim$100 keV
Energetic particles	Flares, shocks	Energies up to $\sim$100 MeV, sometimes up to $\sim$GeV

Adapted from Schwenn (1988).

3.2 SPACE WEATHER EFFECTS OF THE QUASI STEADY-STATE CORONA

3.2.1 Slow and fast solar wind streams and their source regions

Soon after the launch of the first satellites in the years around 1960, the existence of a continuous stream of charged particles, termed solar wind – as had been postulated by Biermann (1951) and Parker (1959) – could be verified through *in situ* measurements based on the invention of spaceflight techniques and instrumentation (Neugebauer and Snyder, 1966; Sonett and Abrahams, 1963). The observed solar wind characteristics showed a systematic two-stream pattern of slow and fast streams, with a 27-day recurrency interval suggestive that their solar source regions rotated with the Sun. For a detailed historical introduction of the early concepts of solar–terrestrial relationships and the discovery of the solar wind, the reader is referred to the article 'The solar atmosphere and space weather' (Bothmer, 2006).

Figure 3.8 (color section) shows observations of the Sun's corona taken at 195 Å, corresponding to temperatures of about 1.5 million K emitted mainly from Fe XII ions, as imaged by the imaging telescope (EIT, see Delaboudinière *et al.*, 1995) onboard SoHO, together with measurements of the geomagnetic activity index Ap (measured by 13 stations world-wide, see Section 3.2.2 and Chapter 7) and the solar wind speed as measured by the WIND satellite for the time period August 27 until September 10, 1996. The two time intervals with a fast and a slow solar wind stream, with speeds at around 400 and 600 km/s, are labeled.

Simultaneous measurements of the solar wind together with X-ray imaging of the solar corona have revealed – since the Skylab era in 1973 – that the sources of fast solar wind streams are 'coronal holes' (CHs) at the Sun. CHs appear as dark regions at the Sun in X-ray and EUV images because the magnetic field lines originating from these areas are rooted with only one end in the solar photosphere, contrar9 to active regions where heated plasma, radiating bright at X-ray and EUV wavelengths, is confined by closed magnetic loops rooted with both ends in underlying bipolar photospheric regions of opposite magnetic polarity. The passage of the fast solar wind stream encountered by the Earth is consistent with the appearance of the coronal hole extension at the central meridian. With a speed of about 600 km/s, it takes roughly 3 days for the solar wind stream to reach the Earth – that is, at times the solar wind has reached Earth's orbit, the solar wind source region, rotating with the Sun, is typically located to the west of the central meridian. In case the low-latitudinal extension of the coronal hole increases in heliolongitude, the time interval of the high-speed flow at 1 AU increases correspondingly. Such low-latitudinal extensions of coronal holes are a prominent feature during the declining phase of the solar activity cycle (e.g., Tsurutani *et al.*, 2006). They can last as quasi-stable co-rotating structures for many months as has been observed – for example, during the Skylab mission in 1974 (Bohlin and Sheeley, 1978).

Due to the Sun's rotation period of 25.4 days, being 27.23 days with respect to Earth, the outward-convected solar magnetic field imbedded in the solar wind gets structured into the pattern of an Archimedian spiral, also termed a Parker spiral, as shown schematically in Figure 3.9. The angle of the magnetic field direction in the ecliptic plane with respect to the Sun–Earth line depends on the solar wind speed, being roughly 45° for a flow speed of 400 km/s, with its polarity being directed outward (+ polarity) or inward (− polarity) depending on the magnetic polarity of the solar source region from which the solar wind stream originates. The angle of the magnetic field direction, φ, is calculated from the solar wind speed V_R at a given distance R with the rotation speed of the Sun Ω_S, as:

$$\varphi = \left(\left|\frac{\Omega_S R}{V_R}\right|\right)$$

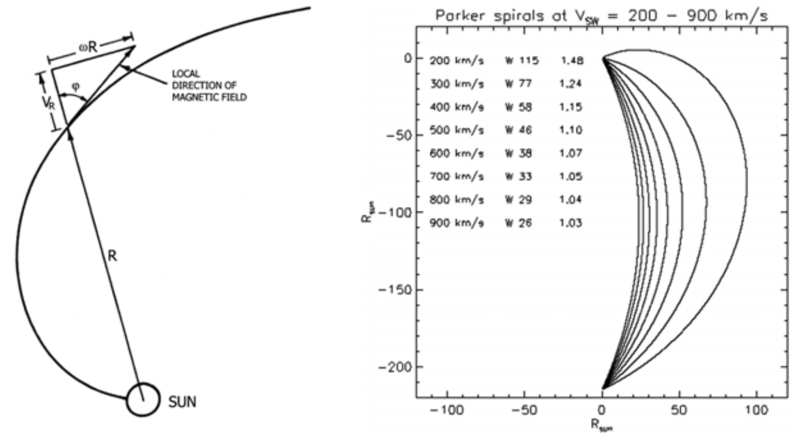


Figure 3.9. Left: schematic geometry of the interplanetary magnetic field (IMF) in the ecliptic plane for a solar wind with speed V_R at a distance R. The situation sketches the observed structure for a solar wind speed of ~400 km/s near 1 AU. The magnetic polarity of the IMF is assumed to be in the anti-sunward direction (i.e., positive). Right: curvature of the Parker spiral at the orbit of Earth for solar wind speeds between 200 and 900 km/s.

The magnetic field swept out by the expanding solar wind is termed the interplanetary magnetic field (IMF). Typical field strengths of the IMF at Earth's orbit are at the order of a few nT and are provided in Table 3.3, together with the average properties of the fast and slow solar wind near Earth's orbit.

As the right EUV image in Figure 3.8 shows, on September 6, 1996 the low-latitude extension of the coronal hole had disappeared behind the limb and the active region, seen to the southeast of the central meridian on August 27, had reached the Sun's west limb. At that time only slow solar wind was encountered at Earth's orbit and the geomagnetic activity index Ap was at low levels. It should be noted here that

Table 3.3. Basic solar wind characteristics near Earth's orbit.

Fast wind	Slow wind
450–800 km/s	<~450 km/s
$n_P \sim 3\,\text{cm}^{-3}$	$n_P \sim 7\text{–}10\,\text{cm}^{-3}$
~95% H, 5% He, minor ions and same number of electrons	~94% H, ~4% He, minor ions and same number of electrons – great variability
$T_P \sim 2 \times 10^5$ K	$T_P \sim 4 \times 10^4$ K
$B \sim 5$ nT	$B \sim 4$ nT
Alfvénic fluctuations	Density fluctuations
Origin in coronal holes	Origin 'above' coronal streamers and through small-scale transients

Adapted from Schwenn (1990).

the origin of the slow solar wind remains still a puzzle, but it is commonly believed to originate from the top of coronal streamers in a ‘drop-like’ manner (Wang and Sheeley, 1998). However, it may often well be that what is believed to be actually slow wind, such as that shown in Figure 3.8, had rather originated from near the edge of a polar coronal hole and expanded non-radially from the higher latitudes down into the ecliptic plane (Bothmer, 1998).

Around times of solar activity minimum, the structure of the heliosphere near the ecliptic plane is dominated by slow solar wind flows and the structure at higher latitudes by fast solar wind flows, as shown in the left part of Figure 3.10 (color section) based on results from the ESA Ulysses mission. Ulysses, launched in 1990, was the first spacecraft to explore the uncharted third dimension of the heliosphere. It orbits the Sun nearly normal to the ecliptic plane at distances between 1 AU and 5 AU (Balogh *et al.*, 2001; Marsden, 2001). The high-inclination orbit was achieved through a gravity assist maneuver at Jupiter. The measurements of the solar wind speed and IMF direction, taken during Ulysses’ first and second orbit around the Sun, are displayed in Figure 3.10. During the first orbit, at higher latitudes outside the ecliptic plane, only fast solar wind with speeds of about 750 km/s was encountered by Ulysses. Slow solar wind with speeds less than 450 km/s was confined to heliospheric regions in a narrow belt from about 20°N to 20°S with respect to the heliographic equator. The dominant magnetic polarity observed by Ulysses in both heliospheric hemispheres is consistent with the expected global magnetic polarity of the Sun during the rising phase of odd cycles: positive in the N, whilst negative in the S. Short-term intervals with faster streams seen at lower latitudes can be attributed to either transient streams (which will be described in detail in Sections 3.3.2 and 3.3.4), to streams from equatorward extensions of polar coronal holes (as shown in Figure 3.8) or to low-latitude short-lived coronal holes that occur near times of solar maximum (i.e., around times of magnetic polarity reversal of the Sun – see Figure 3.23 in color section). The relatively simple structure of the corona completely changes at times of solar maximum, as can be seen in the right top part of the solar image in Figure 3.10. The changing coronal structure is directly connected to variability of the solar magnetic field and the processes of its reversal.

3.2.2 Solar wind impact on the Earth’s magnetosphere

The physical origins of geomagnetic activity are induction currents caused by the solar wind’s electric field impacting the Earth’s magnetosphere (see Chapter 4). The solar wind plasma and IMF impose the electric field $\mathbf{E} = -\mathbf{v} \times \mathbf{B}$ on the Earth’s magnetosphere, causing a complex system of magnetospheric and ionospheric current systems (e.g., Burton, 1975; McPherron, 1979). Its prime components in both hemispheres are the ionospheric polar auroral electrojets and the near equatorial magnetospheric ring current measured through the AE (Auroral Electrojet) and Dst (disturbed storminess) indices (e.g., Bartels and Veldkamp, 1949; Mayaud, 1980; Rostoker, 1972; Siebert, 1971). The current systems are schematically shown in Figure 3.11 (color section).

The electrojet currents are primarily driven by electrons which spiral along magnetic field lines into the ionosphere where the current density **j** is given by $\mathbf{j} = \boldsymbol{\sigma}\mathbf{E}$, with $\boldsymbol{\sigma}$ being the conduction strength and **E** being the electric field that drives the current. The electric field **E** is of the order of several 10^{-3} V/m at times of weak energy input by the solar wind and reaches magnitudes of several 0.1 V/m during large storms (e.g., Schlegel, 2000).

The electrojets and the ring current are the major drivers of the fluctuations of the Earth's magnetic field. Comparable with the 'Richter–Skala' characterizing earthquake magnitudes, the overall geomagnetic activity index Kp ('Kennziffer Planetarisch') is a quasi-logarithmic parameter, given as daily 3-hr values (from 0 to 9), derived from magnetograms recorded by 13 stations primarily located in the northern hemisphere (Bartels and Veldkamp, 1949; Siebert, 1971). It is important to note that Kp is not only made up of the intensity of the electrojet current systems, but that it also depends on their spatial position – that is, large pressure pulses driven by solar wind flows, causing stronger compressions of the Earth's magnetosphere and subsequent movement of the electrojets to lower latitudes, lead to strongly enhanced Kp values, but will appear less pronounced in the ring current index Dst. Hence, a difficulty in an exact global determination of the absolute magnitude of a given magnetic storm is the superposition of compressional effects and current strengths. The Kp index and the magnetospheric ring current index Dst are both measured in units of nano Tesla (nT). The 3-hr ap-values and the daily Ap index represent the linear counterparts of Kp (a description of the conversion method between Kp and ap can be found at *http://www.gfz-potsdam.de*).

The subsolar point up to which the magnetosphere stretches out in space in the sunward direction can be calculated to first order from the pressure balance between the solar wind's dynamic pressure and the pressure provided by the Earth's magnetic field (e.g., Kivelson and Russell, 1995; Parks, 2003) as $2nmV^2 = B^2(r_{MP})/2\mu_0$, with n, V being the density and velocity of the solar wind, m being the proton mass and B being the strength of the Earth's magnetic field. Any pressure variation of the solar wind causes the magnetosphere to flutter in space like a flag – as shown, for example, by the simulations of Goodrich *et al.* (1998). Any variation of the solar wind electric field $\mathbf{E} = -\mathbf{v} \times \mathbf{B}$ leads in turn to fluctuations of the geomagnetic field, either caused through variation of the solar wind speed or its magnetic field magnitude. Note that, for simplicity, dynamic pressure variations due to solar wind density variations are neglected in this chapter in the subsequent considerations of the causes of geomagnetic activity. A detailed overview of solar wind–magnetosphere coupling functions can be found, for example, in Gonzalez (1990).

Solar wind speed is one of the drivers of geomagnetic activity, but it is well known that the energy transfer into the Earth's magnetosphere also depends crucially upon whether the IMF has a southward-directed (antiparallel) component with respect to the ecliptic plane (direction of the magnetospheric field). For an exact treatment, the IMF components must be considered in the GSM (Geocentric Solar Magnetospheric Coordinates) system (Russell and McPherron, 1973a). A systematic introduction to common coordinate systems was provided by Russell (1971). The main driver of geomagnetic activity is referred to, for short, simply as the southward $-B_z$ component

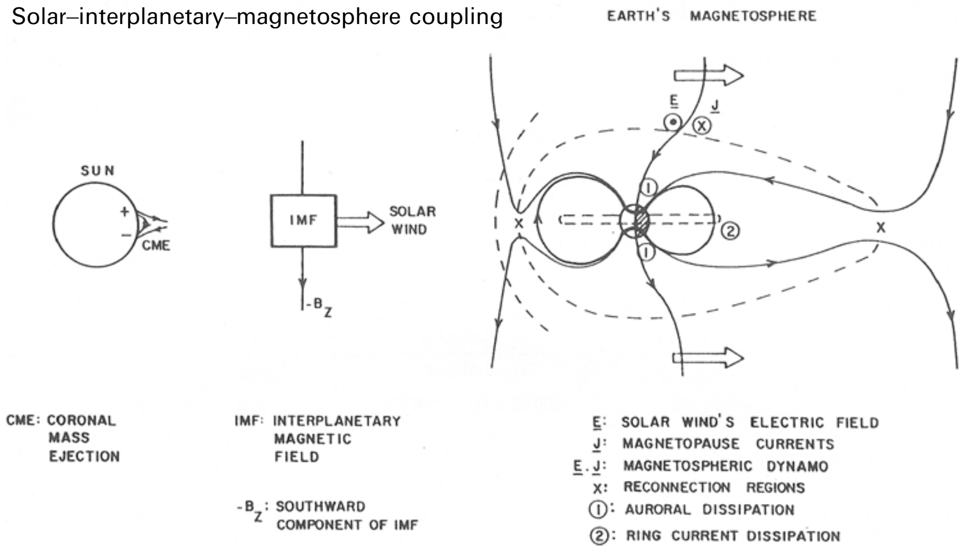


Figure 3.12. Schematic sketch of the magnetic reconnection process of the IMF with the Earth's magnetosphere and the energy injection process into the night-side magnetosphere. Note that the solar wind *per se* does not carry a substantial southward ($-B_z$) component, requiring specific processes – Alfvén waves, stream interactions, coronal mass ejections (CMEs). From Tsurutani and Gonzalez (1997).

of the IMF, leading to magnetic reconnection between interplanetary magnetic fields and magnetospheric fields, as depicted in Figure 3.12 (Tsurutani and Gonzalez, 1997). The reconnection process subsequently leads to plasma injection into the nightside magnetosphere. It should be noted that, should the Earth's magnetic field reverse, a northward component of the IMF would be favorable to reconnection processes.

Figure 3.13 shows the relationship between the interplanetary dawn–dusk electric fields for specific solar wind events (magnetic clouds, see Section 3.3.4) and the peak Dst values in individual geomagnetic storms (Tsurutani *et al.*, 2004). These cases, in which the solar wind speed was low (<400 km/s) – that is, pressure effects on the magnetosphere excited by the solar wind flow can be neglected – show a nearly linear relationship between the electric fields and the Dst values.

3.2.3 Space storms due to co-rotating interaction regions and high-speed flows

The structure of the inner heliosphere near solar minimum resembles the skirt of a 'ballerina', as proposed by Alfvén (1977) and sketched in Figure 3.14 (color section). At some solar radii distance from the Sun, the solar wind in both hemispheres is comprised of the two flows from the large polar coronal holes with opposite magnetic field polarity, yielding the so-called two-sector structure of the IMF. An observer (e.g., located at Earth) would pass through the upper and lower parts of the ballerina skirt during one solar rotation – that is, would record phases of opposite polarity of the IMF.

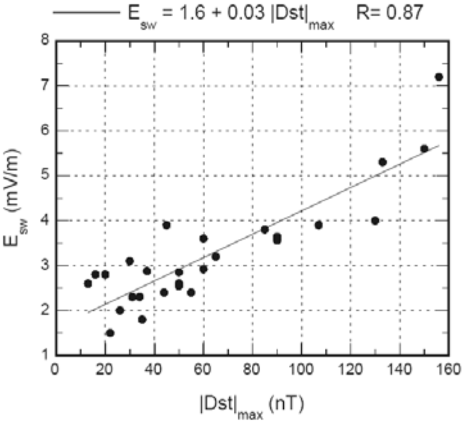


Figure 3.13. Relationship between the interplanetary electric field caused by the solar wind, E_{SW}, and the maximum Dst values of different geomagnetic storms. The individual storms were caused by slow transient solar wind streams with internal southward-directed magnetic fields (slow magnetic clouds). From Tsurutani *et al.* (2004).

A dominant feature of the heliosphere near solar minimum, especially in the declining phase of a solar cycle associated with the fast/slow stream solar wind pattern, is the long-lasting formation of co-rotating interaction regions (CIRs) followed by high speed (>600 km/s) solar wind streams. CIRs are caused by fast solar wind streams catching up slower solar wind streams ahead that had originated in solar longitude westward of the fast streams as viewed from Earth. In the stream interaction process, low-speed wind is compressed in its trailing edge and deflected in the sense of solar rotation, whilst high-speed wind is compressed in its leading portion and slightly deflected towards the opposite direction. Within CIRs the magnetic field magnitude is increased and the field vector may be deflected out of the ecliptic plane (e.g., Tsurutani *et al.*, 2006).

Observations of the corona taken simultaneously with solar wind data obtained by near-Earth satellites have revealed that recurrent – the period of solar rotation is 25.4 days and hence 27.3 days as seen from Earth – geomagnetic storms are caused by fast solar wind streams with speeds typically in the range of 500–800 km/s (Burlaga and Lepping, 1977; Crooker and Siscoe, 1986; Tsurutani, 2001). Contrary to Bartels' earlier belief that recurrent storms are due to magnetic active regions at the Sun, we now know that they stem from coronal holes – that is, rather magnetic 'quiescent' solar regions. Recurrent geomagn%tic storms are dominant especially in the declining phase of sunspot cycles (e.g., Richardson *et al.*, 2001) because at these times large polar coronal holes exhibit persistent low-latitude extensions over time periods of several months (e.g., Tsurutani *et al.*, 2006). Figure 3.15 shows the variation of the Kp index during 1974 organized in the classical 'Kp Musical Diagram' together with a yearly summary plot of the IMF B_z-component, Dst index, solar wind speed and IMF magnitude B. In the declining phase of cycle 21, high-speed solar wind streams from low-latitudinal extensions of the polar coronal holes persisted for months and led to recurrent patterns of enhanced geomagnetic activity.

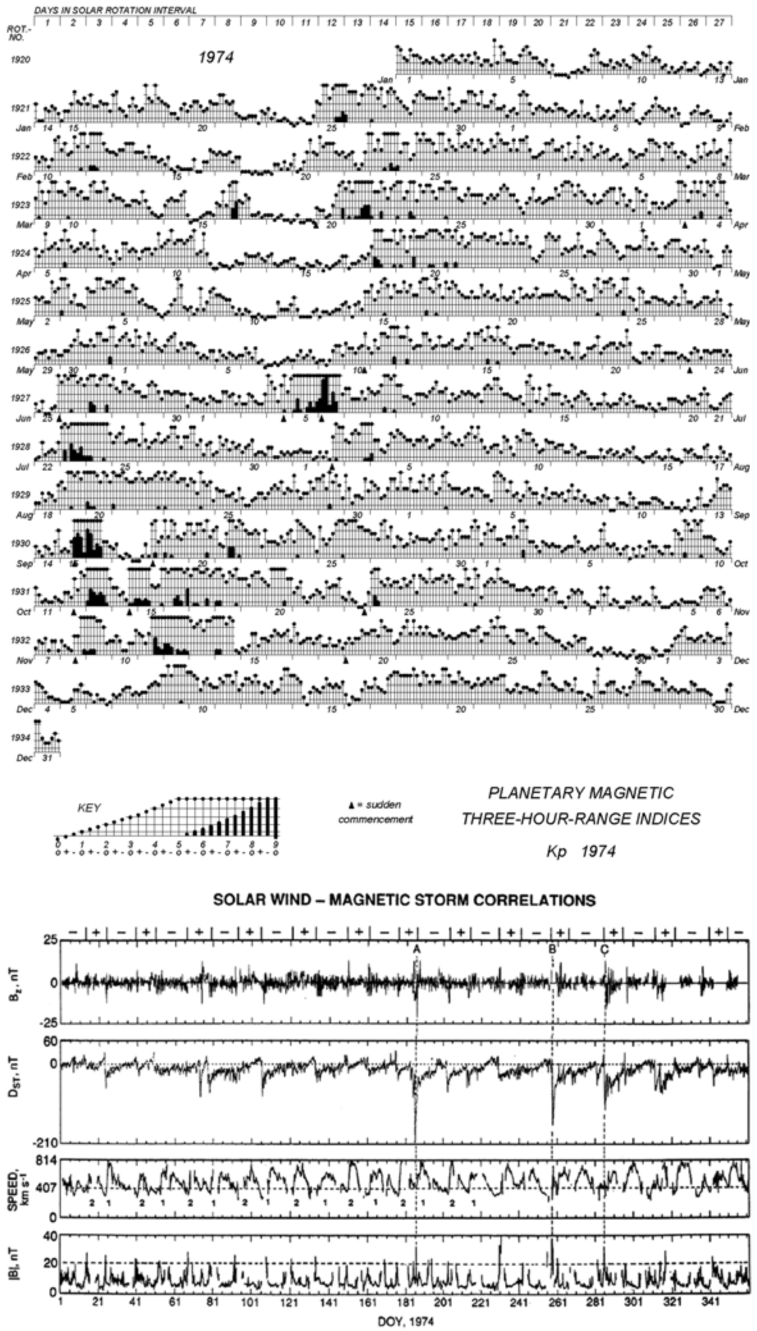


Figure 3.15. Top: the Kp 'musical diagram' for 1974. Bottom: southward component B_z of the IMF, Dst index, solar wind speed and IMF magnitude B in 1974. Labeled on top is the magnetic polarity of the IMF. Bottom diagram courtesy: Tsurutani (2006).

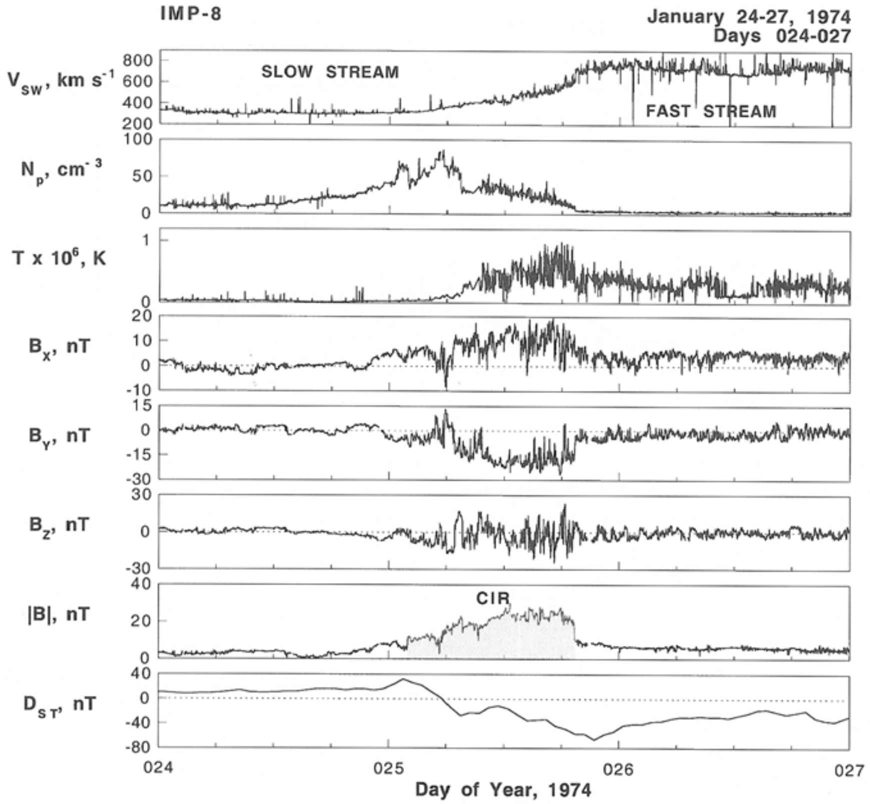


Figure 3.16. Solar wind measurements from IMP (Interplanetary Monitoring Platform) 8 of a co-rotating interaction region (CIR) formed between a slow and fast solar wind stream in January 1974 and the geomagnetic response as provided by the Dst index. Plotted from top to bottom are: solar wind speed, proton density and temperature, Cartesian components and magnitude of the IMF and Dst. From Tsurutani (2000).

Figure 3.16 shows an archetypal example of a geomagnetic storm that was triggered by a CIR and a subsequent high-speed solar wind stream observed in January 1974. Geomagnetic activity peaks within the CIR because of the compression and fluctuations of the IMF and Alfvénic waves (Tsurutani and Gonzalez, 1997). Following the CIR-related storm period, typically lasting for time intervals of less than a day, within the high-speed stream itself geomagnetic activity is triggered at lower levels by large amplitude IMF $-B_z$ fluctuations caused by Alfvénic waves (Tsurutani and Gonzalez, 1997). These fluctuations, as shown in Figure 3.17, stimulate prolonged substorm activities which lead to high-intensity long-duration continuous AE activity (HILDCAA) (Tsurutani and Gonzalez, 1987) in which energetic proton injections into the nightside magnetosphere up to about $L = 4$ (shell parameter of the McIlwain coordinate system, see Chapter 4) occur (Tsurutani *et al.*, 2006). Additionally to the protons, relativistic electrons at MeV energies, sometimes called ‘killer electrons’ because they can lead to surface charging and subsequent discharging

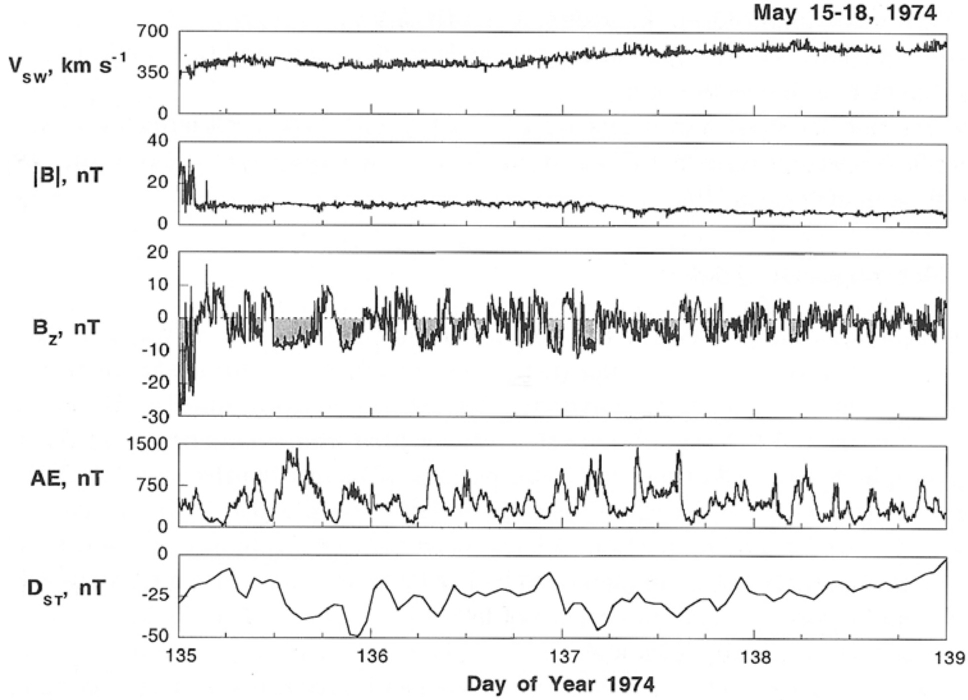


Figure 3.17. Example of a high-intensity long-duration continous AE activity (HILDCAA). From Tsurutani and Gonzalez (1997).

processes that may damage or disable satellite components, have been detected during intervals of high-speed streams (see also Chapter 6 and Baker, 2004).

CIR-related activity and subsequent wave activity are the reasons for the typically observed two-step behavior in recurrent storms (e.g.. Borello Filisetti *et al.*, 1988; Burlaga and Lepping, 1977). CIR-related shocks commonly form at distances around 2 AU from the Sun, though some have been observed by the Helios spacecraft at distances as close as 0.3 AU (Schwenn and Marsch, 1990), depending on whether the plasma gradients of the interacting streams exceed critical thresholds in terms of Alfvén and sound speeds. Charged particles are accelerated by CIR shocks and can stream along the magnetic field lines to heliospheric distances far away from the local acceleration sites, as has been frequently observed during the Ulysses mission (Lanzerotti and Sanderson, 2001). The physics of CIRs in the 3-D heliosphere has been explicitely summarized in reviews on these topics by Balogh and Bothmer *et al.* (1999), Crooker *et al.* (1999), and Gosling and Pizzo (1999). Depending on the spatial structure of the coronal holes, the tilt of the Sun's rotation axis and the structure of the global corona, systematic spatial patterns of compression regions or forward and reverse shock pairs may form (Gosling and Pizzo, 1999) and might have been the cause of systematic out-of-the-ecliptic deflection of the IMF in CIRs – like those reported by Rosenberg and Coleman (1980).

The strength and duration of an individual geomagnetic storm caused by a CIR and related high-speed solar wind stream can be quite variable, depending on the amount of compression of the IMF and the direction of the B_z component as well as its duration in case of a southward direction and, finally, the spatial size of the following high-speed flow (e.g., Richardson *et al.*, 2006). However, field intensities of the IMF at 1 AU within CIRs at 1 AU commonly do not exceed values of about 20 nT, and the variation of the solar wind speed compared with that of the slow solar wind is about a factor of 2. This is the prime reason geomagnetic storms caused by CIRs usually do not exceed Kp values of 7+, as has been inferred by Bothmer and Schwenn (1995) from detailed analysis of satellite data for 43 geomagnetic storms during the years 1960–1990 with peak Kp values of 8− and larger, in agreement with the results of Gosling (1993) and Richardson *et al.* (2001). As will be shown in Sections 3.3.4 and 3.4, all major space storms are caused by solar eruptions. The typical profiles of a CIR-related storm compared with that caused by a transient solar wind stream from a coronal mass ejection (CME, see Section 3.3.3) is shown in Figure 3.18.

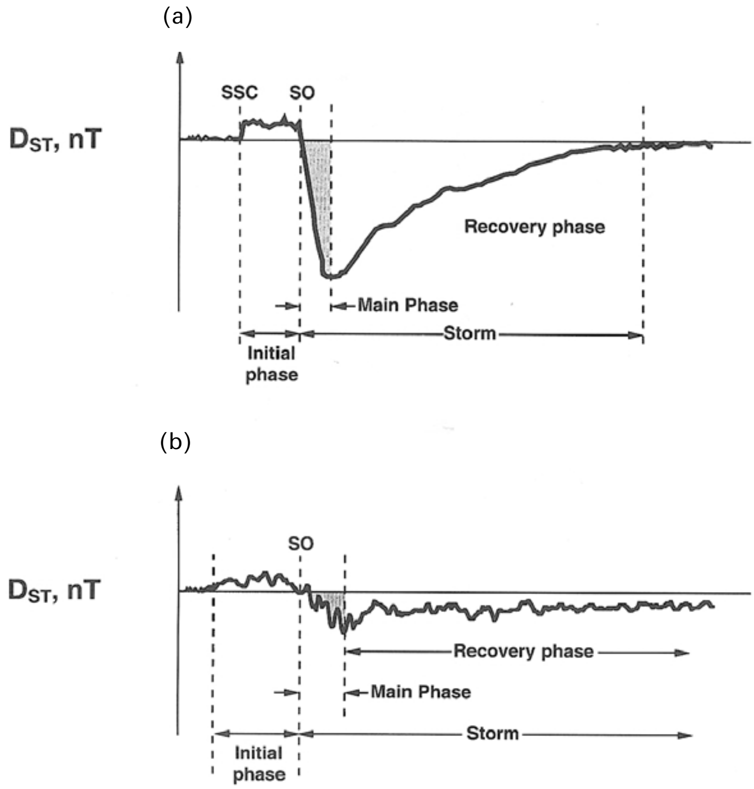


Figure 3.18. Typical Dst profiles for geomagnetic storms generated by an interplanetary coronal mass ejection (a) and a CIR/high-speed stream (b). SSC stands for sudden storm commencement, caused by a short-term compression of the magnetosphere through the transient flow. SO denotes the storm onset. From Tsurutani (2000).

3.3 SPACE WEATHER EFFECTS OF THE DYNAMIC CORONA

3.3.1 The ever changing photospheric magnetic field

Independently of the sunspot phenomenon, the photosphere of the Sun is always occupied by a magnetic field, as can be seen from comparison of SoHO/MDI/EIT (for an overview on the EIT and MDI instruments see Delaboudinière *et al.*, 1995, and Scherrer *et al.*, 1995) images in Figure 3.19 (color section) with GONG (Global Oscillation Network Group) white-light images taken from ground-based observatories. For further comparison, SoHO/EIT 195, 284, 304 Å images and a Catania H_α image have been added. Permanently small-scale changes take place in the photosphere where magnetic bipoles (e.g., Wang and Sheeley, 1989) of various spatial scales emerge, the smaller ones with low intensities replenishing themselves in time periods of about 40 hours are known to give rise to the black and white (salt and pepper) pattern in MDI magnetograms, the so-called magnetic carpet. Figure 3.20 shows how the small-scale magnetic field of the photospheric carpet connects the network on the spatial scale of supergranulation cells, while larger scale magnetic fields extend up into the corona (Aschwanden, 2004).

Figure 3.21 (color section), taken by the TRACE (Transition Region And Coronal Explorer) mission, shows the fine structure of the solar corona in unprece-

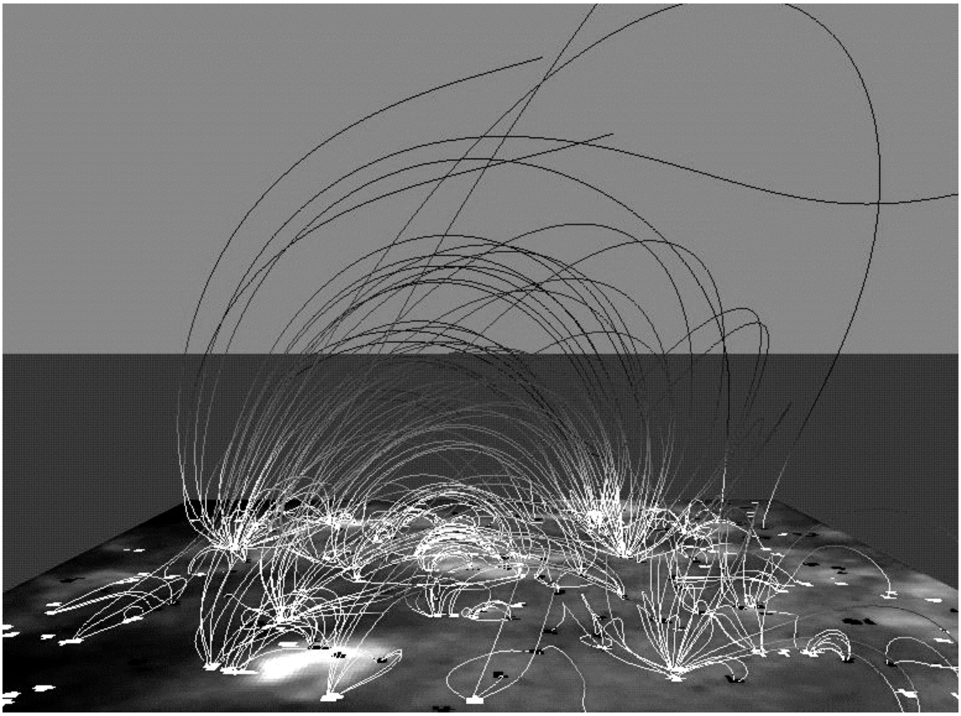


Figure 3.20. SoHO/MDI/EIT illustration of the magnetic carpet. Courtesy: SoHO/MDI/EIT Consortium.

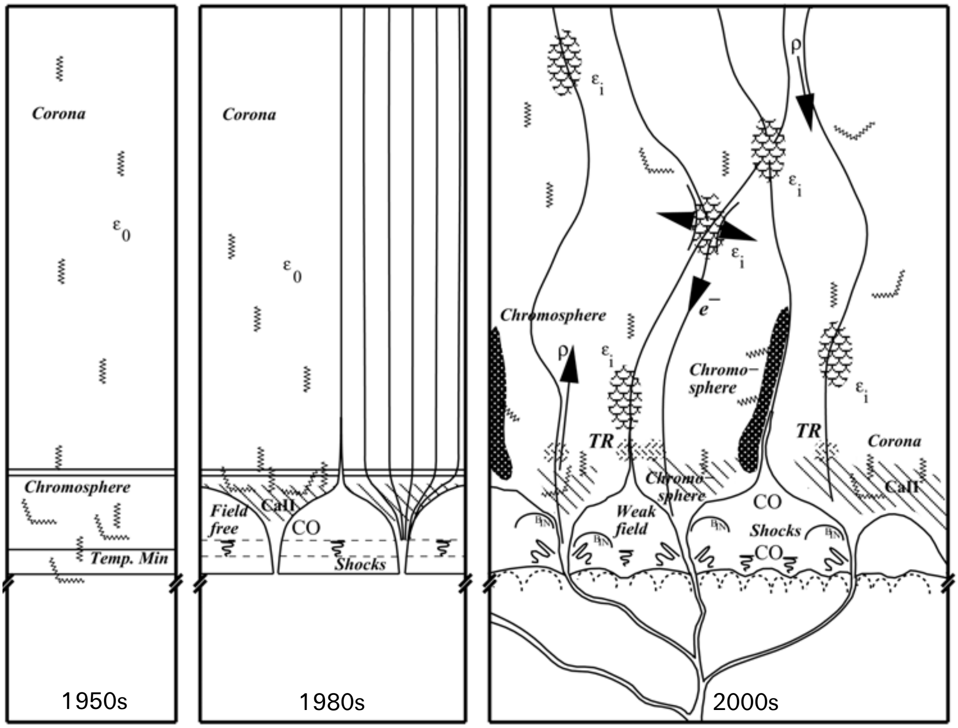


Figure 3.22. Left to right: changing physical concepts describing the structure of the solar corona from a gravitationally stratified atmosphere in the 1950s to an inhomogeneous turbulent profile today. From Aschwanden (2004).

dented spatial resolution. The image (with another color table on the right) of coronal loops over the eastern limb of the Sun was taken in the TRACE 171 Å pass band characteristic of a plasma at 1 MK temperature on November 6, 1999 at 02 : 30 UT. The image was rotated over 90 degrees in the clockwise direction. On the basis of new observations, Figure 3.22 shows how views of the physics of the solar corona have changed in time from a quasi-static, simple, gravitationally stratified solar atmosphere to a complex, highly time-variable system made up out of small-scale magnetic networks (Aschwanden, 2004). Although the Sun's atmospheric layers are ever changing on small scales, most of the time the interplanetary medium seems practically unaffected at the distance of Earth by the Sun's activity on small scales – associated, for example, with micro-flare activity – so that the main role for space weather effects is left to the solar wind from the open regions of the Sun's magnetic field (i.e., coronal hole flows).

With increasing solar activity, more and more magnetic bipoles, with the most intense areas in terms of magnetic flux seen as sunspots, occupy the solar photosphere, as can be seen in the SoHO/MDI images presented in Figure 3.5. Hence, the relatively simple structure of the solar corona described in terms of the ballerina model (Section 3.2.3) changes drastically, and the basic structure of the solar magnetic field and corona comprised of open fields at polar latitudes and closed fields distributed at lower latitudes vanishes. Figure 3.23 (color section) shows the changing structure of the Sun's EUV corona at 195 Å in three different images, taken by SoHO/EIT in 1996 near solar activity minimum, in the increasing phase of the solar cycle in 1998 and in 1999 close to solar activity maximum. At these times, more and more magnetic flux emerges into the photosphere and violent solar eruptions and flares, likely caused through magnetic reconnection processes, start dominating the daily 'solar weather'.

3.3.2 The explosive corona – coronal mass ejections and flares

The observation of a large white-light solar flare on September 1, 1859 by Richard Carrington (1860) and his subsequent conclusion that the flare might have been indicative of solar processes, triggering the major magnetic storm on Earth which occurred about 17 hours later, motivated scientists to try and establish the physical relationships between these two phenomena, without obtaining unambiguous results, today known as the 'solar flare myth' (Gosling, 1993a). The main reason for the long-undiscovered true physical links in solar–terrestrial physics is primarily related to the faintness of the solar corona, being 10^6 times less bright in intensity than the visible solar disk (i.e., the photosphere). Observations of the corona remained elusive until Bernhard Lyot (1939) invented the coronagraph, the first telescope able to detect the faint corona from Earth apart from total solar eclipses. A coronagraph essentially detects photospheric light scattered from free electrons in the hot outer solar atmosphere. This polarized light is also referred to as the Thomson-scattered light of the K-corona, with K denoting the German word *kontinuierlich*. The continuum corona is the prime ingredient of the white-light features visible in coronagraph images. A scientifically highly important feature of the solar corona is that the plasma-β is typically less than 1 – that is, the thermal pressure of the plasma is much smaller than its magnetic pressure and, hence, the ionized atoms and electrons are structured by the Sun's magnetic field.

In the early 1970s, for the first time coronagraphs were developed for space missions and successfully flown on the OSO (Orbiting Solar Observatory) 7 mission and some years later onboard Skylab, subsequently with the P78-1 and Solar Maximum Missions (SMM) and currently on SoHO (e.g., St. Cyr *et al.*, 2000). The first observations of the solar corona, at time cadences of several tens of minutes, recorded by spaceborne coronagraphs yielded a big surprise: The frequent appearance of large coronal 'bubbles', exceeding greatly the Sun's size at some solar radii distance, were propagating outward into space at speeds of several hundreds of km/s in the telescope's fields of view which were about 2–6 solar radii (Hildner *et al.*, 1976; Howard *et al.*, 1982, 1997; Koomen *et al.*, 1974; Sheeley *et al.*, 1985; St. Cyr *et al.*, 1999; Yashiro *et al.*, 2004). These large-scale coronal transients are today commonly referred to as coronal mass ejections or CMEs.

Figure 3.24 (color section) shows a typical, fast CME observed by SoHO on August 5, 1999. The speed of the CME was initially about 700 km/s, but in this case it was accelerated to about 1000 km/s during its outward motion up to distances of at least 10 solar radii, as derived from the measured velocity increase in the field of view

(~2–30 R_S) of the SoHO LASCO (Large Angle Spectrometric Coronagraph, see Brueckner *et al.*, 1995) instrument.

SoHO has so far recorded more than 10,000 CMEs (*http://cdaw.gsfc.nasa.gov/CME_list/*) with unprecedented resolution in space and time, allowing very detailed studies of their white-light structures, origins and kinematics (e.g., Chen *et al.*, 2000; Cremades and Bothmer, 2004). CMEs carry roughly 5×10^{12} to 5×10^{13} kg of solar matter into space (e.g., Howard *et al.*, 1997; Vourlidas *et al.*, 2002). Their speeds are often fairly constant over the first couple of solar radii, with the prime acceleration taking place commonly just within the first solar radii or less (e.g., St. Cyr *et al.*, 1999). However, some CMEs are accelerated sufficiently longer, as was the case in the sample event shown here.

The average speed of CMEs is on the order of 400 km/s, though some are substantially slower and others reach extremely high speeds, exceeding even 3000 km/s (e.g., Gopalswamy *et al.*, 2005). On average, the kinetic energy of a CME is around 10^{23} to 10^{24} J (e.g., Vourlidas *et al.*, 2002), which is comparable with the energy of large solar flares. Their angular widths are in the range of 24° to 72° (Yashiro *et al.*, 2004). During low solar activity, CMEs occur at low heliographic latitudes, but almost all around the Sun at times near solar maximum. Near solar minimum, the daily average CME rate is ~1, while it is ~4 near solar maximum (Yashiro *et al.*, 2004). Table 3.4 summarizes the basic characteristics of CMEs.

CMEs are best associated with eruptive prominences (disappearing filaments) – as shown by Webb and Hundhausen (1987) – and to a lesser extent with solar flares, though the individual phenomena may occur without each other (e.g., Subramanian and Dere, 2001). Gopalswamy *et al.* (2003) found more than 70% of the SoHO/LASCO CMEs during the years 1996–2002 to be associated with prominence eruptions. Flare-associated CMEs seem to be strongly connected to magnetic active regions, as evident from their brightness in low coronal EUV observations and from the enhanced underlying photospheric magnetic flux. This finding seems plausible, as one can easily imagine that stronger changing photospheric flux gives rise to coronal heating and flaring processes in active regions occurring at preferential lower heliographic latitudes in the course of the solar cycle, following the well-known butterfly pattern of sunspots (*http://science.msfc.nasa.gov/ssl/pad/solar/images/bfly.gif*).

In a recent study Zhang *et al.* (2001) analyzed in great detail – for a number of events – the temporal and physical relationship between coronal mass ejections and flares. In these cases the CMEs did slowly evolve through the fields of view of the

Table 3.4. Basic characteristics of CMEs.

Speed	<300–>3000 km/s
Mass	5×10^{12}–5×10^{13} kg
Kinetic energy	10^{23}–10^{24} J
Angular width	~24°–72°
Occurrence frequency	~1–~4 (sol. min.–sol. max.)

From Bothmer (2006).

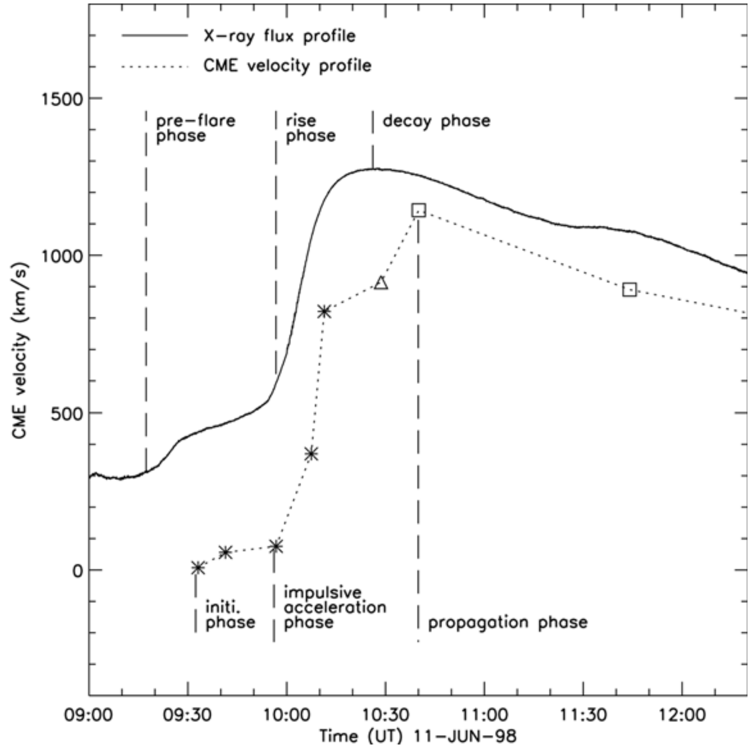


Figure 3.25. Speed–time profile for the CME on June 11, 1998 shown together with the flux profile of the associated X-ray flare. Note the three phases of CME acceleration and flare intensity evolution. From Zhang *et al.* (2001).

LASCO coronagraphs. According to their results, the kinematic evolution of flare-associated CMEs shows a three-phase development: an initiation, an impulsive acceleration and a propagation phase. In the initiation phase the CME slowly rises for a time period of several tens of minutes followed by the onset of the X-ray flare and the impulsive acceleration phase of the CME until, finally, the acceleration ceases and the CME starts propagating farther out at a constant speed, as shown in the velocity–time diagram in Figure 3.25 derived for the CME observed on June 11, 1998. Certainly, future high time-cadence solar observations will shed more light on the physical details of the onset of CMEs.

CMEs develop rapidly into large-scale objects with diameters bigger than the size of the Sun itself, as shown in Figure 3.26 (color section) from the study of structured CMEs performed by Cremades and Bothmer (2004). The typical three-part structure of the CME evident in Figure 3.26, consisting of a bright leading edge, a dark void and a bright trailing core, is evident already in the SoHO/EIT 195 Å images of the low corona (shown at the top). As Cremades and Bothmer (2004) pointed out, CMEs originate from magnetic loop/flux rope systems that likely already existed in the low corona at heights below about 1 solar radii and often expand in a self-similar manner into the field of view (~2–6 R_S) of the LASCO/C2 coronagraph. In the bottom right image of Figure 3.26 the identified source region of the CME is located in a composite SoHO/EIT/MDI image. The purple and blue colors denote regions of opposite magnetic field polarity in the photosphere.

Cremades and Bothmer (2004) found that bipolar regions in the photosphere are generally the underlying source regions of CMEs, independent of whether these were active regions or ones in which magnetic flux was already decaying and had persisted considerably longer in time. At higher latitudes the source regions of CMEs were typically more spatially extended and associated with prominences.

In a systematic study of the CMEs' source region properties, Cremades and Bothmer (2004) found that the 3-D topology of structured CMEs observed in the field of view of LASCO/C2 can be classified according to a basic scheme in which the fundamental parameters are the heliographic position and orientation of the source region's neutral line separating the opposite magnetic polarities. If one assumes that the average orientation of the neutral lines separating bipolar regions as CME sources follows Joy's law, the characteristic white-light shape of a CME seen in the FOV of a coronagraph can be explained naturally through the basic scheme presented in Figure 3.27. CMEs originating from the visible solar disk are seen at the east limb in cross-section and sideways at the west limb. The scheme reverses for CMEs originating at the back-side of the Sun, as viewed from the position of the observer assumed in Figure 3.27. Howard *et al.* (in press) have successfully reproduced the white-light

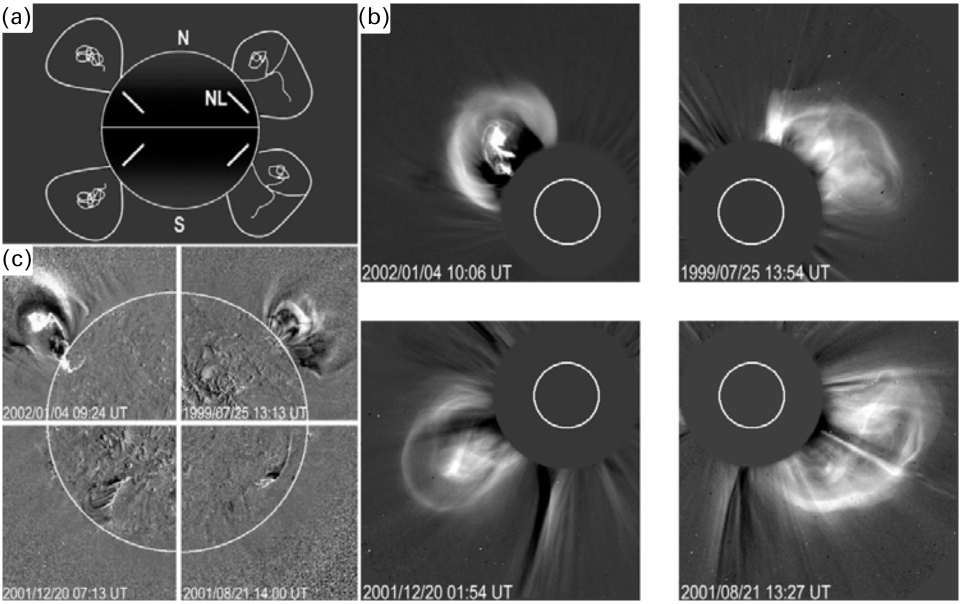


Figure 3.27. (a) Basic scheme showing the extreme cases of CME projection for front-side events. NL stands for neutral line (i.e., polarity inversion line separating the two opposite photospheric polarities). (b) Four projected CMEs seen by SoHO/LASCO C2 representing the scheme. (c) 195-Å signatures identifying the source regions of CMEs. For the northern events eruptive signatures were selected while for the southern ones post-eruptive features are shown. From Cremades and Bothmer (2004).

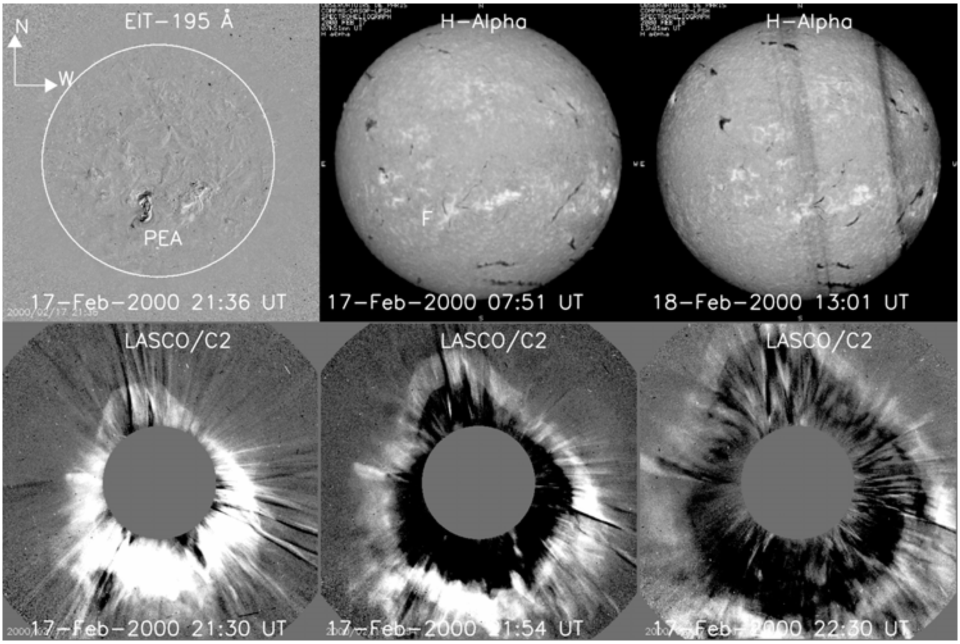


Figure 3.28. Top left: SoHO/EIT 195-Å image showing the post-eruptive arcade which formed after the front-side halo CME observed by LASCO/C2 on February 17, 2000. Middle and right images: Hα images from the Paris/Meudon Observatory showing the disappearance of the associated filament. Bottom images: SoHO/LASCO/C2 images showing the near-Sun development of the halo CME. The speed of the CME was about 600 km/s. Note the asymmetry of the halo in the NE to SW direction. From Tripathi *et al.* (2004).

pattern for the CMEs shown in Figure 3.27 through a graduated cylindrical shell (GCS) model, hence supporting the findings by Cremades and Bothmer (2004) on the 3-D structure of CMEs.

The apparent profile of an individual CME may differ more or less from the basic scheme presented in Figure 3.27 because of the solar variability of the fundamental underlying parameters – for example, many neutral lines are not straight but have rather complicated topologies, especially in active regions. The degree of correspondence with the scheme also depends on the absolute values of source region lengths, which will impose difficulties for small values typically found in compact active regions.

Contrary to the white-light structure of the CMEs shown in the scheme in Figure 3.27, events originating from near the center of the solar disk appear as unstructured halos (Howard *et al.*, 1982), as shown in the bottom sequence of images in Figure 3.28. The middle and right images at the top show the disappearance of a filament in H_α, the left image shows a post-eruptive EUV arcade that developed after the CME's onset in its low coronal source region (Tripathi, Bothmer and Cremades, 2004). Multi-wavelength observations of the CME's source region are shown in Figure 3.29 (color section) based on EUV 195 Å images from SoHO/EIT, soft X-ray observations from Yohkoh and H_α-images from the French Observatory at Paris/Meudon (*http://bass2000.obspm.fr/home.php*). The view is complemented by SoHO/MDI magnetograms.

From the SoHO observations it seems obvious that CMEs originate from localized spatial source regions in the two solar hemispheres. In the case presented here the front-side halo CME did several days later pass Earth's orbit, as identified from WIND and ACE (Advanced Composition Explorer) solar wind data (Bothmer, 2003; Yurchyshin, 2001). The calculated orientation of the CME's major axis, as inferred from white-light observations, was found to lie almost normal to the ecliptic plane, in agreement with the expected orientation of the filament in the CME's source region. In this case the magnetic field configuration was consistent with that expected from MDI observations (Bothmer, 2003).

The low-corona EUV signatures of CMEs on the solar disk can be used to discriminate whether they are front-sided or back-sided events (e.g., Tripathi, Bothmer and Cremades, 2004; Zhukov and Auchère, 2004). These features include 'EIT waves' and 'dimmings' (Figure 3.30). The coronal waves seen by EIT typically propagate at speeds of several hundreds of km/s, but are not seen for all CMEs (Klassen *et al.*, 2000; Thompson, 2000; Wang, 2000). According to Tripathi *et al.*, EUV post-eruptive arcades seen for a couple of hours after the onset of a CME are a definitive CME proxy, even without the availability of simultaneous coronagraph observations. Intensity brightening in soft X-rays near the onset time of CMEs (as seen in Yohkoh observations, often of sigmoidal structure), EUV dimmings and prominence eruptions are other good CME proxies (e.g., Canfield *et al.*, 1999).

It is important to summarize that solar flares emit short-term flashes of EM radiation over a wide spectral range which at the time of their observation cause effects on the Earth's ionosphere and atmosphere, as pointed out in Chapters 7 and 13, while CMEs are responsible for the convection of magnetized solar plasma into interplanetary space, with the fastest (>1000 km/s) CMEs typically causing the most intense interplanetary disturbances and, in case of the presence of a southward B_z at Earth's orbit, the strongest magnetic storms as well (e.g., Bothmer, 2004; Bothmer and Schwenn, 1995; Gosling, 1993; Tsurutani, 2001). Depending on the speed of the CME, the delay of the geomagnetic storm with respect to the solar event ranges from less than a day to several days (e.g., Brueckner *et al.*, 1998).

3.3.3 Interplanetary consequences of coronal mass ejections – shocks and ICMEs

The first *in situ* measurements of the solar wind already showed that, apart from slow and fast solar wind streams as described in Section 3.2.1, the interplanetary medium is frequently disrupted by transient flows, often associated with interplanetary shock waves discernible as strong discontinuities in which all plasma parameters (velocity, density, temperature) and the magnetic field intensity abruptly increase (e.g., Gosling *et al.*, 1968). Figure 3.31 shows an interplanetary shock wave detected by Helios 1 on May 13, 1981 (Sheeley *et al.*, 1985). The plasma speed abruptly increased from 600 km/s to over 1200 km/s.

Simultaneous operations of the Solwind coronagraph onboard the P78-1 satellite

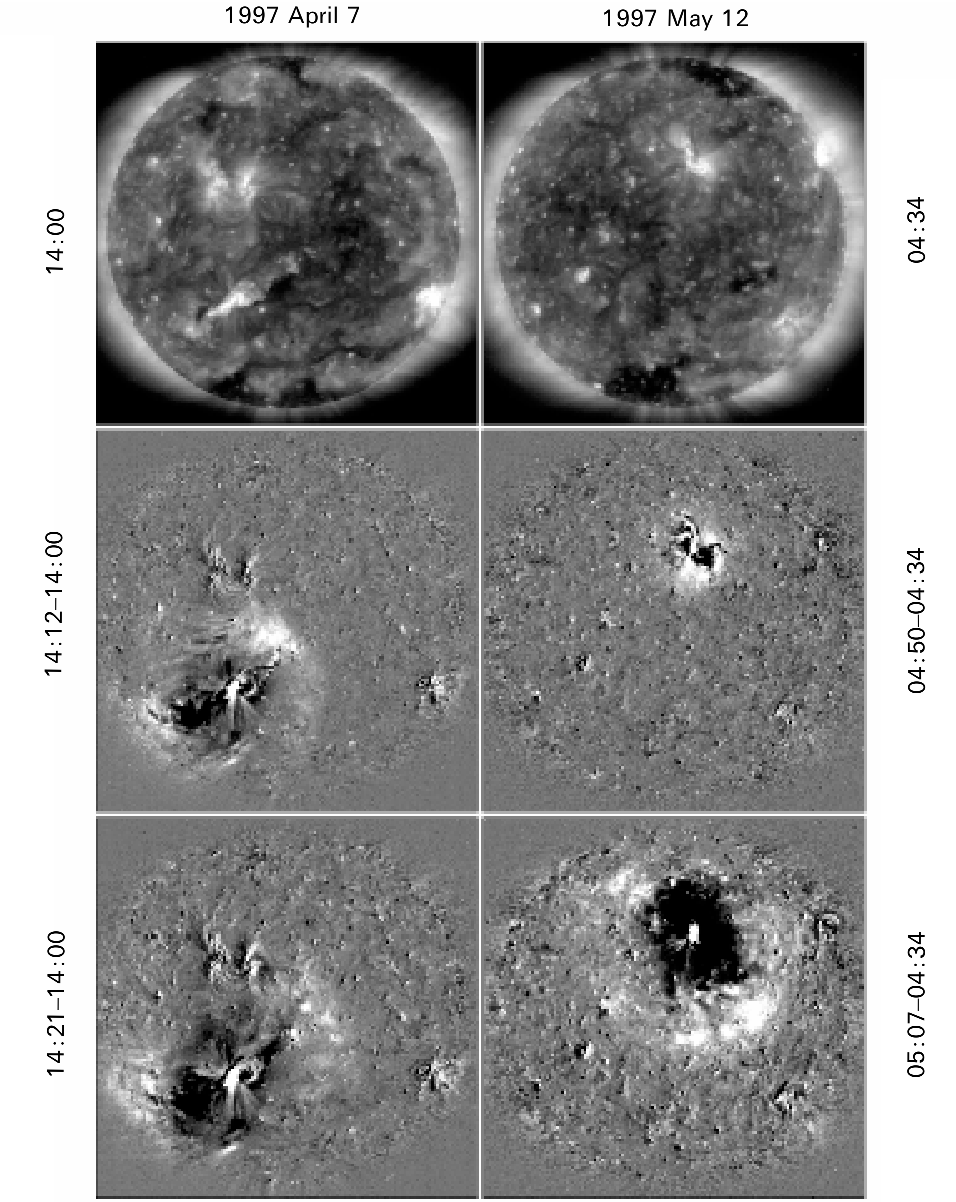


Figure 3.30. EIT waves imaged by SoHO/EIT at 195 Å in the solar corona associated with CMEs on April 7, 1997 and May 12, 1997. From Wang (2000).

and the German/US sun-orbiting spacecraft Helios 1, which explored the *in situ* characteristics of the inner heliosphere over the range 0.3–1 AU in the ecliptic plane, together with its sister spacecraft Helios 2, allowed for the first time during the years 1979–1982 to directly study the interplanetary effects of CMEs (Bothmer and

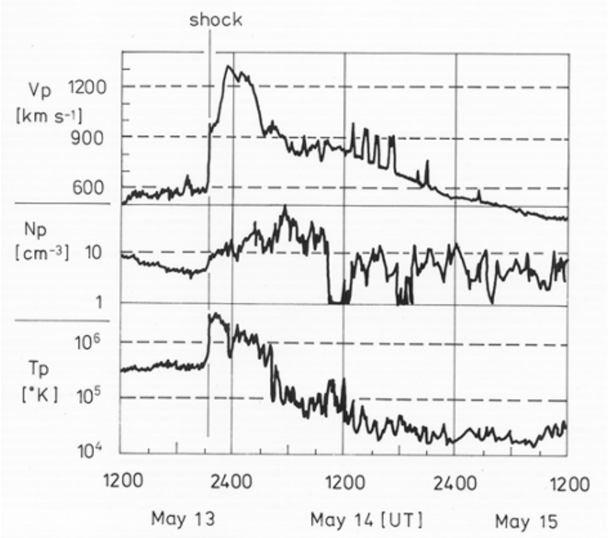


Figure 3.31. An interplanetary shock wave detected by Helios 1 on May 13, 1981. Solar wind parameters form top to bottom: proton bulk speed, density and temperature. From Sheeley *et al.* (1985).

Schwenn, 1996; Sheeley *et al.*, 1985). Sheeley *et al.* (1985) found that 72% of the interplanetary shock waves detected by Helios 1 were associated with large, low-latitude mass ejections on the nearby limb, with most of the associated CMEs having had speeds in excess of 500 km/s, some even having had speeds in excess of 1000 km/s. These observations clarified that CMEs are the sources of interplanetary shock waves and not solar flares, as was commonly believed until then (e.g., Chao and Lepping, 1974). Today it is well known that flares and CMEs can occur without each other, although the most intense events commonly occur jointly, and the concrete physical relationships between the two phenomena is a subject of ongoing research (e.g., Harrison, 1986; Zhang *et al.*, 2001).

Often an interplanetary shock wave was found to be followed several hours later by a transient solar wind stream with unusual plasma and magnetic field signatures, likely being the driver of the shock wave (Bothmer and Schwenn, 1996; Burlaga *et al.*, 1981). Systematic analyses of the wealth of satellite data obtained since the beginning of the space age has made it possible to establish reliable identification criteria of transient magnetized plasma flows as interplanetary consequences of CMEs (e.g., Gosling, 1990). To distinguish these CMEs in the solar wind from their solar counterparts, they are termed interplanetary coronal mass ejections or ICMEs (e.g. Cane and Richardson, 2003).

The classic plasma, magnetic field and suprathermal particle signatures of ICMEs at 1 AU are helium abundance enhancements, unusual ion and electron temperatures and ionization states (e.g., He^{+}, Fe^{16+}), higher than average magnetic field strengths ($>$10 nT), low variance of the magnetic field, smooth rotations of the magnetic field direction over time periods of several hours, bi-directional suprathermal ($>$40 eV)

Table 3.5. Basic characteristics of ICMEs at 1 AU.

Speed	300–>2000 km/s
Interplanetary shocks ahead of ICMEs	For CMEs with speeds >400 km/s
Magnetic field intensities	<10–>100 nT
Radial extension	0.25 AU ≈ 24 hours
Radial expansion with distance R from Sun	$\sim R^{+0.8}$ (R in AU)
Helical magnetic field structure	Magnetic cloud type ICMEs (1/3)
Plasma-β	<1 (especially in magnetic clouds)

From Bothmer (2006).

electron streaming, bi-directional suprathermal ion flows and plasma composition anomalies (e.g., Gosling, 1993 and references therein; Henke *et al.*, 1998, 2001). The detection of bi-directional electrons (BDEs) inside ICMEs suggests that the magnetic field lines may still be rooted in the Sun's photosphere at both ends (e.g., Bothmer *et al.*, 1996, 1997a; Gosling, 1993a). Table 3.5 summarizes the basic characteristics of ICMEs at 1 AU.

3.3.4 Examples of space storms driven by CMEs/ICMEs

Figure 3.32 shows measurements of the ICME that caused the major magnetic storm on July 15/16, 2000 (Bothmer, 2003; Lepping *et al.*, 2001). During this storm the Ap and Kp indices reached their maximum values. The plot in Figure 3.32 shows the time profile of the magnetic field magnitude B, the latitudinal angle θ (+90° corresponds to ecliptic north), the azimuthal angle φ (0° corresponds to the sunward direction, measured positive in the counterclockward direction), the plasma bulk velocity V, the proton density N_P and the thermal velocity of the protons V_{th}, as measured by the WIND spacecraft (courtesy D. Berdychevsky, from Lepping *et al.*, 2001). Data gaps were replaced by measurements from the GEOTAIL satellite (*http://www-spof.gsfc nasa.gov/istp/geotail/geotail.html*).

On July 15, 2000, the plasma parameters abruptly rose and the plasma speed increased to about 1100 km/s. This corresponds to the arrival of the interplanetary shock wave at 14:35 UT, labeled with a solid thick line in the figure. Several hours later on the same day at around 19:00 UT, it was followed by the ICME, as identified from the unusually high magnetic field strength and smooth southward to northward rotation of the magnetic field direction with respect to the ecliptic plane. The rotation started on July 15 at about 19:00 UT and lasted until July 16, 09:00 UT. The time period of the strongest magnetic disturbances measured at Earth corresponds with the time interval of the occurrences of strong southward magnetic field components caused by the ICME, even without taking details of the WIND orbit (i.e., the propagation time of the ICME from WIND to the magnetopause) into account. The magnetic storm starts with the arrival of the shock wave that triggers turbulent IMF fluctuations, followed by its main phase when the southward-directed magnetic field is encountered at the leading edge of the ICME. The magnetic field strength at the

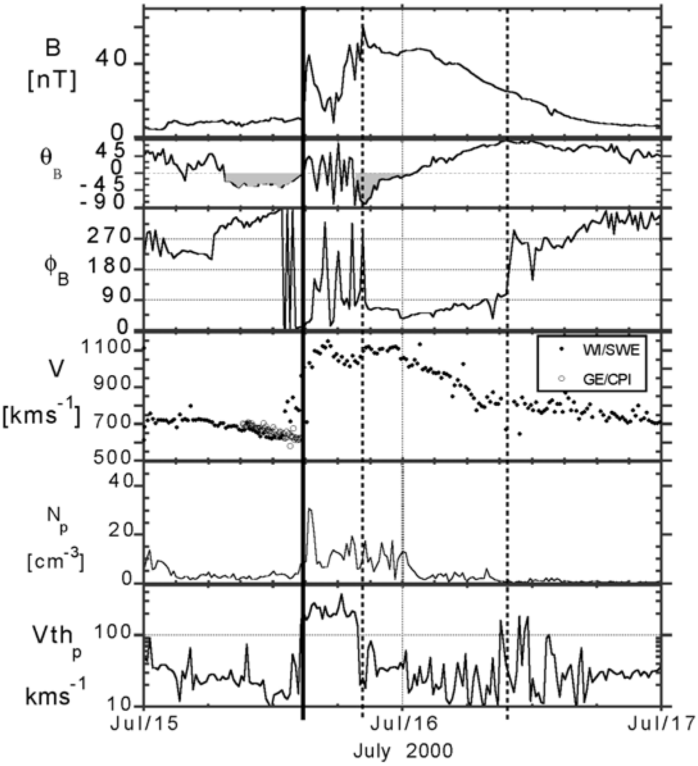


Figure 3.32. The ICME observed on July 15/16, 2000 by the WIND spacecraft. Data gaps are substituted with data from the Geotail satellite. The interplanetary shock ahead of the ICME is labeled by a solid line, dashed lines mark the boundaries of the ICME itself. Within the magnetic cloud type ICME, the magnetic field direction rotated from south to north, being directed eastward at its center. The ICME is of type SEN (i.e., it has left-handed magnetic helicity). Displayed solar wind parameters from top to bottom: magnetic field magnitude B, polar and azimuthal angles θ_B and φ_B, solar wind speed V, proton density N_P and thermal speed V_{th}. Courtesy: Berdychevsky, from Lepping *et al.* (2001).

nose of an ICME is typically considerably enhanced (e.g., Bothmer and Schwenn, 1998) if the ICME is fast and the speed gradient, with respect to the ambient solar wind, is large so that the interplanetary shock can be driven by the ICME far out into the heliosphere, even to distances of many tens of AU or its outer boundaries (Richardson *et al.*, 2006). The ICME's leading edge undergoes strong compression effects due to its interaction with the ambient slower moving plasma ahead. From the solar wind speed and duration of the ICME derived from the WIND measurements, the radial size of the ICME is estimated as ~0.3 AU. Although the spatial extent of the ICME is not known in the direction out of the ecliptic plane, it can be assumed from the large radial size that it is very large-scale in this direction too. Figure 3.33 sketches the propagation and deflections of the IMF by a fast ICME in the inner heliosphere, as viewed from a meridional perspective (Gosling, 1990). The deflections of the IMF in

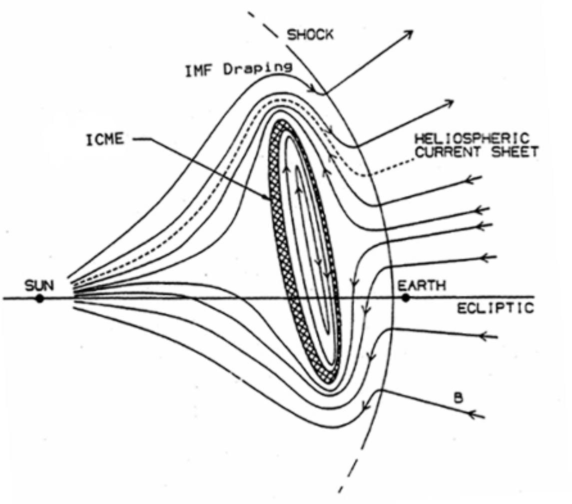


Figure 3.33. Idealized sketch of a fast ICME in the inner heliosphere viewed normal to the ecliptic plane. Note that – depending on the observer's position with respect to the center of the ICME and the Heliospheric Current Sheet (HCS) – different characteristic out-of-ecliptic variations of the IMF will be observed and that the shock and IMF deflections could also be observed even without hitting the ICME. The different signatures would lead to different signatures of geomagnetic storms or even not cause one. From Gosling (1990).

the region between the shock wave and the leading edge of the ICME, in which a high turbulent plasma regions forms, is termed 'draping of the IMF in the ICME sheath' (McComas *et al.*, 1988).

ICMEs that exhibit large-scale helical internal magnetic field configurations – like the one shown in Figure 3.32 – are termed magnetic clouds (Klein and Burlaga, 1982). In magnetic cloud type ICMEs the plasma-β is typically much smaller than 1, independent of the distance to the Sun in the inner heliosphere (Bothmer and Schwenn, 1998). According to Bothmer and Rust (1997) and Bothmer and Schwenn (1994, 1998), magnetic cloud type ICMEs can possess four different orientations of the magnetic field, as characterized by systematic rotations with respect to the ecliptic plane: SEN, SWN, NES, NWS, with N, S, E, W denoting the subsequent direction in which the magnetic field is directed at the leading edge of the ICME, its center and its trailing edge. These orientations characterize the helicity or chirality of a large-scale cylindrical magnetic flux tube in agreement with the self-consistent MHD model of a cylindrical magnetic flux tube shown in Figure 3.34 (Bothmer and Rust, 1997; Bothmer and Schwenn, 1998; Goldstein, 1983). If one points the fingers of the observer's hand towards the direction of the magnetic field in the leading edge of the ICME and the thumb along the magnetic field direction at the center of the ICME (the axis) and one further takes into account that the direction at the trailing edge is opposite to the direction at its leading edge, one can classify SEN- and NWS-type ICMEs as left-handed helical structures and SWN- and NES-type ICMEs as right-handed helical structures (Bothmer and Rust, 1997). According to this classification,

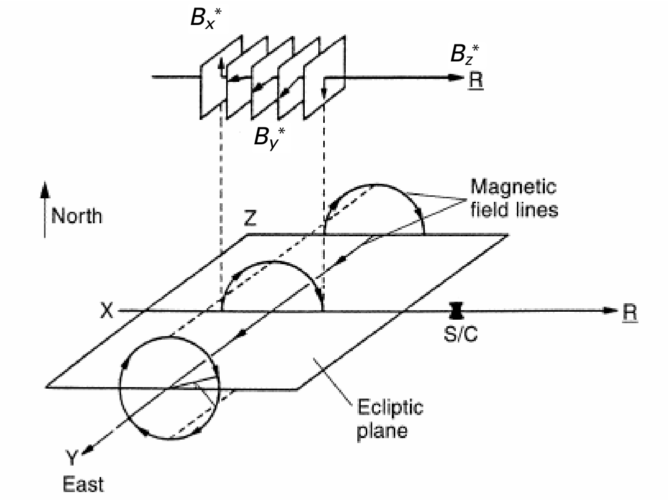


Figure 3.34. Idealized MHD model – a large-scale cylindrical flux tube – explaining the magnetic signatures observed during the passage of an ICME. From Bothmer and Schwenn (1998), after Goldstein (1983).

the ICME on July 15/16, 2000, is of type SEN – that is, it possesses left-handed magnetic chirality.

Magnetic cloud type ICMEs are of specific importance in the context of geomagnetic storms because their internal magnetic field structure can lead to large southward components of the IMF at 1 AU (e.g., Zhang and Burlaga, 1988). From a study of magnetic clouds in solar cycle 23, observed by the WIND and ACE satellites, Bothmer (2003) concluded that magnetic clouds trigger geomagnetic storms basically in two different ways or in combinations of these two ways: (1) through their specific internal magnetic field configuration, not only by SN or NS rotations, but also through the cloud's axis orientation when it is highly inclined and possessing a southward field direction (see also Mulligan *et al.*, 1998); and (2) through draping of the ambient IMF, especially in case of fast ICMEs driving shock waves ahead. The energy transfer from the solar wind to the magnetosphere is most efficient if ICMEs are associated with long-lasting (several hours) strong (<-10 nT) components of the magnetic field at 1 AU (e.g., Bothmer and Schwenn, 1995; Gonzalez and Tsurutani, 1987; Tsurutani, 2001) and activity is further amplified by ICMEs with high speeds (especially those >1000 km/s). It should be noted that in contrast – in case of northward IMF (i.e., $+B_z$-values associated with passage of an ICME) – a decrease in geomagnetic activity occurs (e.g., Veselovsky *et al.*, 2005).

ICMEs cause the highest V and lowest $-B_z$-values at Earth's orbit – the reason they are the drivers of all major geomagnetic storms with $\mathrm{Kp} > 7+$ (Bothmer and Schwenn, 1995). Bothmer (2004) analyzed – in the framework of a European Union project (INTAS 99-727) – the causes of all geomagnetic storms with disturbed days with Ap values >20 nT during the years 1997–2001, based on an unprecedented coverage of solar wind measurements provided by near-Earth satellites. The results

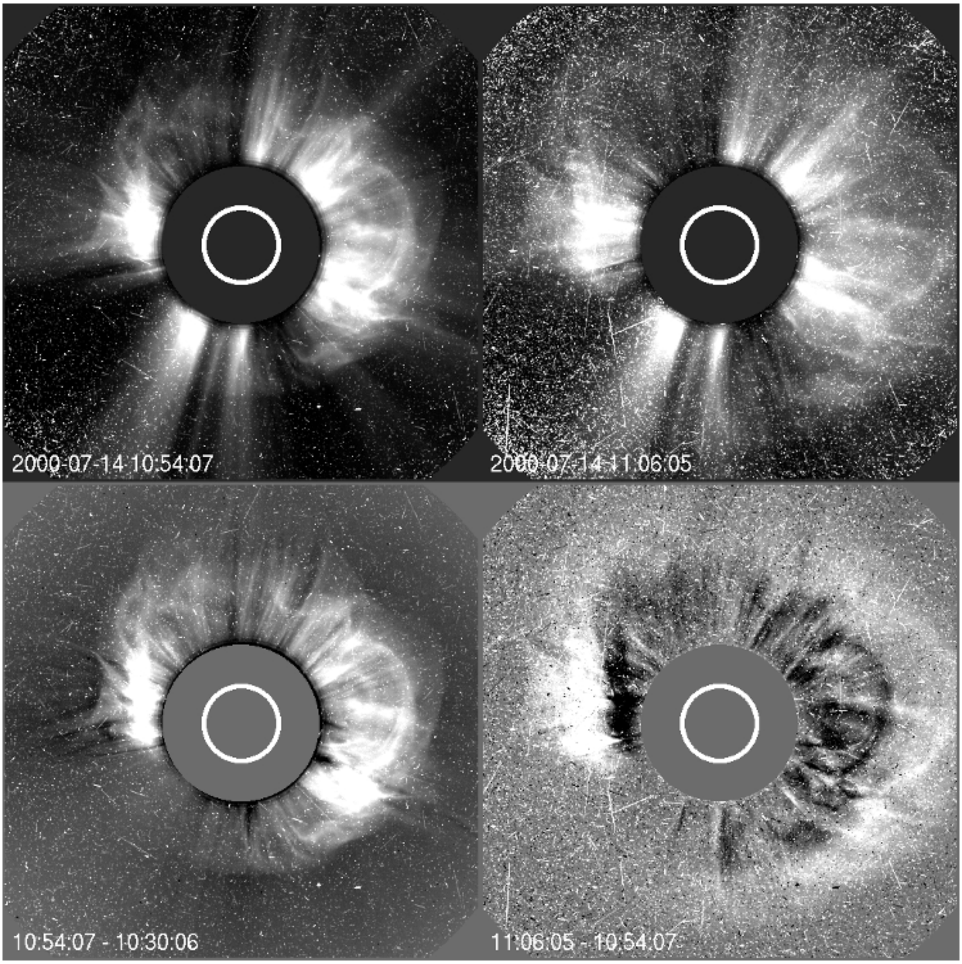


Figure 3.35. SoHO/LASCO C2 observations of the halo CME on July 14, 2000.

confirm that ICMEs are the prime drivers of major storms, as proposed by Bothmer and Schwenn (1996) in an earlier study of all storms with Kp values greater than 8−, and also that super-intense storms are often triggered by multiple interacting ICMEs.

Fast ICMEs driving shock waves from close to the Sun out into the heliosphere are also capable of accelerating charged particles to energies up to MeV or even GeV (e.g., Reames, 1999). Figures 3.35, 3.36 and 3.37 (color section) show the CME originating at the Sun on July 14, 2000, that caused the major geomagnetic storm on July 15/16, the EUV post-eruptive arcade that formed in this source region and a mosaic of solar observations including MeV electron and proton measurements from SoHO/COSTEP. COSTEP is the COmprehensive SupraThermal and Energetic Particle analyzer (e.g., Bothmer *et al.*, 1997b). The time delay at 1 AU to the CME onset time at the Sun is typically less than 15 minutes for the electrons and less than about 30

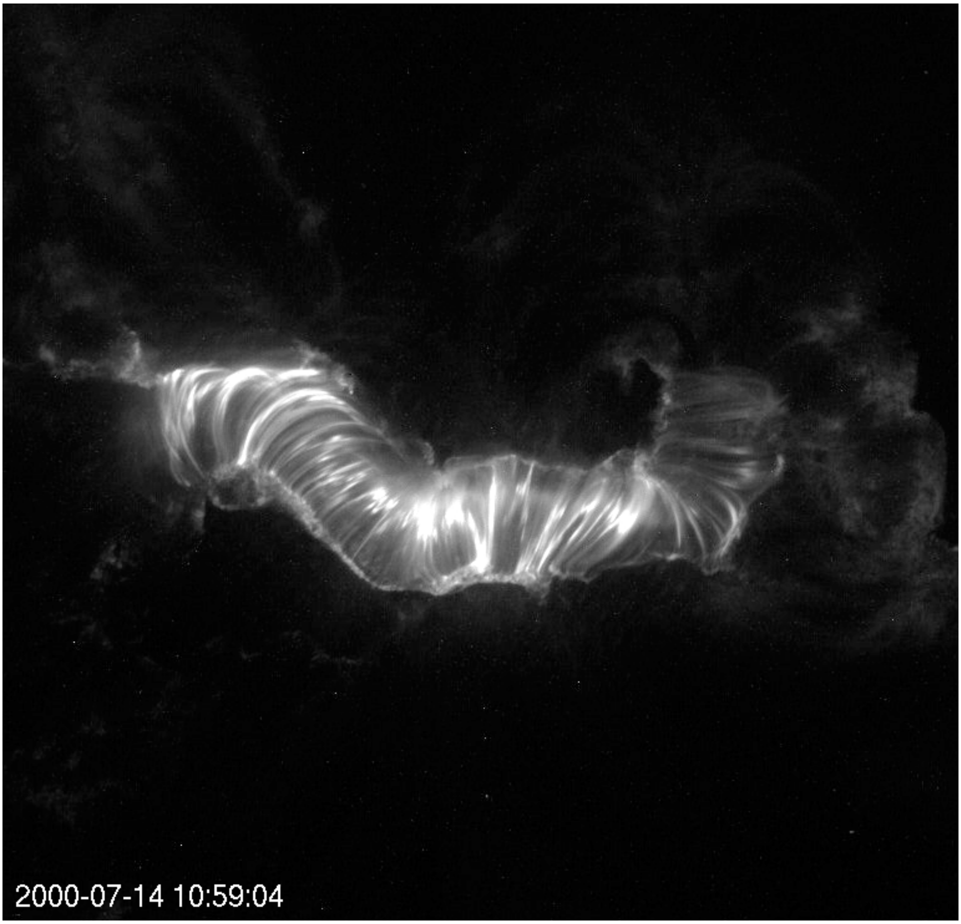


Figure 3.36. TRACE observations at 195 Å of the post-eruptive arcade in the CME's solar source region on July 14, 2000.

minutes for the protons at MeV energies for a prompt event. Protons of energies from 50–100 MeV from such solar energetic particle (SEP) events cause 'particle snowstorms' in the images from the SoHO optical telescopes when they pass through the CCDs, as can be seen in the LASCO/C2 and C3 images in Figure 3.37 (color section). They can degrade spacecraft hardware components, as can be seen in the plot of the efficiency of the SoHO solar panels in Figure 3.38 from their power decrease after the July 2000 event and after other major particle events (Brekke *et al.*, 2006). How high-energy protons affect electronic chips in the form of single-event upsets (SEUs) is schematically shown in Figure 3.39.

The largest geomagnetic storms in solar cycle 23, as measured by the peak Kp values (see also Table 3.8 in Section 3.5), were caused by two superfast (>2000 km/s, see *http://cdaw.gsfc.nasa.gov/CME_list/*) CMEs observed by SoHO/LASCO on

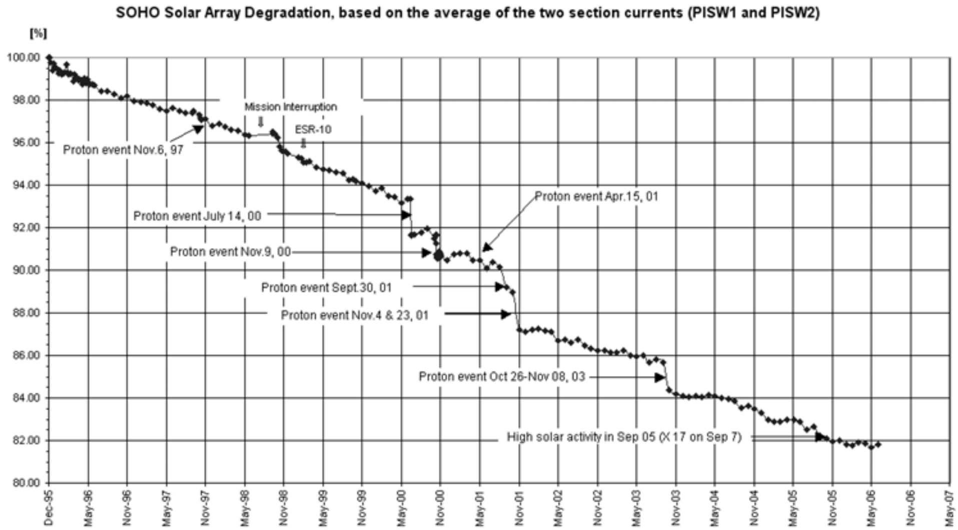


Figure 3.38. Effects of the July 14, 2000 solar energetic particle event and other major particle events on the solar panels of the SoHO spacecraft. Courtesy: Brekke *et al.* (2006).

October 28 and 29, 2003 (Figure 3.40, color section). It is worth pointing out that such high-speed CMEs are rare. Out of the more than 10,000 events listed in the SoHO/LASCO CME catalog for the years 1996–2006, only 36 had speeds in excess of 2000 km/s and just 25 of them reached speeds greater than 2500 km/s. Such superfast CMEs play a major role in terms of intense solar energetic particle events, as will be described in Section 3.3.5.

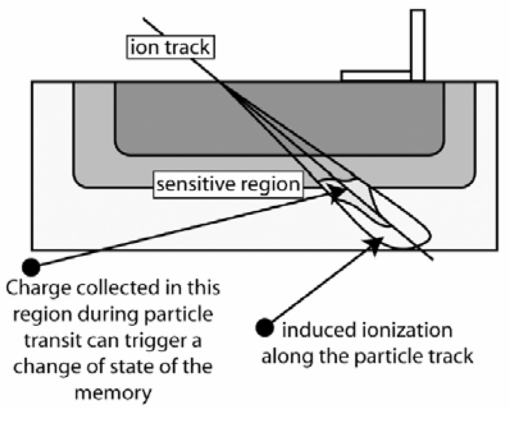


Figure 3.39. Example for ion interactions causing single-event upsets (SEUs). From Baker (2004), adapted from Robinson (1989).

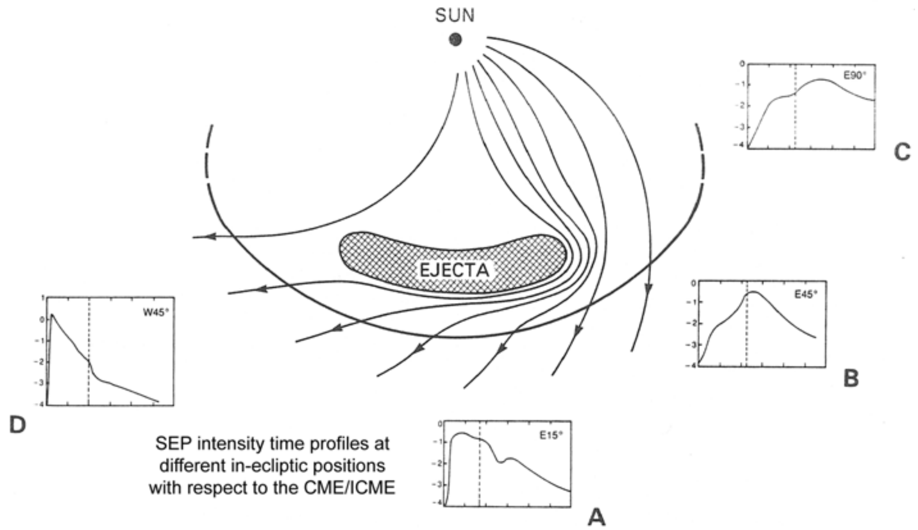
3.3.5 Major SEP events, CME-driven shocks and radio-wave signatures

Intense solar energetic particle (SEP) events represent a serious threat to manned spaceflights to the Moon and Mars, during extravehicular activities on the International Space Station, and to air crews and passengers (see Chapters 5 and 11). Fortunately, commonly only a couple of times during a solar cycle do very intense SEP events with proton fluxes exceeding 10^{10} protons/cm^2 occur (ANSER, 1996), but unfortunately at the present time their origin is poorly understood and their forecast constitutes a big challenge to modern research.

We describe here two SEP events that occurred on October 28 and 29, 2003 and on January 20, 2005 (Figures 3.41a and 3.41b, color section), which were among the strongest in solar cycle 23. The close association between fluxes of energetic (at MeV energies) electrons and protons at 1 AU following CMEs at the Sun is well known (e.g., Bothmer *et al.*, 1997). It is now commonly assumed that MeV particles are primarily caused by shock acceleration mechanisms (e.g., Reames, 1999; Tylka, 2006) whereas particle acceleration in magnetic reconnection processes at the Sun seems to generate beams of particles at lower energies – as observed, for example, during the impulsive phase of a flare (see, e.g., Aschwanden *et al.*, 2006; Klassen *et al.*, 2000 and references therein). It is obvious that the key trigger exciting a strong shock wave in the coronal plasma and in the solar wind is the speed of the CME, as has been shown by Gopalswamy *et al.* (2005). High-speed CMEs ($> \sim$1500 km/s) play a crucial role in the acceleration of particles up to GeV energies, which can be registered as short-time cosmic ray intensity increases by neutron monitors on Earth (e.g., Gopalswamy *et al.*, 2005). Such ground level enhancements (GLEs) are caused by interaction of incoming ions with particles in the Earth's atmosphere.

Figure 3.41a shows the rapid increase in intensity of energetic protons on October 29, 2003 as measured by the GOES satellite in different energy channels. The enhanced particle flux lasted through the end of the day when a second particle event occurred. The time period of a decrease in particle intensity after 12:00 UT on October 29 was caused by the geospace transit of the ICME because its internal magnetic field structure is less transparent to energetic particles (see also Bothmer *et al.*, 1997). The two sudden proton flux increases can be associated with two superfast front-side halo (FH) CMEs labeled FH CME 1 and FH CME 2 in the second panel of Figure 3.41a, which represent height–time diagrams for the CME's leading edges tracked in the field of view of the LASCO C2 coronagraph during the displayed time period. Both CMEs were associated with strong X-ray flares measured by GOES (Figure 3.41a, third panel from top). Interplanetary magnetic field data from the ACE (Advanced Composition Explorer) satellite used to identify the subsequent ICME are shown in the bottom panel of Figure 3.41a.

Figure 3.41b provides the same measurements as described in Figure 3.41a, but for the SEP event on January 20, 2005. The intensity–time profile is more impulsive and shorter-lasting compared with the October 2003 events and the spectrum was much harder (Tylka, 2006). Particle energies reached GeV levels and the event was detected as GLE. In fact, in terms of GeV protons it was the largest GLE since 1956 (Tylka, 2006). According to a detailed analysis of the SoHO, ACE, SAMPEX, GOES



SEP intensity time profiles at different in-ecliptic positions with respect to the CME/ICME

Figure 3.42. Intensity-time profiles measured in different solar energetic particle events with a different relative location with respect to the CME source region at the Sun. Adapted from Cane *et al.* (1988).

and RHESSI data and ground-based neutron monitor and radio-wave data, the superfast front-side halo CME had a speed exceeding 3000 km/s, such that it caused a strong shock wave with an Alfvénic Mach number of ~3, capable of accelerating the particles to GeV energies (Gopalswamy *et al.*, 2005; Tylka, 2006). The estimated distance from the Sun at which the protons were accelerated was estimated as 2.6 solar radii from the Sun's center, which interestingly corresponds to the coronal regime where the transition from open to closed magnetic field lines is expected. The composition of the energetic particles measured by ACE, SAMPEX and WIND supports the conclusion that the particles were accelerated out of the low corona and solar wind.

According to the results of Cane *et al.* (1988) the magnetic connection to the onset site of a CME and during its further interplanetary evolution can naturally explain the different intensity–time profiles observed by satellites in individual SEP events, as schematically shown in Figure 3.42. In agreement with the time–intensity profiles in October 2003 and January 2005, the sources of the October CMEs were situated near the solar disk center, whereas the January 2005 CME originated near the Sun's west limb (Figure 3.43). Due to the presence of higher energies of particles, the January 2005 SEP event was a far more serious threat to astronauts than the October 2003 SEP events (see also Chapter 11). It is well known that the SEP event in August 1972, which occurred between the Apollo 16 and 17 missions, would have been lethal to any astronauts on the Moon.

The largest solar proton events since the time they were systematically recorded are listed in Table 3.6. Note that the January 20, 2005 event is not included. The

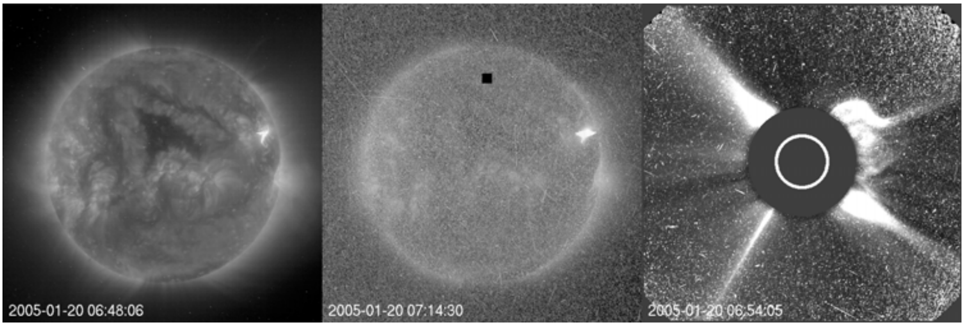


Figure 3.43. SoHO/EIT/LASCO observations of the CME, its source region and proton 'snowstorm' on January 20, 2005. The left image taken at 195 Å at 06:48 UT is a preflare image showing the bright active region to the northwest. The second EIT image shows the flaring region and image contamination by the proton snowstorm, followed by a LASCO/C2 image of the CME.

occurrence times of events listed in Table 3.6 show that, in principle, a SEP event can occur at any given time during a solar cycle – for example, the SEP events in 1994 and 2005 occurred not far from solar activity minimum. Another important aspect of Figure 3.42 is that the best magnetic connection is established with west-limb CMEs. This is the reason the solar source regions of CMEs that cause GLEs are predominately west-limb events (see Gopalswamy *et al.*, 2005). However, it must also be taken into account that, when considered more precisely, the magnetic connection at times of a solar event depends on the curvature of the Parker spiral at 1 AU at that time – that is, on the solar wind speed at 1 AU – so that the source region of the quasi steady-state solar wind is also important in terms of space weather forecasts. Since westward-directed CMEs do not necessarily expand to geospace, their occurrence does not necessarily imply the existence of a large southward B_z-component of the IMF near the Earth. Therefore, SEP events are often observed without the occurrence of geomagnetic storms. The speed at which energetic particles propagate to the Earth, often not much slower than the photons, can be seen in Figure 3.43 where the proton snowstorm is observed within minutes after the onset of the solar eruption. Tylka (2006) has shown that the particles were likely accelerated by the CME-driven shock at a distance of about 2.5 solar radii.

It is commonly assumed that it is in the upstream region of the shock wave driven by a fast CME in the low corona and interplanetary medium where the process of electron acceleration takes place. The accelerated electrons produce the radio emission near the electron plasma frequency as well as its second harmonic known as metric (at frequencies from some tens of MHz to several hundreds of MHz) and kilometric (at frequencies from several kHz up to about ten MHz) radio bursts (e.g., Klassen *et al.*, 2002). Though many questions remain to be answered about the origin and characteristics of, say, type II radio emissions – for example, the role of possible particle acceleration by blast waves initiated during the onset of flare/CME events – they can be considered as reliable indicators of shock-associated CMEs with speeds

Table 3.6. The 25 largest solar proton events measured in geospace between January 1976 and September 2005. Proton fluxes are integral 5-min averages for energies >10 MeV, given in particle flux units (p.f.u.), measured by GOES spacecraft at geosynchronous orbits (1 p.f.u. $=$ 1 particle/(cm^2 s sr)). Different detectors, onboard various GOES spacecraft, have taken the data since 1976. More details are given at *http://umbra.nascom.nasa.gov/SEP/* The full list of proton events has been prepared by the U.S. Department of Commerce, NOAA, Space Environment Center, Boulder, CO.

Proton event			Associated flare and location of AR		
Start (day/UT)	Maximum	Proton flux (p.f.u. @ >10 MeV)	Flare max. (loc./day UT)	Importance (X-ray/opt.)	Location
1978					
Sep 23/10:35	Sep 24/04:00	2,200	Sep 23/10:23	X1/3B	N35W50
1982					
Jul 11/07:00	Jul 13/16:15	2,900	Jul 09/07:42	X9/3B	N17E73
1984					
Apr 25/13:30	Apr 26/14:20	2,500	Apr 25/00:05	X13/3B	S12E43
1989					
Mar 08/17:35	Mar 13/06:45	3,500	Mar 06/14:05	X15/3B	N35E69
Mar 17/18:55	Mar 18/09:20	2,000	Mar 17/17:44	X6/2B	N33W60
Aug 12/16:00	Aug 13/07:10	9,200	Aug 12/14:27	X2/2B	S16W37
Sep 29/12:05	Sep 30/02:10	4,500	Sep 29/11:33	X9/EPL	S26W90
Oct 19/13:05	Oct 20/16:00	40,000	Oct 19/12:58	X13/4B	S27E10
Nov 30/13:45	Dec 01/13:40	7,300	Nov 30/12:29	X2/3B	N26W59
1991					
Mar 23/08:20	Mar 24/03:50	43,000	Mar 22/22:47	X9/3B	S26E28
Jun 04/08:20	Jun 11/14:20	3,000	Jun 04/03:52	X12/3B	N30E70
Jul 07/04:55	Jul 08/16:45	2,300	Jul 07/02:23	X1/2B	N26E03
1992					
May 09/10:05	May 09/21:00	4,600	May 08/15:46	M7/4B	S26E08
Oct 30/19:20	Oct 31/07:10	2,700	Oct 39/18:16	X1/2B	S22W61
1994					
Feb 20/03:00	Feb 21/09:00	10,000	Feb 20/01:41	M4/3B	N09W02
2000					
Jul 14/10:45	Jul 15/12:30	24,000	Jul 14/10:24	X5/3B	N22W07
Nov 08/23:50	Nov 09/15:55	14,800	Nov 08/23:28	M7/mult.	N00-10W75-80
2001					
Sep 24/12:15	Sep 25/22:35	12,900	Sep 24/10:38	X2/2B	S16E23
Oct 01/11:45	Oct 02/08:10	2,360	Oct 01/05:15	M9	S22W91
Nov 04/17:05	Nov 06/02:15	31,700	Nov 04/16:20	X1/3B	N06W18
Nov 22/23:20	Nov 24/05:55	18,900	Nov 22/23:30	M9/2N	S15W34
2002					
Apr 21/02:25	Apr 21/23:20	2,520	Apr 21/01:51	X1/1F	S14W84
2003					
Oct 28/12:15	Oct 29/06:15	29,500	Oct 28/11:10	X17/4B	S16E08
2005					
Jan 16/02:10	Jan 17/17:50	5,040	Jan 15/23:02	X2	N15W05
May 14/05:25	May 15/02:40	3,140	May 13/16:57	M8/2B	N12E11

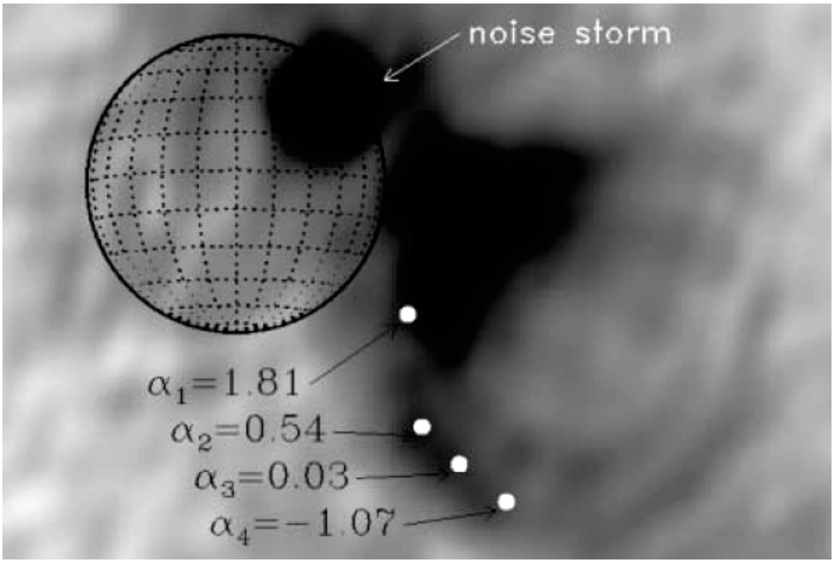


Figure 3.44. Snapshot map of the radio CME at a frequency of 164 MHz at the time of maximum flux (April 20, 1998 at 1013 : 23 UT, Nançay Radioheliograph). Background emission from the Sun has been subtracted. The radio CME is visible as a complex ensemble of loops extended out to the southwest. From Bastian *et al.* (2001).

$> \sim 500$ km/s (Cane *et al.*, 1987). Bastian *et al.* (2001) have used radioheliograph measurements at a frequency of 164 MHz to image a CME at radio waves for the first time. The radio-emitting CME loops visible in Figure 3.44 are the result of nonthermal synchrotron emission from electrons with energies of ~0.5–5 MeV interacting with magnetic fields of ~0.1 to a few gauss. They appeared nearly simultaneously with the onset of a shock-associated type II radio burst, type III radio bursts and the initiation of a solar energetic particle event.

Figures 3.45 (color section) and 3.46 from Klassen *et al.* (2002) show the typical time history of optical, radio-wave and energetic particle measurements for a shock-associated west-limb CME. The MeV electrons measured at Earth's orbit by SoHO/COSTEP were released during or after, but never simultaneously with the onset of type II bursts and CMEs. The time delay between type II burst onset and electron event release ranged from 11.5 to 45 minutes. Thus, the electrons were released either at the end of shock-associated (SA) type II bursts or somewhat later. Most likely they were released when the associated type II burst and the CME reached a certain height, *h*, above the photosphere ($h \sim 1$–$4\ R_S$), such that the expanding CME and its shock wave reached magnetic field lines connected to the observer at Earth's orbit. During the subsequent evolution of the CME in the heliosphere, it drove a shock wave ahead of it as long as the speed gradient of the CME with respect to the ambient solar wind flow was sufficiently large. However, particle acceleration is most efficient close to the Sun since plasma density rapidly decreases with distance from the Sun and because the speed of the CME decreases due to its interaction with slower plasma ahead of it.

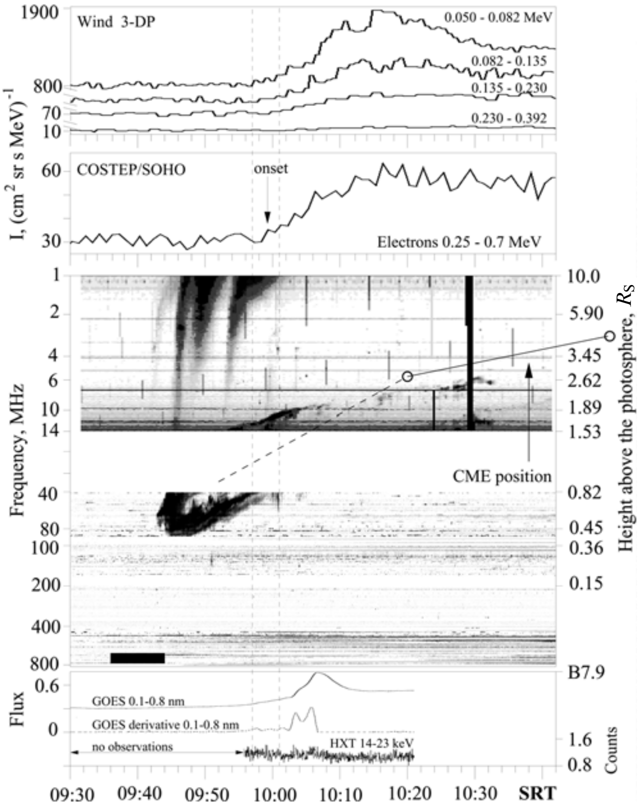


Figure 3.46. Relation between electron intensities in the range 0.050–0.7 MeV and electromagnetic emission at/close to the Sun for the event on May 18, 1998. The coronal type II burst, the CME and the electron event start after the filament eruption (see Figure 3.45). The energetic electrons in all channels were released simultaneously, 16.5 min after the type II and the SA (shock-accelerated) type III bursts onset. Top two panels: electron intensities observed by WIND 3-DP and SoHO/COSTEP instruments. The vertical dashed lines show the onset time interval of electrons detected in the range 0.050–0.392 MeV. The arrow indicates the onset time of electrons in the range 0.25–0.7 MeV. Middle panel: dynamic radio spectrum (800–1 MHz) overlaid with CME trajectory. A type II burst occurs between 86–6 MHz at 09:43–10:42 SRT. From the type II onset, the intense SA type III bursts escape from its lanes. At frequencies above those of type II no other type III, IV were observed. Open circles represent the CME heights (right scale), dashed line is backward extrapolation of the CME trajectory. Black bar denotes the time interval of the filament eruption. Bottom panel: soft X-ray flux (GOES: 1–8 Å, W m^{-2} 10^6), its time derivative (dotted, flux × 50) and hard X-rays measured by Yohkoh. The soft X-ray flare starts after the type II onset. The hard X-rays show – in agreement with the temporal behavior of the soft X-ray derivative – a weak enhancement at 10:03–10:07 SRT after the onset of the electron event. From Klassen *et al.* (2002).

Figure 3.47 (color section) shows the intensity–time profile for electrons and protons observed for the front-side halo CME on April 7, 1997. As expected, the particles measured at the highest energies by SoHO/COSTEP are detected first, with the $\sim$5 MeV protons following the $\sim$0.5 MeV electrons. In contrast to the sudden intensity drop of the MeV particles, the peak intensities of the ~100 keV protons are observed in the upstream region of the interplanetary shock driven by the CME a couple of days later during its passage at 1 AU (Bothmer *et al.*, 1997, 1999). The 'leakage' of upstream shock-accelerated keV protons could be used to track the arrival of the halo CME/ICME similarly to the use of kilometric radio waves shown in Section 3.5. For further details on the different radio-wave signatures associated with solar eruptions the reader is referred to the summaries presented by Aschwanden (2004) and Schwenn (2006).

3.4 SPACE STORMS OVER THE SOLAR CYCLE – TIMES OF OCCURRENCE AND IMPORTANCE OF SOLAR, HELIOSPHERIC AND MAGNETOSPHERIC MODULATIONS

Although space weather forecasts are required on a daily basis, the dependence of the origin and characteristics of geomagnetic storms on the solar cycle phase yields important clues that can be added together to help establish realistic forecasts of space weather in the near future.

The variability of the solar photospheric magnetic field on various temporal and spatial scales shapes the global structure of the overlying corona and drives solar activity during the Sun's 11-year cycle. Input conditions that produce magnetic fluctuations measured at the Earth's surface can only be determined on the basis of satellite observations of the solar wind ahead of the Earth's magnetosphere. Since the beginning of the space age (end of the 1950s to the early 1960s), large databases were compiled which provided scientists with an invaluable resource to investigate the interplanetary causes of space storms and to study a unique set of correlated observations of the Sun, interplanetary space and geospace. Amongst the first systematic studies of the interplanetary causes of geomagnetic storms based on satellite data are those by Gosling (1993b). They analyzed the associations between Earth passage of interplanetary disturbances associated with CMEs and geomagnetic storms for major (Kp 8− to 9), large (Kp 7− to 7+), medium (Kp 6− to 6+) and small (Kp 5− to 5+) storms between August 1978 and October 1982. This was the period around solar activity maximum when the ISEE 3 satellite was operating directly upstream from the Earth. Gosling (1993b) found that all 14 of the major storms during the interval studied were associated with the passage of shock disturbances, and in 13 cases the ICME driving the shock was encountered as well. This reflects the fact that the shock itself does not commonly produce a long-lasting southward IMF component, rather this is primarily associated with the sheath region between the shock and ICME and with the internal magnetic field of the ICME itself. The level of geomagnetic activity stimulated by the shock/ICME in different storms was found to be directly related to the magnitude of the flow speed, magnetic field strength and southward field component associated with the event. These relationships reflect the fact that energy is transferred from the solar wind to the Earth's magnetosphere primarily by means of magnetic reconnection between the IMF and the terrestrial magnetic field at the dayside magnetopause (see Chapter 4). The rate of reconnection, and presumably also the rate at which energy is transferred to the magnetosphere, depends both on solar wind flow speed and the magnetic field strength and orientation. The association of geomagnetic activity with shocks and ICMEs becomes less and less pronounced at lower levels of geomagnetic activity – that is, for medium and small storms.

The solar and interplanetary causes of the five largest geomagnetic storms between 1971 and 1986 were analyzed by Tsurutani *et al.* (1992). Analyzing the satellite data of the solar wind, they found that these five storms were caused by transient fast solar wind flows that were driving shock waves ahead of them. The solar sources could not be investigated in depth, but intense solar flares indicative of strong solar eruptions (fast CMEs) were associated with all events. The key ingredient that was found in the solar wind data was the long-lasting (several hours) extreme magnitude of the southward-directed IMF (which is usually of the order of 5 nT at 1 AU in regular solar wind flows). The enhanced IMF variability was more pronounced than solar wind variability. The southward IMF component at 1 AU that triggered the geomagnetic storms was caused either by draping of the IMF in the sheath region between the shock and subsequent ICME (see Section 3.3.4) and/or by the strong southward internal magnetic field of the ICME itself. The intensity of the geomagnetic storms was amplified in cases when the ICME-driven shocks ram solar wind flows with small pre-existing southward fields and compress those fields.

These two studies – which yielded results for a few major storms around solar activity maximum – were followed by an extended study by Bothmer and Schwenn (1995) who analyzed the interplanetary causes of all major (Kp $\geq 8-$) geomagnetic storms during the years 1966–1990 (when satellite solar wind data without major gaps were taken). The results of this study showed that 41 of the 43 analyzed storms during that time interval were found to be caused by shock-associated ICMEs, one storm was caused by a slow-moving ICME of the magnetic cloud type followed by a CIR and only one by a CIR itself. Thus, independently of the solar cycle phase, major geomagnetic storms are driven by fast (shock-associated) ICMEs. A similar result was obtained using the Dst index as an indicator of geomagnetic storm occurrence. The maximum values of Kp in the individual events were directly related to the peak southward components of the IMF. Draping of the IMF near the front part of ICMEs and/or the magnetic field configuration of the magnetic cloud type (in the ICME itself) were the sources of the extreme negative B_z values. The intensity of the IMF southward component was often substantially amplified at the front and rear parts of magnetic clouds due to their interaction with the ambient solar wind. This was found to be of particular importance for cases when a magnetic cloud type ICME was followed by a CIR or by an interplanetary shock, especially during those very disturbed interplanetary conditions produced by a sequence of ICMEs (through multiple ICMEs or MICMEs). The prime reason CIRs commonly do not trigger major geomagnetic storms is that they are associated with minor magnetic field strengths of shorter duration at 1 AU.

Different contributions of ICMEs and CIRs have been investigated in detail for the years 1972–2000 by Richardson *et al.* (2002). Their results support the previous findings: the most intense storms are nearly solely caused by ICMEs, as shown in Figure 3.48. The authors used almost the same classification as Gosling *et al.* (1993b). The different occurrence rates (storms/year) of small, medium, large and major geomagnetic storms in 1972–2000 caused by ICMEs and co-rotating streams are

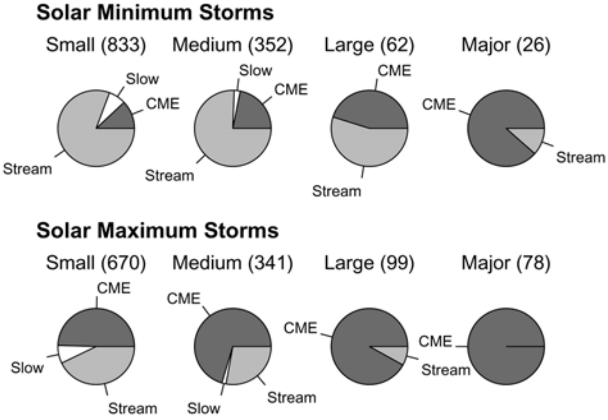


Figure 3.48. Frequency distribution for small, medium, large and major geomagnetic storms as classified by the Kp index, during the years 1972–2000 at different phases of the solar cycle as inferred from analysis of solar wind data. From Richardson *et al.* (2001).

shown in Figure 3.49. ICME-associated storms clearly dominate around times of solar activity maximum and most large and major storms are caused by them. Storms related to co-rotating streams and CIRs are dominant at medium to small storm intensity levels, especially in the declining phase of the solar cycle. Interestingly, the frequency distribution of intense storms over the solar cycle shows a two-peak frequency distribution (as also reported by Gonzalez and Tsurutani, 1990), with peaks before and after the sunspot number maximum, as can be seen in Figure 3.49. The short decrease at times of solar activity maximum might be related to the latitudinal variation in CME source region position (Figure 3.50, Gopalswamy *et al.*, 2003) and to its free expansion to higher latitudes in the absence of polar coronal holes that can systematically deflect CMEs to lower latitudes (see Figure 3.51, Cremades and Bothmer, 2004; Cremades, Bothmer and Tripathi, 2006).

If one ignores the strength of intensity and only takes into account the number of geomagnetically disturbed days (e.g., with Ap $\geq$40), then the picture of enhanced geomagnetic activity in the declining phase of the solar cycle becomes very pronounced, as can be seen in Figure 3.52. These weaker storms are caused by CIRs followed by high-speed streams from coronal holes, as described in Section 3.2.3.

Arguably, the most detailed study of geomagnetic storms based on the unprecedented set of interplanetary measurements from the IMP, WIND and ACE satellites was that undertaken within the EU–ESA/INTAS projects 99-727 and 03-51-6206: the solar and interplanetary sources of all geomagnetic storms in solar cycle 23 with intensity levels of Ap >20 were analyzed. The project website (*http://dbserv.sinp.msu.ru/apev*) includes a catalog of all identified space storms including solar and interplanetary data, which serves as an invaluable tool for space weather researchers. The results of these projects are shown in Table 3.7 and Figure 3.53, which summarize the sources identified from analysis of the interplanetary data for all storms with intensity levels of Ap >20 in 1996–2001 (Bothmer, 2004). Note that there were no such storms observed in 1996, so Table 3.7 and Figure 3.53 start to list events recorded since

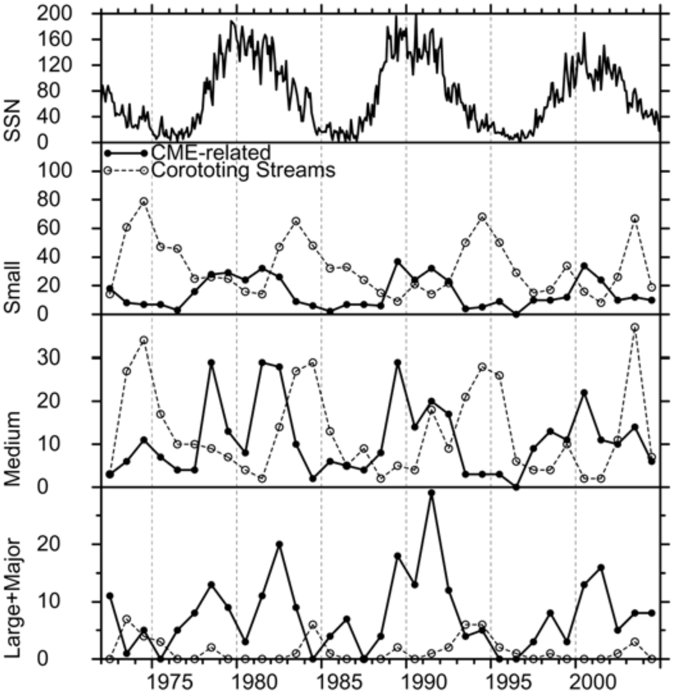


Figure 3.49. Occurrence rates (storms/year) of small, medium, large and major geomagnetic storms in 1972–2005 associated with ICMEs and co-rotating streams displayed together with the sunspot number. From Richardson (2006).

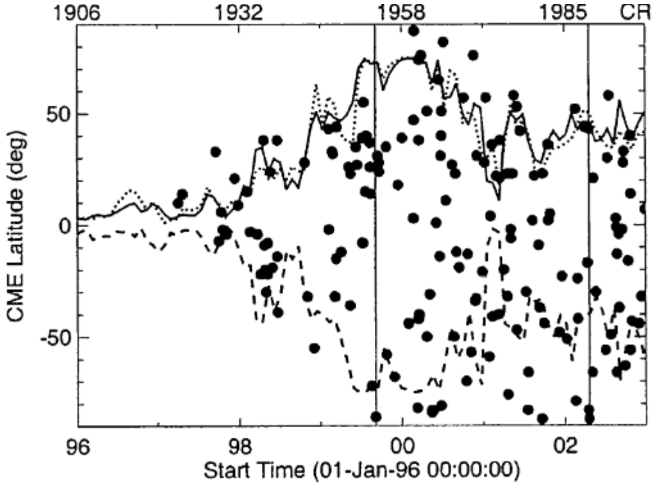


Figure 3.50. Latitudes of prominence eruption-associated CMEs (filled circles) in the northern (dotted line) and southern (dashed line) hemisphere. The solid lines represents an average value. The vertical lines denote the time interval of the observations of high-latitude CMEs. From Gopalswamy *et al.* (2003).

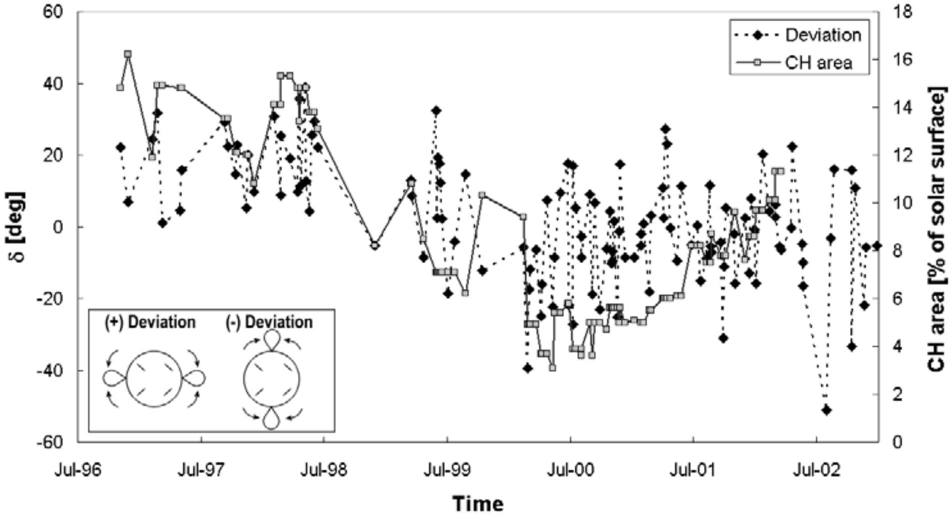


Figure 3.51. Comparison of the deflection angles δ measured for the CMEs' centers with respect to their low coronal source regions with the spatial area of the polar coronal holes at the Sun in 1996–2002. A positive angle corresponds to a deflection towards the ecliptic plane. Note that around times of the Sun's magnetic polarity reversal in 2000, the polar coronal holes have vanished, and that during that time the CME deflection reflects an unsystematic pattern. From Cremades and Bothmer (2004).

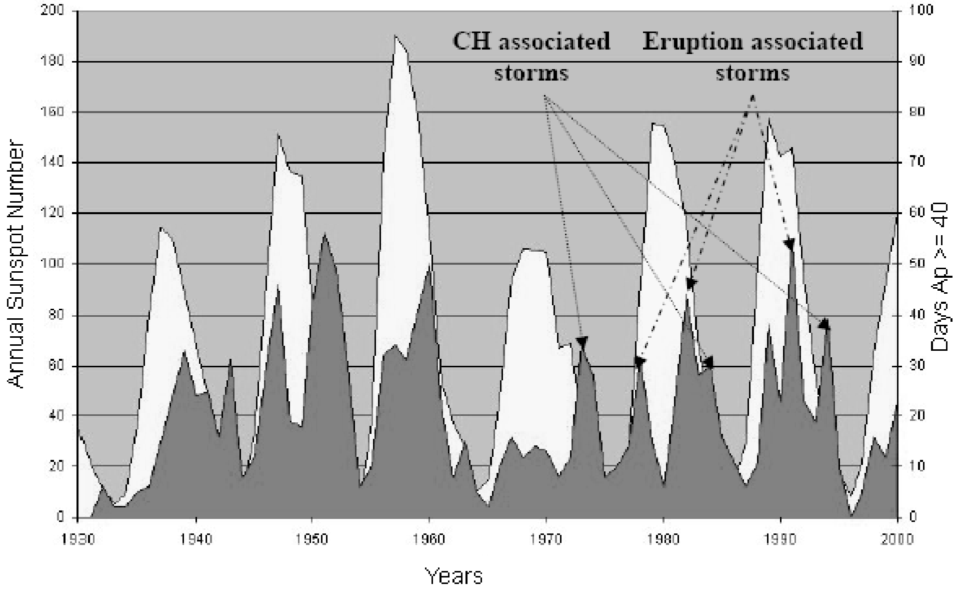


Figure 3.52. Solar cycle variation of the number of geomagnetically disturbed days with Ap $\geq$40. The sunspot number curve is shown in white. The few time intervals that were dominated by different solar drivers of the storms are indicated. Ap diagram adapted from J. Allen, *http://www.ngdc.noaa.gov/stp/GEOMAG/image/APStar_2000sm.gif*

Table 3.7. Causes of geomagnetic storms with Ap >20 during 1996–2001.

Cause of storm	Number of days with Ap >20 during 1997–2001	Number of individual storms	Typical Ap range
Slow solar wind	8	8	<30
CIR/CH	90	55	<60
Combined ICME/CIR	18	11	<150
ICME	101	81	<170
MICMEs	38	30	<200
Total number	*255*	*185*	

Acronyms: co-rotating interaction region (CIR); coronal hole (CH); interplanetary coronal mass ejection (ICME); multiple ICMEs (MICMEs). From Bothmer (2004).

1997. According to this study, the solar wind drivers of geomagnetic storms with intensity levels of Ap >20 can be classified in five different categories as follows:

1. Slow solar wind near the heliospheric current sheet associated with low levels of geomagnetic activity (storm intensity commonly Ap <30).
2. Co-rotating interaction regions (CIRs) followed by coronal hole (CH) high-speed (500–750 km/s) plasma flows (storm intensity commonly Ap <60).
3. Interplanetary counterparts of CMEs (ICMEs) interacting with CIRs (storm intensity commonly Ap <150).
4. Southward IMF components caused by the IMF draping ahead of ICMEs or through their internal magnetic field configuration, or both (storm intensity commonly Ap <170).
5. Multiple ICMEs (storm intensity commonly Ap <200).

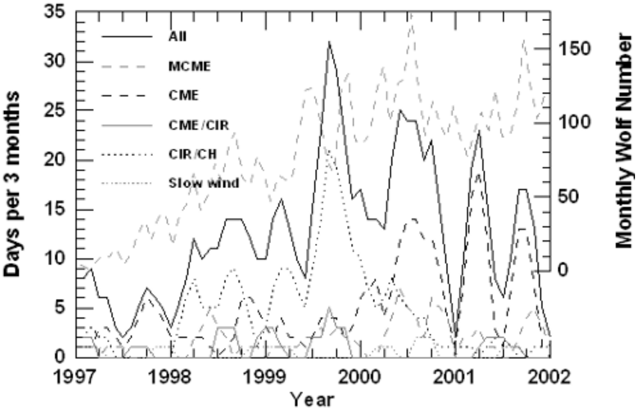


Figure 3.53. Frequency distribution of the different solar/interplanetary drivers of geomagnetic storms with Ap >20 in 1997–2001 compared with the sunspot number as an indicator for the phase of the solar activity cycle. From Bothmer (2004).

Most disturbed days during the investigated period occurred in 1999, 2000 and 2001. Whereas in mid-1999 many storms (although of lower intensity) were still caused by CHs, this picture dramatically changed in 2000 when geomagnetic activity drivers started to become completely ICME-dominated. It is well known that CHs and associated CIRs play major roles as triggers of geomagnetic storms in the decreasing phase of solar activity, but their role in the rising phase of the cycle seems so far to have been underestimated.

Besides the characteristics of photospheric and coronal source regions of quasi steady-state solar wind flows and transient streams due to solar eruptions (CMEs), stream interactions taking place in the heliosphere during the evolution of individual flows up to 1 AU can lead to amplification or weakening of geo-effectiveness in individual events – as described for the formation of CIRs in Section 3.2.3. CIRs can also amplify the geo-effectiveness of ICMEs, as is shown in the example in Figure 3.54 in which an ICME of the magnetic cloud type has been compressed in its trailing

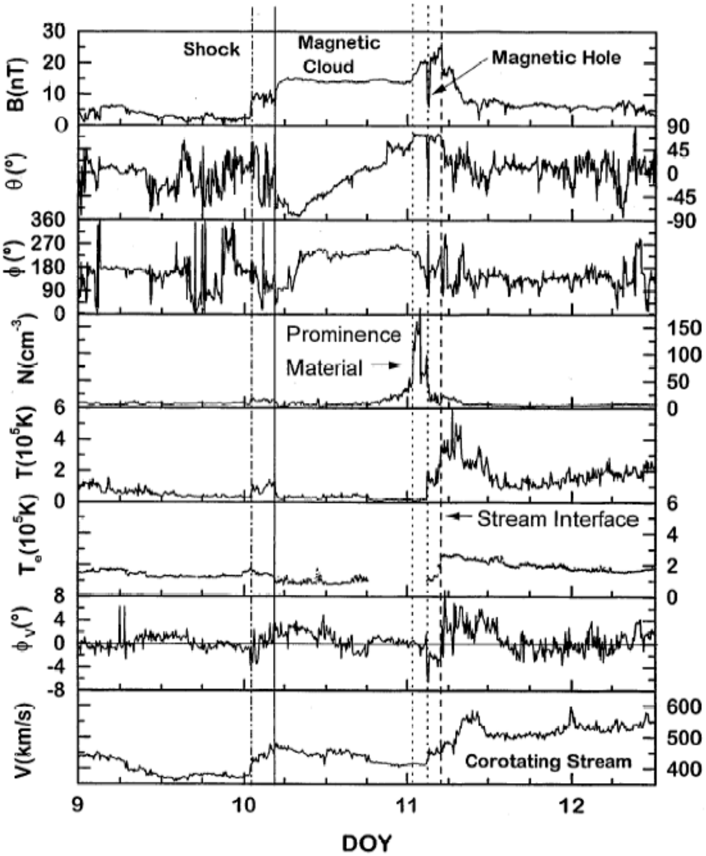


Figure 3.54. Solar wind and magnetic field parameters from January 9 to 12, 1997 showing a magnetic cloud type ICME that was overtaken by a co-rotating stream from a coronal hole. From Burlaga *et al.* (1997).

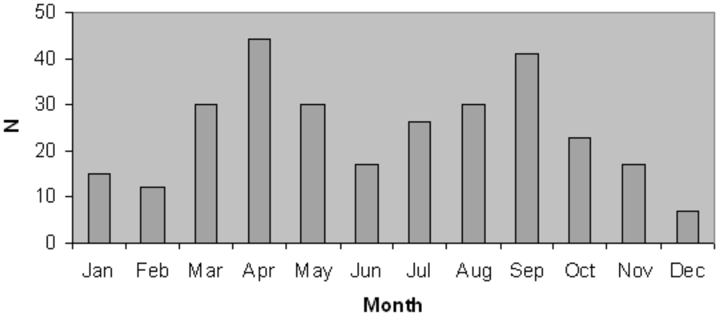


Figure 3.55. Monthly occurrence rates of geomagnetic storms with intensity levels of Kp ≥8− in 1932–2006.

part by a co-rotating high-speed solar wind stream from a coronal hole overtaking the ICME, as observed by the WIND satellite (Burlaga *et al.*, 1998). It is interesting to note that the January 1997 ICME event produced only a moderate geomagnetic storm, but it has become a very famous space weather event in the media because of its association with the malfunction of an AT&T telecommunication satellite (*http://www-istp.gsfc.nasa.gov/istp/cloud_jan97/*). The stream interaction led to compression of the ICME's trailing part, so in the case of a longer interval of the southward field, substantial amplification of the geo-effectiveness of the ICME would have been caused. In this way, multiple transient solar wind streams can cause quite complicated flow structures in the heliosphere (Burlaga *et al.*, 1987).

Beyond the solar and heliospheric modification of solar wind input conditions, the structure of the Earth's magnetosphere also plays a key role in modulating geo-effectiveness. The occurrence rate of intense geomagnetic storms shows a pronounced semiannual variation, as shown in Figure 3.55 where the monthly frequencies of geomagnetic storms with levels of Kp ≥8− are plotted for the years 1932–2006. Most intense storms occur at times near the equinoctial months of March and September, with a clear depression in the number of storms during the solstitial months in June and December. The seasonal variation of geomagnetic activity has been known for a long time (e.g., Bartels, 1963; Gonzalez *et al.*, 1993) and various mechanisms have been proposed in controversy. Historically, the first was the so-called axial mechanism (e.g., Cortie, 1913, 1916; Priester and Cattani, 1962) which was based on the fact that the Earth reaches its maximum heliographic latitude of +7.2° and −7.2° on approximately September 6 and March 5, respectively. The idea was that low-latitude active regions cause solar eruptions that are more likely to reach the Earth. However, today we know that CMEs originating from latitudes around 40° in both solar hemispheres do pass geospace (e.g., Bothmer and Rust, 1997; Bothmer and Schwenn, 1998), so that latitudinal variation does not seem to play the key role in the origin of semiannual variation.

Another early proposed concept to explain the seasonal variation of intense geomagnetic storms was based on variation in the effectiveness of the interaction between the solar wind and the Earth's magnetosphere that depends on the orientation of the Earth's magnetic dipole relative to the Sun–Earth line. In this theory, known as the equinoctial mechanism, storms would be expected to be favorably observed to occur near March 21 and September 23. Another theory, the Russell–McPherron model, takes into account the inclination of the solar magnetic equator with respect to the ecliptic plane, leading to the largest southward components of the IMF (assuming a nominal Parker spiral configuration) in the GSM coordinate system around April 5 and October 5 (Russell and McPherron, 1973b). In this model geomagnetic activity is higher for solar wind streams possessing a negative IMF polarity (i.e., the field is directed towards the Sun) in spring and a positive IMF polarity (i.e., the field is directed away from the Sun) in fall (e.g., Crooker and Siscoe, 1986). It should be noted that, ideally, this holds for streams having the same solar wind speed and that in the case of different speeds the Russell–McPherron effect can be wiped out.

Due to the proximity predicted by the different models for the dates of expected maximum geomagnetic activity and also to the intrinsic variation in the seasonal pattern from year to year, it has been a challenging task to determine the individual contributions of different mechanisms to semiannual variation. As has been shown by Gonzalez *et al.* (1993), the seasonal phase of maximum activity varies with the level of geomagnetic activity. This fact supports the assumption that different solar wind conditions have different intrinsic properties (CIRs, fast solar wind streams, ICMEs) and stimulate geomagnetic activity differently (e.g., Huttunen *et al.*, 2002). Of course, Alfvénic fluctuations causing southward IMF components within high-speed streams, amplified or weakened through the Russell–McPherron effect, do not impose large pressure pulses on the Earth's magnetosphere as in the case of CIRs or fast ICMEs. Extremely fast ICMEs can even compress the day-side magnetopause to a distance as close as about 6 R_E such that geostationary satellites unexpectedly stay for some period of time in the solar wind upstream of the Earth's magnetosphere, as was the case in the March 1989 geomagnetic storm.

It is worth noting that the Kp and Dst indices react differently to solar wind dynamic pressure (e.g., Huttunen *et al.*, 2005, see Section 3.2.2). A scenario that can possibly explain the pronounced semiannual variation of intense geomagnetic storms has been proposed by Crooker *et al.* (1992). The authors provide convincing evidence that intense southward IMFs responsible for great storms can reside in the post-shock plasma preceding fast ICMEs and maybe as well in the internal fields of ICMEs themselves. According to this model, post-shock southward field geo-efficiency results from a major increase in the Russell–McPherron effect through a systematic pattern of compression and draping within the ecliptic plane (see Figure 3.42). Differential compression from the shock increases the Parker spiral angle and, consequently, the azimuthal IMF component that projects as the southward IMF component onto Earth's dipole axis. Southward fields in post-shock flows ahead of ICMEs become strongest at the spring (fall) equinox in ICMEs emerging from toward (away) sectors. We note here that, more precisely, the prime role belongs to the IMF polarity the ICME-driven shock runs into, rather than to the IMF polarity the ICME originated from. ICME source region polarities determine the internal magnetic field configuration of magnetic clouds (Bothmer and Rust, 1997; Bothmer and Schwenn, 1998). The efficiency of the Russell–McPherron effect for post-shock flows is supported by the results of Phillips *et al.* (1992). They have shown that geo-effectiveness associated with shocks and ICMEs is ordered by plasma bulk speed, southward field and pre-existing conditions. When such events are ordered by pre-event GSEQ (geocentric solar equatorial) components, the seasonal effect is clear, with the Dst index being largest near the equinox. Contrarily, no such trend was found for slower ICMEs that are not associated with shocks. Besides the semiannual variation obvious from Figure 3.55, it appears that seasonal variation with slightly enhanced geomagnetic activity in July to September is statistically present. A possible explanation for this seasonal variation could be the uneven hemispheric distribution of the stations from which Kp is derived – 11 of the current total number of 13 Kp stations are located in the northern hemisphere.

Although solar wind input conditions are modulated and intense storms favor the equinox months, an intense storm (Kp $\geq$8−) can in principle occur at any given time of the year and at any phase of the solar cycle (see also Tsurutani *et al.*, 2003). This can be seen from the storm occurrence times of the 25 largest geomagnetic storms (in terms of their peak Kp index) in 1932–2006 listed in Table 3.8. Intense storms occurred in January 1960 and 1949 or in July 1959 and 1946 – that is, not during the equinoctial months favored by seasonal variation. Moreover, sometimes intense storms even occur not far from solar activity minima, as in February 1986 or November 2004. The intensity of an individual geomagnetic storm is in principle independent of the phase and strength of a solar cycle. These findings yield the natural conclusion that space weather forecasting has to be performed on a daily basis, with fundamental solar and heliospheric observations and modeling required, as will be described in the final section of this chapter.

3.5 SOLAR OBSERVATIONS AND MODELING FOR SPACE WEATHER FORECASTS

Since the advent of the space age new scientific observations other than from the Earth's surface have provided us with a new view on the physics of the Sun, its dynamic atmosphere and effects on interplanetary space, including geospace. Besides solar EM radiation, the Sun permanently emits magnetized plasma, either in a quasi steady-state form of slow and fast solar wind streams, or, occasionally, in an explosive manner in the form of coronal mass ejections (CMEs) that evolve in the interplanetary medium as ICMEs (interplanetary coronal mass ejections). Solar wind streams of any type may interact with each other, and compression effects may significantly amplify preexisting southward components of the IMF which are favorable for triggering geomagnetic storms.

To know at any given moment of time the energy state of the interplanetary medium, especially that of the Earth's magnetosphere, which is permanently impacted by the solar wind, it is of crucial importance to provide reliable forecasts of solar wind conditions for a few day in advance. This task requires continuous monitoring of the evolution of solar photospheric magnetic flux as the ultimate driver of the variety of known solar activity phenomena and continuous – if possible in high time resolution –

Table 3.8. The largest 25 geomagnetic storms between January 1932 and July 2006. The last column provides the total sum of all ap-values greater than or equal to 179 (equivalent to Kp $\geq 8-$).

Year	Days	Month	Total sum of 3-hour intervals with ap $\geq$179
1989	13, 14	3	2824
1941	18, 19	9	2736
1960	12, 13	11	2529
1940	29, 30, 31	3	2459
1960	6, 7	10	2451
1960	31, 1	1	2444
1940	24, 25	3	2429
1946	22, 23	9	1944
1986	7, 8, 9	2	1923
1991	24, 25, 26	3	1879
1959	15	7	1807
2003	29, 30	10	1765
1957	4, 5	9	1615
1946	28	3	1579
1946	26, 27	7	1507
1967	25, 26	5	1515
1982	13, 14	7	1515
1941	5	7	1479
2004	9, 10	11	1451
1949	25, 26	1	1422
1958	8	7	1415
1941	1	3	1414
2000	15, 16	7	1386
1958	11	2	1379
1959	17, 18	7	1365

imaging of the corona, preferentially at white-light and EUV wavelengths. These observations can than be used to model the solar wind outflow to predict the expected flow structures and solar wind conditions at Earth's orbit. These conditions are also of prime importance to discover into which global stream structure a suddenly released CME evolves and to provide information into which magnetic background the energetic charged particles that may have been accelerated in association with solar flares and CMEs will propagate.

3.5.1 Modeling the quasi steady-state corona and solar wind

Modeling of the quasi steady-state corona and solar wind may in its easiest way be established based on the so-called potential field source surface (PFSS) model in which the coronal magnetic field is assumed to be current-free ($\nabla \times \mathbf{B} = 0$) (e.g., Hoeksema, 1984; Schrijver and de Rosa, 2003; Wang and Sheeley, 1992). The data input for these models are ground- and space-based magnetograms. New magnetohydrodynamic (MHD) codes are able to take – to some extent – contributions from solar active regions into account (Aschwanden *et al.*, 2006). The structure of the global solar corona has been calculated – for example, based on the Magnetohydrodynamics Around a Sphere (MAS) model developed by the SAIC (Science International Corporation) group – for the range 1–30 R_S based on the strength of the radial magnetic field $B_r(\theta, \varphi)$ as a function of solar latitude (θ) and longitude (φ) provided through full disk synoptic (the data cover the time period of a full solar rotation) magnetograms and coronal temperature $T_e(\theta, \varphi)$ and density $n_e(\theta, \varphi)$ values. Modeling of the solar corona and its changing structure is shown in Figures 3.56 and 3.57 (both in color section) (from Balogh and Bothmer *et al.*, 1999; modeling by Linker *et al.*, 1999, and Mikic *et al.*, 1999). This model has been used successfully to estimate solar wind speeds at the orbit of the Ulysses spacecraft during the Whole Sun Month Campaign in 1996, as shown in Figure 3.56.

Wang *et al.* (1997) modeled the solar wind expansion in the heliosphere by taking into account the locations, areal sizes, rotation and solar cycle evolution of coronal holes. Solar wind flows from the coronal holes can then be reproduced by applying extrapolation techniques to measurements of the photospheric magnetic field and its expansion to estimate the bulk speed, mass and energy densities of solar wind plasma in the heliosphere. Odstrcil *et al.* (2002, 2003) developed a time-dependent solar wind model, the so-called ENLIL code, for the range 21.5 R_S to 1.6 AU for a heliographic latitude range of $\sim 60^\circ$ in the inner heliosphere. The model is also based on solar magnetograms and, in addition, uses the MAS or WSA (Wang–Sheeley–Arge) code to locate inner radial boundary conditions. A detailed summary on the current modeling efforts was given by Aschwanden *et al.* (2006). It should be noted that, unfortunately, the magnetograms currently obtained are not providing precise values of the photospheric magnetic field at heliographic latitudes beyond about 60° and commonly just provide measurements of the longitudinal component of the magnetic field along the line of sight. Figure 3.58 (color section) shows how the method by Wang *et al.* was applied by the U.S. NOAA National Space Environment Laboratory, Boulder, CO, to forecast solar wind speed and IMF polarity at Earth's orbit and to compare the predictions with *in situ* data from the ACE satellite (*http://www.sec.noaa.gov/ws/*). Predicted quasi-stationary solar wind input conditions can then be used to forecast expected geomagnetic activity – similar to the methods used in Figure 3.59 (color section) where solar wind parameters were used to model the expected impact of an ICME on November 20–21, 2003 – in terms of the Dst index according to the models by O'Brien and McPherron (2000) and Wang *et al.* (2003). This event was one of the strongest storms in solar cycle 23, as identified in the Dst index (Huttunen *et al.*, 2005). The O'Brien and McPherron model assumes that the ring current injection and ring current decay parameter are controlled by the solar wind electric field. The Wang *et al.* (2003) model includes the influence of solar wind dynamic pressure in the injection function and the decay parameter. Recently, Mikic and Linker (*http://shadow.adnc.net/corona/mar06eclipse/mar06eclipse.html*) successfully modeled the global structure of the solar corona a week ahead of the solar eclipse on March 29, 2006.

3.5.2 Forecasting coronal mass ejections and solar energetic particle events

Since fast coronal mass ejections (CMEs) cause the most intense space storms and solar energetic particle (SEP) events (see Sections 3.3.4 and 3.5), it is currently the most challenging task to predict the onset and interplanetary consequences of these CMEs. In terms of space storms one is especially interested in the prediction of superfast (2000–3000 km/s) front-side halo CMEs (SFHCMEs). As was shown by Cremades and Bothmer (2004) and Tripathi, Bothmer and Cremades (2004), commonly CMEs originate from localized bipolar regions in the photosphere separating opposite polarity magnetic fields (see Section 3.3.2). Photospheric bipolar regions appear in various spatial sizes and in a variety of intensities and lifetimes, sometimes they occur within a few hours, and to date the physical characteristics of the magnetic regions that produce CMEs are still poorly understood. Figure 3.60 (color section) from Cremades, Bothmer and Tripathi (2006) shows the source regions of CMEs during 1996–2002 together with the evolution of the longitudinal component of the photospheric magnetic field. CMEs originating from active regions (marked red) follow the typical butterfly pattern seen for the sunspots, whereas CMEs originating from quiescent regions (marked green) typically occur at the higher latitudes. Magnetograms thus provide an important base to identify possible source regions of CMEs, however the reasons for their onsets are not yet understood and need much further investigation in the future. So far the onset of CMEs can only be guessed in advance, based on solar magnetograms or solar white-light observations showing the appearance of new active regions at the east limb and the distribution of active and quiescent regions on the visible solar disk. In contrast to observations from Earth, SoHO/MDI provides continuous observations of the photospheric magnetic field. The MDI observations in Figure 3.61 show how, occasionally, photospheric magnetic flux of strong intensity can emerge even at times near solar activity minimum within a relatively short time, as was the case in January 2005. The emerging flux consequently led to major solar eruptions and SEP events as described in Section 3.3.5. Figure 3.62 (color section) shows an example of how the small-scale emergence of magnetic flux in the source region of a prominence-associated CME observed by SoHO on September 12, 2000 may have caused its eruption (Bothmer and Tripathi, 2006). The threshold for magnetic field data was taken as ±200 G in the analysis of the MDI data at a time resolution of 96 minutes, to allow identification of significant changes in magnetic field uniquely. From the observations shown in Figures 3.60 and 3.62 it appears obvious that emerging photospheric flux at various spatial scales is a key trigger of solar activity. Newly emerging flux may already be detected by means of helioseismology methods at the far side of the Sun invisible to direct observations from Earth, as shown in Figure 3.63. Here, a new active region with increased magnetic flux appers as a large-scale local inhomogeneity to the propagation of solar seismic waves. However, in contrast to filament eruptions and CMEs associated with emerging flux, cases have also occurred where annihilation of magnetic flux of opposite polarity was observed along a filament channel (separating regions of opposite magnetic polarity) shortly before the eruption of a filament and CME (Bothmer and Tripathi, 2006). This scenario is referred to as a flux cancellation process (e.g., Linker *et al.*, 2003). Studies

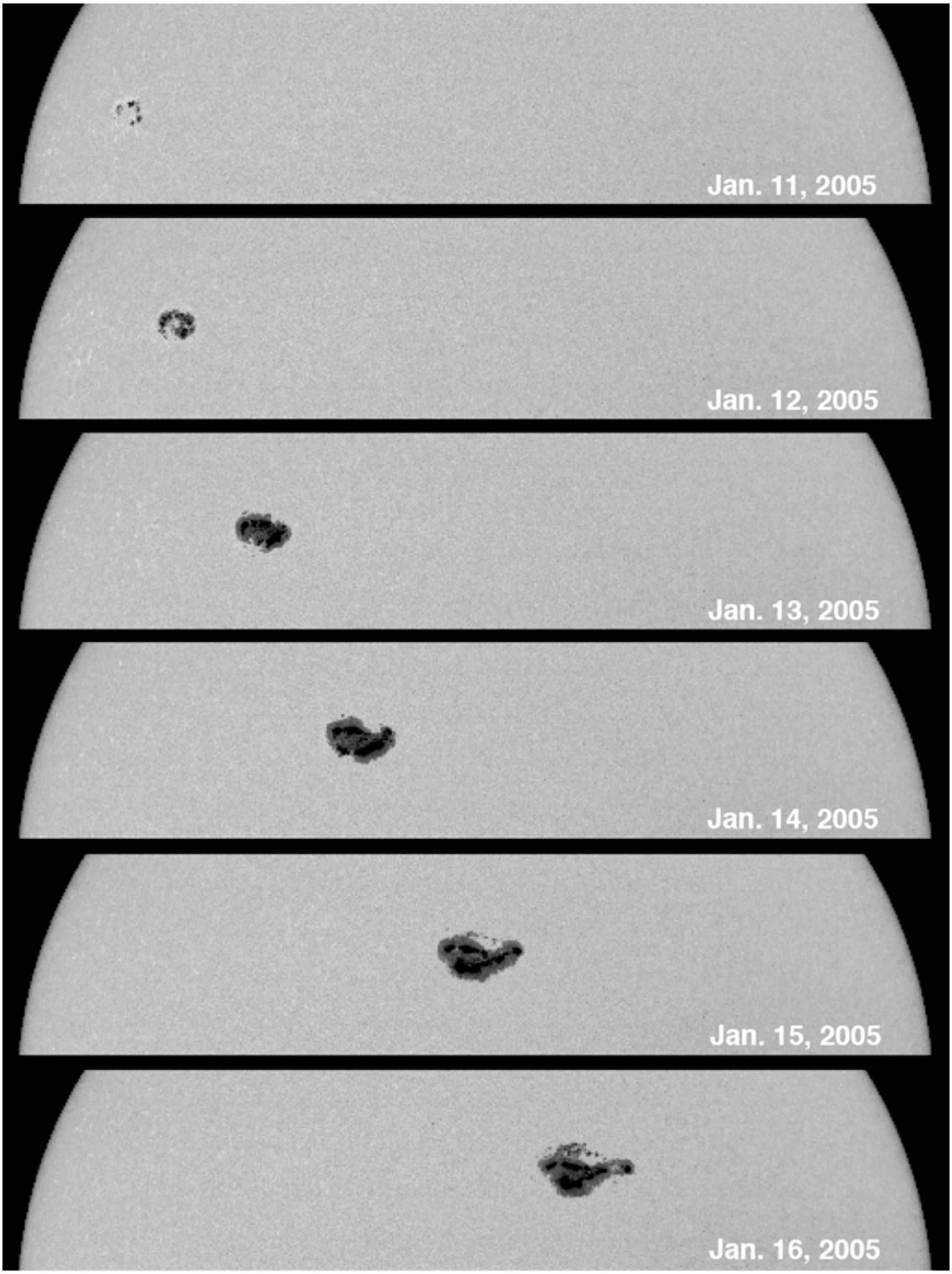


Figure 3.61. SoHO/MDI white-light observations of the development of a sunspot region between January 11 and 16, 2005 when the Sun was approaching solar activity minimum. The active region associated with the emerging magnetic flux caused several fast CMEs, one was associated with a GLE. Courtesy: SoHO/MDI Consortium.

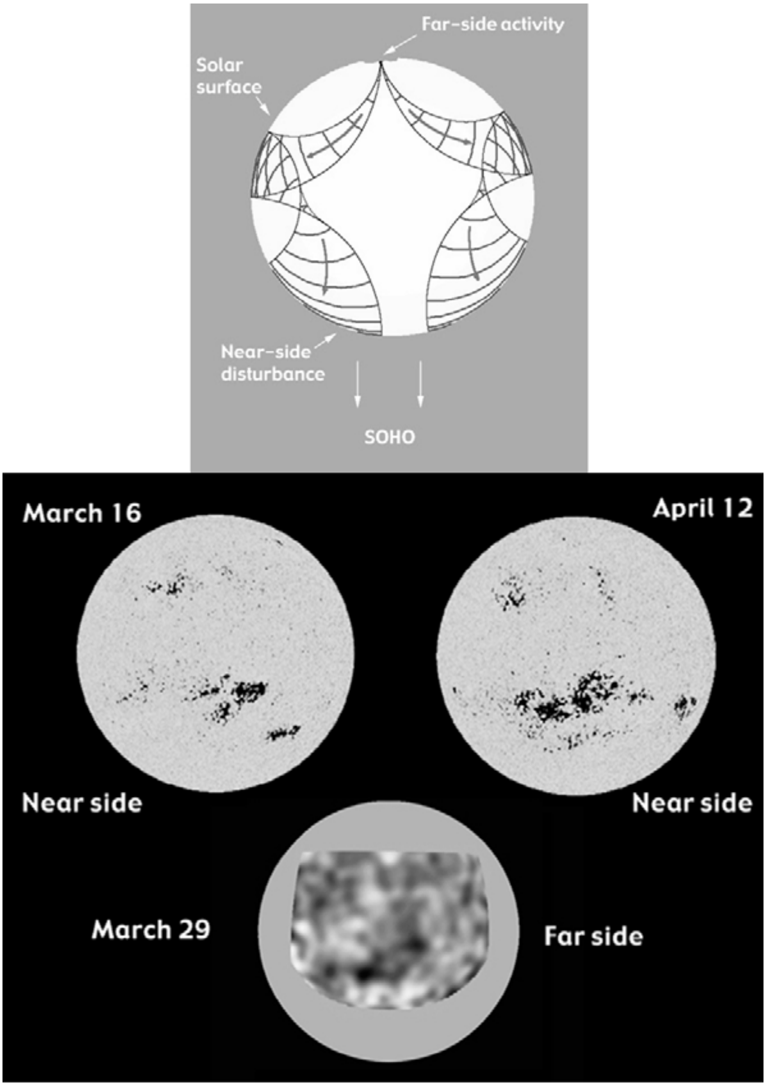


Figure 3.63. Top: an active region on the side of the Sun facing away from the Earth causes sound waves, represented by arcs, that travel through the interior, bounce once off the surface, and reach the side facing the Earth (the near side). The waves generate ripples on the near side surface and are reflected back toward the active region. An active region reveals itself because it possesses very strong magnetic fields that speed up the sound waves. Waves that pass through an active region have a round trip travel time about 12 seconds shorter than the average of 6 hours. The difference becomes evident when sound waves shuttling back and forth get out of step with one another. Bottom: three-panel image, showing absolute magnetic field strength of the same feature (upper left) one-half solar rotation before, and (upper right) one-half solar rotation after the (below, center) holographically imaged far-side region. Courtesy: NASA and ESA (*http://science.nasa.gov/headlines/y2000/ast09mar_1.htm*).

of the relationship of photospheric and coronal observations at a higher time resolution have only been facilitated recently through SoHO. It can be expected that new missions – like STEREO (Solar TErrestrial RElations Observatory) and Solar-B – operating simultaneously and providing disk and limb observations, will yield unprecedented observations of the connection of physical processes of the overlying corona with the evolution of the underlying photospheric magnetic field. Possible thresholds in terms of the magnetic flux in the photosphere that are required to cause CMEs are not known yet but likely exist, since without the presence of bipolar regions covering a substantial area in the photosphere no CMEs seem to occur.

Besides the unresolved question about what causes the onset of a CME is the question about its tentative speed and propagation time to Earth (e.g., dal Lago *et al.*, 2004; Gopalswamy *et al.*, 2001; Huttunen *et al.*, 2005; Schwenn, 2006; Zhang *et al.*, 2003). To determine the arrival times of ICMEs at 1 AU based on coronagraph observations is a difficult task because of the projection effects inherent in white-light observations (e.g., Cremades and Bothmer, 2004). Front-side disk-centered (i.e., likely Earth-directed) CMEs watched from Earth (or the L1 orbit of SoHO) unfortunately appear as unstructured halos (see Figure 3.28). In such cases a reliable radial propagation speed is hard to derive. Dal Lago *et al.* (2004) and Schwenn *et al.* (2005) have developed the model shown in Figure 3.64, which allows forecasting the arrival time of CMEs at Earth's orbit by discriminating between the expansion and propagation speeds of CMEs originating from different source locations at the Sun (limb, near-limb, halo events). A thorough analysis based on CMEs observed by SoHO/LASCO (Figure 3.64) yielded the following estimate for the travel time (T_{tr}) of halo CMEs to 1 AU (for magnetic cloud type ICMEs and shocks see Huttunen *et al.*, 2005):

$$T_{\mathrm{tr}} = 203 - 20.77 * \ln(V_{\mathrm{exp}}) \qquad (T_{\mathrm{tr}} \text{ in hours, } V_{\mathrm{exp}} \text{ in km/s})$$

It seems plausible to assume that those CMEs which propagate into slower ambient solar wind within the inner heliosphere are those which are decelerated most on their way to Earth's orbit. Systematic studies of the properties of CMEs/ICMEs in the inner heliosphere are required to further help understand better their heliospheric evolution (e.g., Forsyth and Bothmer *et al.*, 2006). The proton density, N_{P}, decrease inside magnetic cloud type ICMEs was studied by Bothmer and Schwenn (1998) using data from the Helios 1 and 2 spacecraft for the inner heliosphere between 0.3 and 1.0 AU, with the result that the density inside these ICMEs decreased in proportion to the distance from the Sun as $R^{-2.4}$. The radial size, s, of these magnetic cloud type ICMEs was found to increase proportionally with distance, R, from the Sun as $s(R) = 0.24 \times R^{0.78}$ (R in AU). A quantitative determination of the evolution of the magnetic field strength of ICMEs up to 1 AU and realistic simulations of compression effects due to interactions with the ambient solar wind still remain to be modeled. Figure 3.65 shows a rare case in which kilometric type II radio-wave observations from the WIND spacecraft have been used to track the CME/ICME in January 1997 from the Sun to beyond Earth.

The internal magnetic field configuration and spatial orientation of CMEs/ICMEs may be predicted according to the scenario proposed by Bothmer and Rust (1997) and Bothmer and Schwenn (1994, 1998), in which the helical magnetic flux

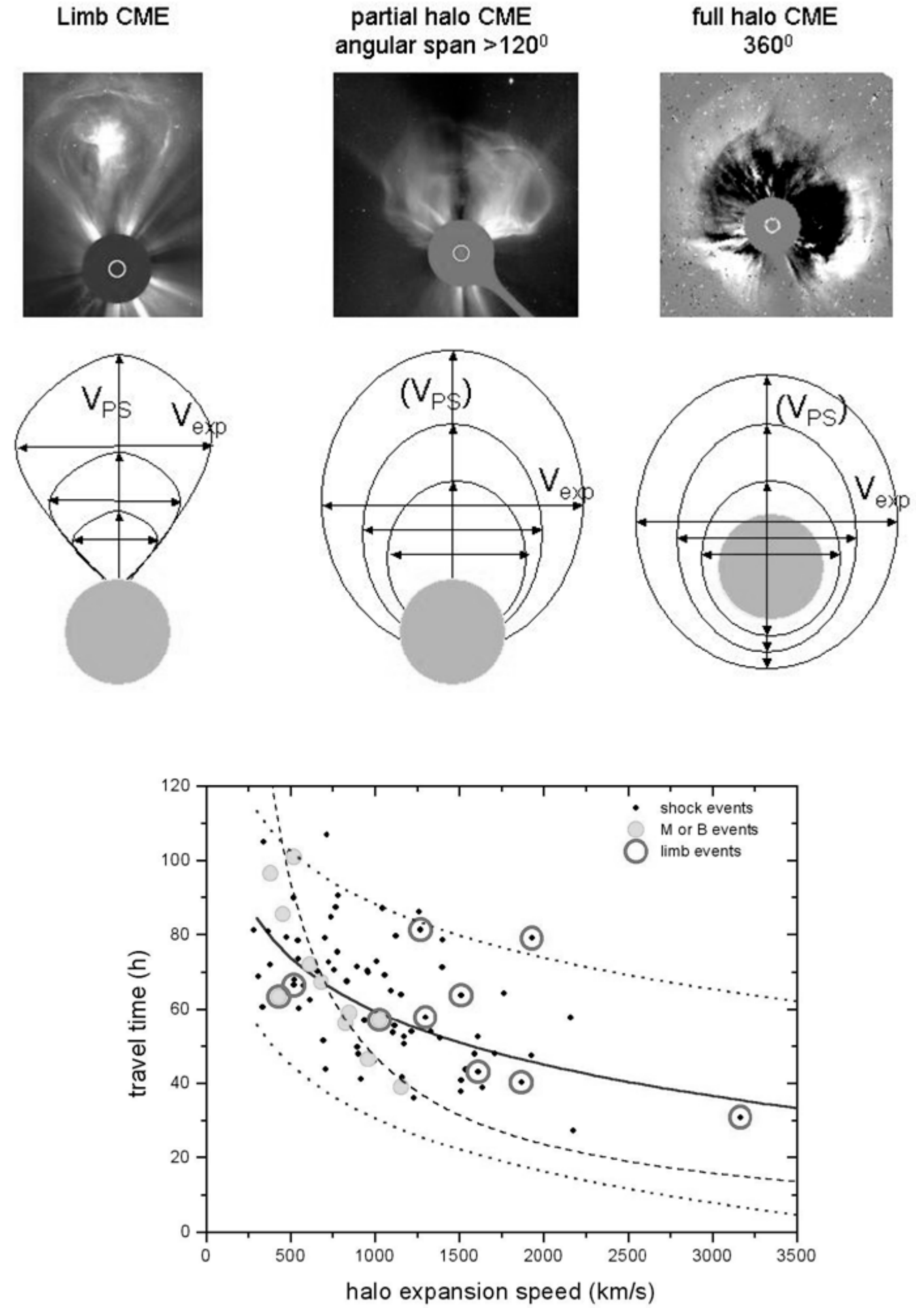


Figure 3.64. Top images: sketch showing the direction of the propagation speed (V_{PS}) and that of the expansion speed ($V_{\exp}$) for CMEs at the limb, near the limb and for halo events. Bottom: comparison of the travel time of CMEs to 1 AU with the calculated halo expansion speeds. From Dal Lago *et al.* (2004).

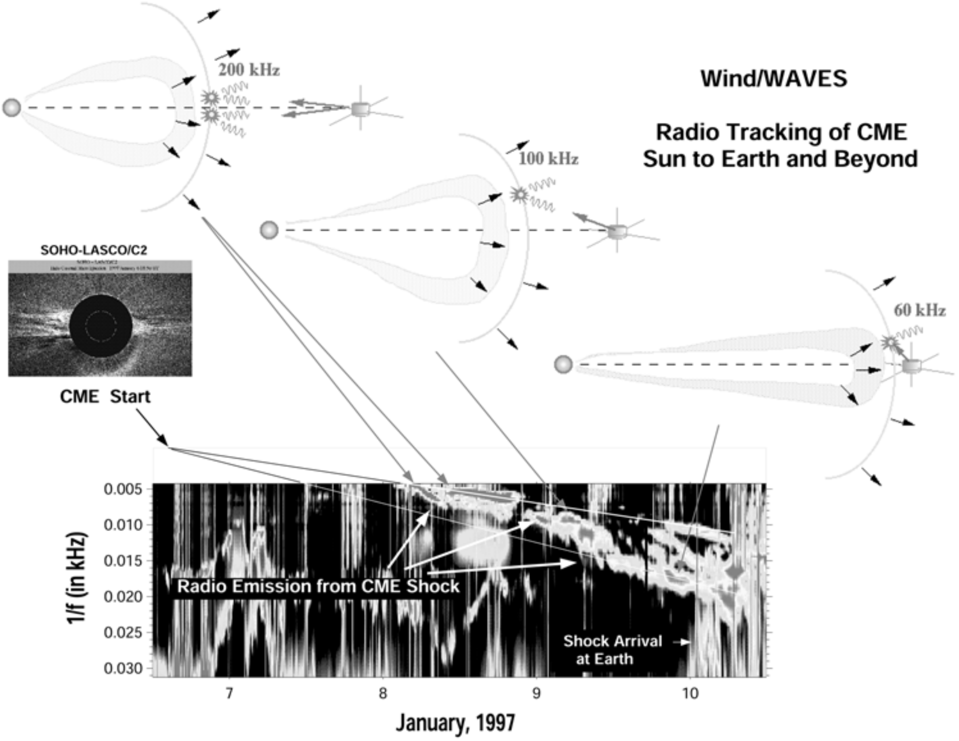


Figure 3.65. Radio tracking of a CME-driven shock from the Sun to beyond Earth by the WIND/WAVES instrument for the January 6, 1997 CME. Propagation of the shock is observed as an interplanetary type II emission. Since the radial distance of the type II emission (proportional to $1/f$) and thus the speed of the shock front depends on the density model and geometry chosen, the sketches at the top could only be constructed after the shock passage by the Earth. Courtesy: M. Kaiser, NASA/GSFC, Science Definition Team Report for the NASA STEREO Mission, 1997.

rope structure of CMEs/ICMEs can be inferred from the underlying magnetic polarity in the CME's photospheric source regions, orientation of the associated filaments and post-eruptive arcades and the handedness preference in the two different solar hemispheres (Figure 3.66). A recent study under development indicates that magnetic cloud type ICMEs with opposite orientations and magnetic configurations to those expected from Figure 3.66 result from those events which originate at the Sun from bipolar regions which do not show the Joy and Hale typical patterns expected for the given solar cycle, such as bipolar regions of opposite magnetic polarity and with orientations reversed from systematic hemispheric inclinations or from quadrupolar magnetic field regions at times when bipolar regions are located close to each other. In these cases hemispheric-handedness is opposite to that expected according to the simple scheme shown in Figure 3.66. This is an encouraging result in that it seems likely in the near future to reliably predict the internal magnetic structure and spatial orientation of ICMEs from the Sun to 1 AU. Future studies should help establish

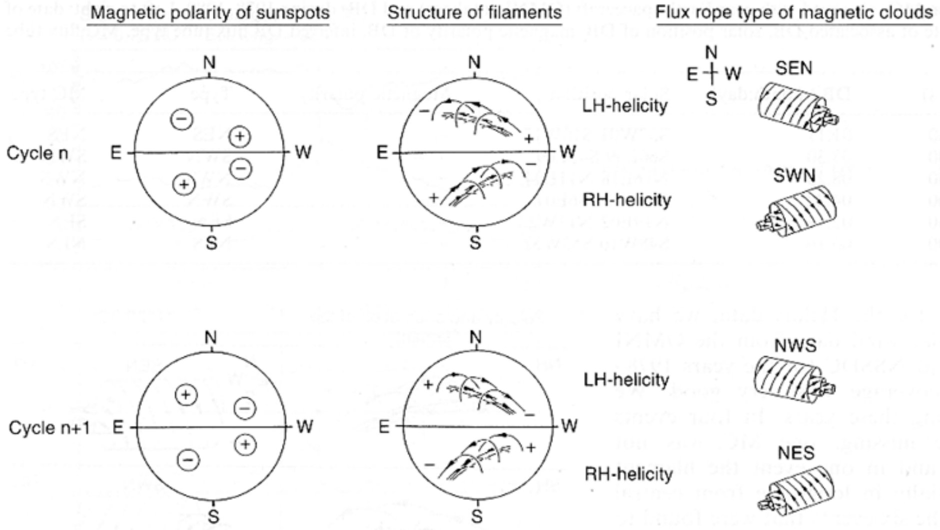


Figure 3.66. Solar cycle dependence of the magnetic field structure of filaments at the Sun and that of the corresponding MCs in the interplanetary medium. Note that for simplicity the MCs are oriented horizontally with respect to the ecliptic plane and that the cycles do not indicate any overlaps nor do they take into account the situation in magnetic regions in the photosphere that reveal a more complex field configuration (e.g., in quadrupolar regions). From Bothmer and Rust (1997), Bothmer and Schwenn (1998).

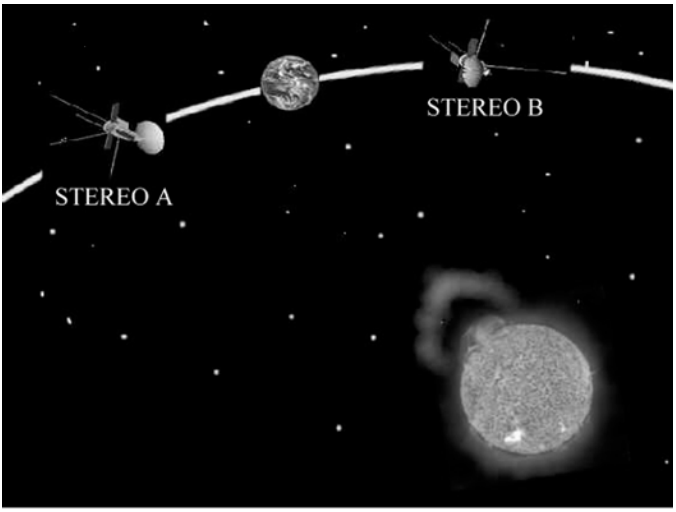
quantitative forecasts for the field strength, diameter and arrival time of ICMEs. It should be noted that the effects of the global structure of the corona should also be taken into account, since CMEs are deflected in their direction of propagation with respect to their low coronal source regions in the presence of polar coronal holes, as shown in Figure 3.51 (Cremades and Bothmer, 2004). What has been learned in the past years from SoHO and Yohkoh observations is that there exist reliable signatures in the solar corona, such as coronal dimmings, EIT waves, post-eruptive arcades, filament eruptions, observable especially at EUV and X-ray wavelengths, from which the onset of a CME can be deduced relatively certainly (Tripathi, Bothmer and Cremades, 2004; Zhukov, 2005; Zhukov and Auchère, 2004), whereas flares can occur without CMEs and sigmoids are signatures of CMEs if a sigmoid to arcade restructuring occurs.

Future research is certainly needed to better understand the onset of CMEs, their 3-D structure and interplanetary evolution as ICMEs. A new mission dedicated to help unravel these questions and help clarify the space weather effects of CMEs/ICMEs is currently under preparation for launch in 2006: NASA's STEREO consists of two suitable, nearly identical spacecraft, equipped with optical telescopes (coronagraphs, EUV imagers and interplanetary cameras) that will allow for the first time study of the 3-D structure of CMEs and their evolution from the Sun to Earth's orbit and beyond, simultaneously from new vantage points from the Sun–Earth line (*http://stp.gsfc.nasa.gov/missions/stereo/stereo.htm*).

Figure 3.67 shows schematically the orbit of the two STEREO spacecraft and the set of near-Sun imagers called SCIP (Sun-Centered Imaging Package), including the entrance aperture mechanism SESAMe (SECCHI Experiment Sun Aperture Mechanism (Howard *et al.*, 2000), the command sent for opening the telescope doors is 'open sesame'). The twin spacecraft will drift by 22° per year in opposite directions in near 1-AU orbits with respect to the Sun–Earth line. Scientists hope that STEREO will approximately operate until 5 years from launch – that is, until 2011 which is the next solar maximum. STEREO will allow for the first time real-time space weather predictions with forecast times from $\sim$1 to 5 days, depending on the speed of the earthward-propagating CMEs/ICMEs. The two STEREO satellites will also sample the *in situ* plasma and magnetic field characteristics of ICMEs and the flows of solar energetic particles (e.g., Luhmann *et al.*, 2005). The physical mechanisms of particle acceleration are at the moment poorly understood, but they are of high importance in order to forecast radiation hazards to astronauts on the ISS and for manned missions to Mars and the Moon (see Chapters 5 and 11) – it is worth pointing out here that Moon and Mars have no shielding atmosphere like the Earth – and also to airline crews and passengers. Though the travel time of high-energy particles to Earth's orbit is only several tens of minutes, and short-time exposure might not be predictable, long-time (several hours to days) exposure for humans can be avoided through real-time CME/flare alerts based on routine observations of the evolution of the photospheric magnetic field and imaging of the corona. Finally, Table 3.9 provides a brief payload concept for a future space mission operating in low Earth orbit (LEO) that could provide the most important observations needed to help establish reliable space weather forecasts.

3.6 ACKNOWLEDGEMENTS

V. B. acknowledges the support of the project Stereo/Corona by the German Bundesministerium für Bildung und Forschung through the Deutsche Zentrum für Luft- und Raumfahrt e.V. (DLR, German Space Agency). Stereo/Corona is a science and hardware contribution to the optical imaging package SECCHI, developed for the NASA STEREO mission to be launched in 2006. Further information can be found at *http://stp.gsfc.nasa.gov/missions/stereo/stereo.htm*. A. Z. acknowledges support from the Belgian Federal Science Policy Office through the ESA-PRODEX programme. Both authors acknowledge the support of the EU–ESA/INTAS projects 99-727 and 03-51-6206. We thank all the members of the SoHO/LASCO/EIT/MDI consortium who built the instruments and provided the data used in this chapter. LASCO, EIT and MDI images are courtesy of the SoHO consortium. SoHO is a project of international cooperation between ESA and NASA. The *CME Catalog* is generated and maintained by NASA and the Catholic University of America in cooperation with the Naval Research Laboratory. We acknowledge the use of H_α data from the Observatory of Meudon. The NSO/Kitt Peak data shown are produced cooperatively by NSF/NO AO, NASA/GSFC and NOAA/SEL. We also would like to thank the Yohkoh, TRACE, WIND and ACE consortia in the same way as the SoHO consortium.



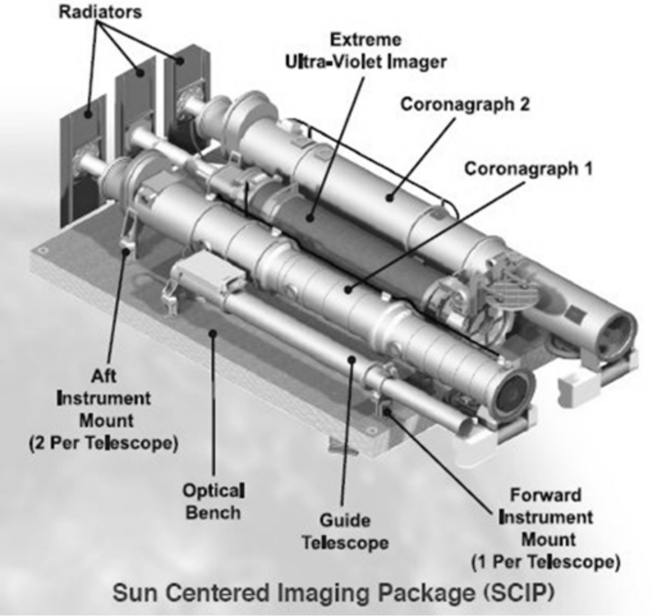


Figure 3.67. Top: schematic of the orbit of the STEREO spacecraft A (ahead) and B (behind). Courtesy: NASA STEREO Consortium. Bottom: the SCIP (Sun-Centered Imaging Package) of the SECCHI (Sun–Earth Connection Coronal and Heliospheric Investigation) imaging package for the NASA STEREO Mission. SESAMe are the SECCHI experiment Sun apperture mechanisms (reclosable doors). Courtesy: SECCHI Consortium.

Table 3.9. Prime payload for solar observations enabling reliable space weather forecasts.

Instrument	Wavelength for observations, field of view and time cadence	Space weather target
EUV imager	195 Å, 171 Å full disk, $\lesssim 5$ min.	Detection of coronal transients, active regions, CME onsets and post-eruptive signatures, heliographic location of CME source region
Coronagraph	White-light, 1.5–10 R_S, $\lesssim 10$ min.	Detection of halo and partial halo CMEs and provision of speed estimates
Magnetograph (if not supplied through ground-based observations)	Ni I 6768, full disk scans, <3 hours	Detection of newly emerging flux, active regions, sunspots, flux evolution, magnetic structure of CME source regions

Thanks go further to Dr. Hebe Cremades, Dr. Durgesh Kumar Tripathi and Ralf Kotulla who helped solve technical issues and/or provided results from their work. I would finally like to thank Daniel Berdychevsky and Ronald Lepping for providing the solar wind data plot for the Bastille event.

3.7 REFERENCES

Alfvén, H., Electric currents in cosmic plasmas, *Rev. Geophys. Space Phys.*, **15**, 271, 1977.

Allen, J.H. and D.C. Wilkinson, Solar–terrestrial activity affecting systems in space and on Earth, *Solar–Terrestrial Predictions – IV: Proceedings of a Workshop at Ottawa, Canada May 18–22, 1992*, J. Hruska, M.A. Shea, D.F. Smart, and G. Heckman (eds.), p. 75, 1992.

Allen, J.H., H. Sauer, L. Frank and P. Reiff, Effects of the March 1989 solar activity, *EOS*, **70-46**, 1486–1488, 1989.

ANSER, *Foundations of Solar Particle Event Risk Management Strategies*, prepared under NASA Grant NAGW-4166 for the NASA Mission from Planet Earth Study Office and Life and Microgravity Science and Applications Office, ANSER Suite 800, 1215 Jefferson Davis Highway, Arlington, VA 22202, 1996.

Aschwanden, M., *Physics of the Solar Corona: An Introduction*, Springer-Praxis, 2004.

Aschwanden, M., S.K. Antiochos, J.W. Cook, R.A. Howard, J.T. Karpen, J.A. Klimchuk, N.R. Sheeley, L.A. Fisk, T.I. Gombosi, N. Lugaz *et al.*, Theoretical modeling for the STEREO Mission, *Space Sci. Rev.*, in press, 2006.

Badalyan, O.G., V. Obridko and N.J. Sykora, *Solar Physics*, **199**, 421, 2001.

Baker, D.N., Specifying and forecasting space weather threats to human technology, in *Effects of Space Weather on Technology Infrastructure*, I.A. Daglis (ed.), Kluwer Academic, pp. 1–25, 2004.

Balogh, A., V. Bothmer (Co-Chairs), N. Cooker, R.J. Forsyth, G. Gloeckler, A. Hewish, M. Hilchenbach, R. Kallenbach, B. Klecker, J.A. Linker *et al*., The solar origin of corotating interaction regions and their formation in the inner heliosphere (Report of Working Group 1), *Space Sci. Rev*., **89**, 141–178, 1999.

Balogh, A., R.G. Marsden and E.J. Smith, *The Heliosphere near Solar Minimum: The Ulysses Perspective*, Springer-Praxis, 2001.

Bartels, J., Terrestrial–magnetic activity and its relations to solar phenomena, *Terr. Magn. Atmosph. Electr*., **37**, 1–52, 1932.

Bartels, J., Discussion of time-variations of geomagnetic activity indices Kp and Ap, 1932–1961, *Ann. Geophys*., **19**, 1–20, 1963.

Bartels, J. and J. Veldkamp, Geomagnetic and solar data: International data on magnetic disturbances, *J. Geophys. Res*., **54**, 295–299, 1949.

Bastian, T.S., M. Pick, A. Kerdraon, D. Maia and A. Vourlidas, The coronal mass ejection of 1998 April 20: Direct imaging at radio wavelengths, *Astrophys. J*., **558**, L65–L69, 2001.

Benevolenskaya, E.E., A.G. Kosovichev, J.R. Lemen, P.H. Scherrer and G.L. Slater, Large-scale solar coronal structure in soft X-rays and their relationship to the magnetic flux, *Astrophys. J*., **571**, L181–L185, 2002.

Biermann, L., Kometschweife und solare Korpuskularstrahlung, *Z. Astrophys*., **29**, 274, 1951.

Bohlin, J.D. and N.R. Sheeley Jr., Extreme ultraviolet observations of coronal holes: II. Association of holes with solar magnetic fields and a model for their formation during the solar cycle, *Solar Physics*, **56**, 125–151, 1978.

Borello Filisetti, O., G. Lovera, V. Mussino, M. Parisi and M. Storini, On the geomagnetic response to high-speed solar wind streams, *Ann. Geophys*., **6**, 627–634, 1988.

Bothmer, V., Solar corona, solar wind structure and solar particle events, *Proc. of ESA Workshop on Space Weather, November 1998*, ESA WPP-155, ISSN 1022-6656, pp. 117–126, 1999.

Bothmer, V., Sources of magnetic helicity over the solar cycle, *Proc. ISCS 2003 Symposium, 'Solar Variability as an Input to the Earth's Environment'*, ESA SP-535, p. 419, 2003.

Bothmer, V., The solar and interplanetary causes of space storms in solar cycle 23, *IEEE Transactions on Plasma Science*, **32**, 4, 2004.

Bothmer, V., The solar atmosphere and space weather, in *Solar System Update*, P. Blondel and J.W. Mason (eds.), Springer-Praxis, pp. 1–53, 2006.

Bothmer, V. and Rust, D.M., *The Field Configuration of Magnetic Clouds and the Solar Cycle*, AGU Geophys. Monogr. 99, pp. 139–146, 1997.

Bothmer, V. and Schwenn, R., Eruptive prominences as sources of magnetic clouds in the solar wind, *Space Sci. Rev*., **70**, 215–220, 1994.

Bothmer, V. and R. Schwenn, The interplanetary and solar causes of major geomagnetic storms. *J. Geomagn. Geoelectr*., **47**, 1127–1132, 1995.

Bothmer, V. and Schwenn, R., Signatures of fast CMEs in interplanetary space, *Adv. Space Res*., **17**, 319–322, 1996.

Bothmer, V., and Schwenn, R., The structure and origin of magnetic clouds in the solar wind, *Annales Geophysicae*, 1–24, 1998.

Bothmer, V. and D. Tripathi, Photospheric field evolution in the source regions of coronal mass ejections, *SOHO 17 Conference Proceedings*, in press, 2006.

Bothmer, V., M.I. Desai, R.G. Marsden, T.R. Sanderson, K.J. Trattner, K.-P. Wenzel, J.T. Gosling, A. Balogh, R.J. Forsyth, B.E. Goldstein, Ulysses observations of open and closed magnetic field lines within a coronal mass ejection, *Astron. Astrophys*., *316*, 493–498, 1996.

Bothmer, V., B. Heber, H. Kunow, R. Müller-Mellin, G. Wibberenz, J.T. Gosling, A. Balogh, A. Raviart and C. Paizis, The effects of coronal mass ejections on galactic cosmic rays in the high latitude heliosphere: Observations from Ulysses' first orbit, *Proc. 25th Int. Cosmic Ray Conference*, Space Research Unit, **1**, 333–336, 1997a.

Bothmer, V., A. Posner, H. Kunow, R. Müller-Mellin, B. Heber, M. Pick, B.J. Thompson, J.-P. Delaboudinière, G.E. Brueckner, R.A. Howard *et al.*, Solar energetic particle events and coronal mass ejections: New insights from SOHO, *Proc. 31st ESLAB Symposium*, ESA SP-415, p. 207, 1997b.

Brekke, P., B. Fleck, S.V. Haugan, T. van Overbeek, H. Schweitzer and M. Chaloupy, Space weather effects on SOHO and its role as a space weather watchdog, *Proc. 8th Spacecraft Charging Technology Conf., October 20–24, 2003, Huntsville, AL*, in press, 2006.

Brueckner, G.E., R.A. Howard, M.J. Koomen, C.M. Korendyke, D.J. Michels, J.D. Moses, D.G. Socker, K.P. Dere, P.L. Lamy, A. Llebaria *et al.*, The Large Angle Spectroscopic Coronagraph (LASCO), *Solar Phys.*, **162**, 357–402, 1995.

Brueckner, G.E., J.-P. Delaboudiniere, R.A. Howard, S.E. Paswaters, O.C. St. Cyr, R. Schwenn, P. Lamy, G.M. Simnett, B. Thompson, D. Wang, Geomagnetic storms caused by coronal mass ejections (CMEs): March 1996 through June 1997, *Geophys. Res. Lett.*, **25**, 3019, 1998.

Burlaga, L.F., Interaction with the earth, *Space Sci. Rev.*, **17**, 327, 1975.

Burlaga, L.F. and J.H. King, Intense interplanetary magnetic fields observed by geocentric spacecraft during 1963–1975, *J. Geophys. Res.*, **84**, 6633, 1979.

Burlaga, L.F. and R.P. Lepping, The causes of recurrent geomagnetic storms, *Planet. Space Sci.*, **25**, 1151–1160, 1977.

Burlaga, L., Sittler, E., Mariani, F. and Schwenn, R., Magnetic loop behind an interplanetary shock: Voyager, Helios and IMP 8 observations, *J. Geophys. Res.*, **86**, 6673–6684, 1981.

Burlaga, L.F., L. Klein, N.R. Sheeley Jr., D.J. Michels, R.A. Howard, M.J. Koomen, R. Schwenn and H. Rosenbauer, A magnetic cloud and a coronal mass ejection, *J. Geophys. Res.*, **9**, 1317–1320, 1982.

Burlaga, L.F., K.W. Behannon and L.W. Klein, Compound streams, magnetic clouds, and major geomagnetic storms, *J. Geophys. Res.*, **92**, 5725–5734, 1987.

Burlaga, L., R. Fitzenreiter, R. Lepping, K. Ogilvie, A. Szabo, A. Lazarus, J. Steinberg, G. Gloeckler, R. Howard, D. Michels *et al.*, A magnetic cloud containing prominence material: January 1997, *J. Geophys. Res.*, **103**, 277–285, 1998.

Burton, R.K., R.L. McPherron and C.T. Russell, An empirical relationship between interplanetary conditions and Dst, *J. Geophys. Res.*, **80**, 4204, 1975.

Cane, H.V. and I.G. Richardson, Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002, *J. Geophys. Res.*, **108**, 1156, doi:10.1029/2002JA009817, 2003.

Cane, H.V., N.R. Sheeley Jr. and R.A. Howard, Energetic interplanetary shocks, radio emission, and coronal mass ejections, *J. Geophys. Res.*, **92**, 9869–9874, 1987.

Cane, H.V., D.V. Reames and T.T. von Rosenvinge, The role of interplanetary shocks in the longitude distribution of solar energetic particles, *J. Geophys. Res.*, **93**, 9555–9567, 1988.

Canfield, R.C., H.S. Hudson and D.E. McKenzie, Sigmoidal morphology and eruptive solar activity, *Geophys. Res. Lett.*, **26**(6), 627, 1999.

Carrington R.C., Description of a singular appearance in the Sun on September 1, 1859, *Monthly Not. Royal Astron. Soc.*, **20**, 13, 1860.

Chao, J.K. and R.P. Lepping, A correlative study of ssc's, interplanetary shocks, and solar activity, *J. Geophys. Res.*, **79**(13), 1799, 1974.

Chen, J., R.A. Santoro, J. Krall, R.A. Howard, R. Duffin, J.D. Moses, G.E. Brueckner, J.A. Darnell and J.T. Burkepile, Magnetic geometry and dynamics of the fast coronal mass ejection of 1997 September 9, *Astrophys. J.*, **533**, 481, 2000.

Cortie, A.L., Sunspots and terrestrial magnetic phenomena, 1898–1911, *Monthly Not. Roy. Astron. Soc.*, **73**, 52–60, 1913.

Cortie, A.L., Sunspots and terrestrial magnetic phenomena, 1898–1911, *Monthly Not. Roy. Astron. Soc.*, **76**, 15–18, 1916.

Cremades, H. and Bothmer, V., On the three-dimensional configuration of coronal mass ejections, *Astron. and Astrophys.*, **422**, 307, 2004.

Cremades, H., V. Bothmer and D. Tripathi, Properties of structured coronal mass ejections in solar cycle 23, *Adv. Space Res.*, in press, 2006.

Crooker, N.U. and G.L. Siscoe, The effects of the solar wind on the terrestrial environment, in *Physics of the Sun*, Vol. III, D. Reidel, p. 193, 1986.

Crooker, N.U., E.W. Cliver and B.T. Tsurutani, The semiannual variation of great geomagnetic storms and the postshock Russell–McPherron effect preceding coronal mass ejecta, *Geophys. Res. Lett.*, **19**, 429–432, 1992.

Crooker, N.U. and J.T. Gosling (Co-Chairs), V. Bothmer, R.J. Forsyth, P.R. Gazis, A. Hewish, T.S. Horbury, D.S. Intriligator, J.R. Jokipii, J. Kóta *et al.*, CIR morphology, turbulence, discontinuities, and energetic particles (Report of Working Group 2), *Space Sci. Rev.*, **89**, 179, 1999.

Daglis, I.A. (ed.), *Effects of Space Weather on Technology Infrastructure*, NATO Science Series, II. Mathematics, Physics and Chemistry, p. 176, 2004.

Dal Lago, A., L.E.A. Vieira, E. Echer, W.D. Gonzalez, A.L.C. de Gonzalez, F.L. Guarnieri, N.J. Schuchand R. Schwenn, Comparison between halo cme expansion speeds observed on the Sun, the related shock transit speeds to Earth and corresponding ejecta speeds at 1 AU, *Solar Physics*, **222**(2), 323, 2004.

Delaboudinière, J.-P., G.E. Artzner, J. Brunaud, A.H. Gabriel, J.F. Hochedez, F. Millier, X.Y. Song, B. Au, K.P. Dere, R.A. Howard *et al.*, EIT: Extreme-Ultraviolet Imaging Telescope for the SOHO Mission, *Solar Physics*, **162**, 291–312, 1995.

Dere, K., J. Wang and Y. Yan (eds.), Coronal and stellar mass ejections, *Proc. IAU Symposium 226, Beijing*, 2004.

Dikpati, M., G. de Toma and P.A. Gilman, Predicting the strength of solar cycle 24 using a flux-transport dynamo-based tool, *Geophys. Res. Lett.*, **33**, L05102, doi:10.1029/2005GL025221, 2006.

Eddy, J.A., Historical evidence for the existence of the solar cycle, in *TheSolar Output and Its Variation*, O.R. White (ed.), Associated University Press, Boulder, CO, 1977.

Feynman, J. and S.F. Martin, The initiation of coronal mass ejections by newly emerging magnetic flux, *J. Geophys. Res.*, **100**, 3355, 1995.

Fleck, B. and Svetska, Z. (eds.), The first results from SOHO, *Solar Physics*, **170**, 1; **172**, 2, Kluwer Academic, 1997.

Fleck, B., Domino, V. and Poland, A., The SOHO mission, *Solar Physics*, **162**, 1, Kluwer Academic, 1995.

Forbes, T.G., A review on the genesis of coronal mass ejections, *J. Geophys. Res.*, **105**, 23153, 2000.

Forsyth, R.J. and V. Bothmer (Co-Chairs), C. Cid, N.U. Crooker, T.S. Horbury, K. Kecskemety, B. Klecker, J.A. Linker, D. Odstrcil, M.J. Reiner *et al.*, ICMEs in the inner heliosphere: Origin, evolution and propagation effects, *Space Sci. Rev.*, in press, 2006.

Froehlich, C., Solar irradiance variability, in *Solar Variability and Its Effects on the Earth's Atmosphere and Climate System*, AGU Monograph Series, J.M. Pap and P. Fox (eds.), AGU, Washington, D.C., 2003.

Gibson, S.E. and B.C. Low, Three-dimensional and twisted: An MHD interpretation of on-disk observational characteristics of coronal mass ejections, *J. Geophys. Res.*, **105**, A8, 18187, 2000.

Goldstein, H., On the Field configuration in magnetic clouds, in *Solar Wind Five*, NASA Conf. Publ. CP-2280, pp. 731–733, 1983.

Gonzalez, W.D., A unified view of solar wind–magnetosphere coupling functions, *Planetary and Space Science*, **38**(5), 627, 1990.

Gonzalez, W.D. and A.L.C. Gonzalez, Dual-peak solar cycle distribution of intense geomagnetic storms, *Planet. Space Sci.*, **38**, 181, 1990.

Gonzalez, W.D. and B.T. Tsurutani, Criteria of interplanetary parameters causing intense magnetic storms (Dst $< -100\,\mathrm{nT}$), *Planet. Space Sci.*, **35**, 1101, 1987.

Gonzalez, de A.L.C., W.D. Gonzalez, S.L.G. Dutra and B.T. Tsurutani, Periodic variation in geomagnetic activity: A study based on the Ap index, *J. Geophys. Res.*, **98**, 9215–9231, 1993.

Goodrich, C.C., J.G. Lyon, M. Wiltberger, R.E. Lopez and K. Papadopoulos, An overview of the impact of the January 10–11, 1997 magnetic cloud on the magnetosphere via global MHD simulation, *Geophys. Res. Lett.*, **25**(14), 2537, 1998.

Gopalswamy, N., A. Lara, S. Yashiro, M. L. Kaiser and R. A. Howard, Predicting the 1-AU arrival times of coronal mass ejections, *J. Geophys. Res.*, **106**, 29207, 2001.

Gopalswamy, N., Lara, A., Yashiro, S., Nunes, S. and Howard, R.A., Coronal mass ejection activity during solar cycle 23, in *Solar Variability as an Input to the Earth's Environment*, ESA SP-535, pp. 403–414, 2003.

Gopalswamy, N., H. Xie, S. Yashiro and I.Usoskin, I. Coronal mass ejections and ground level enhancements, *Proc. 29th International Cosmic Ray Conference*, Vol. 1, pp. 169–173, 2005.

Gosling, J.T., Coronal mass ejections and magnetic flux ropes in interplanetary space, in *Physics of Magnetic Flux Ropes*, E.R. Priest, L.C. Lee and C.T. Russell (eds.), AGU Geophys. Monogr. 58, p. 343, 1990.

Gosling, J.T., The solar flare myth, *J. Geophys. Res.*, **98**, A11, 18937, 1993a.

Gosling, J.T., Coronal mass ejections: The link between solar and geomagnetic activity, *Phys. Fluids*, **B5**, 2638, 1993b.

Gosling, J.T. and D.J. McComas, Field line draping about fast coronal mass ejecta: A source of strong out-of-the-ecliptic interplanetary magnetic fields, *Geophys. Res. Lett.*, **14**, 355–358, 1987.

Gosling, J.T. and V. Pizzo, Formation and evolution of corotating interaction regions and their three-dimensional structure, *Space Sci. Rev.*, **89**(1–2), 21, 1999.

Gosling, J.T., J.R. Asbridge, S.J. Bame, A.J. Hundhausen and I.B. Strong, Satellite observations of interplanetary shock waves, *J. Geophys. Res.*, **73**, 43, 1968.

Gosling, J.T., S.J. Bame, D.J. McComas and J.L. Phillips, Coronal mass ejections and large geomagnetic storms, *Geophys. Res. Lett.*, **17**, 901–904, 1990.

Gosling, J.T., J. Birn and M. Hesse, Three-dimensional reconnection and the magnetic topology of coronal mass ejection events, *Geophys. Res. Lett.*, **22**, 869–872, 1995.

Gringauz, K.I., V.V. Bezrukikh and V.D. Ozerov, Results of measurements of the concentration of positive ions in the atmosphere, using ion traps mounted on the Third Soviet Earth Satellite, in *Artificial Earth Satellites*, L.V. Kurnosova (ed.), New York, Plenum Press, Vol. 6, p. 77, 1961.

Harrison, R.A., Solar coronal mass ejections and flares, *Astron. and Astrophys.*, **162**, 283, 1986.

Henke, T., J. Woch, U. Mall, S. Livi, B. Wilken, R. Schwenn, G. Gloeckler, R. von Steiger, R.J. Forsyth and A. Balogh, Differences in the $O+7/O+6$ ratio of magnetic cloud and noncloud coronal mass ejections, *Geophys. Res. Lett.*, **25**, 3465, 1998.

Henke, T., J. Woch, R. Schwenn, U. Mall, G. Gloeckler, R. von Steiger, R.J. Forsyth and A. Balogh, Ionization state and magnetic topology of coronal mass ejections, *J. Geophys. Res.*, **106**, 597, 2001.

Hildner, E., J.T. Gosling, R.M. MacQueen, R.H. Munro, A.I. Poland and C.L. Ross, Frequency of coronal transients and solar activity, *Solar Physics*, **48**, 127–135, 1976.

Hoeksema, J.T., Structure and evolution of the large scale solar and heliospheric magnetic fields, PhD thesis, Stanford University, CA, 1984.

Howard, R.A., D.J. Michels, N.R. Sheeley and M.J. Koomen, The observation of a coronal transient directed at earth, *Astrophys. J.*, **263**, L101–L104, 1982.

Howard, R.A., G.E. Brueckner, O.C. St. Cyr, D.A. Biesecker, K.P. Dere, M.J. Koomen, C.M. Korendyke, P.L. Lamy, A. Llebaria, M.V. Bout *et al.*, Observations of CMEs from SOHO/LASCO, *Coronal Mass Ejections*, AGU Geophysical Monograph 99, p. 17, 1997.

Howard, R.A., A.F. Thernisien, A. Vourlidas, C. Marque and N. Patel, Modeling of CMEs for the STEREO Mission, *Proc. Solar Wind*, **11**, in press.

Hundhausen, A.J., C.B. Sawyer, L. House, R.M.E. Illing and W.J. Wagner, Coronal mass ejections observed during the Solar Maximum Mission: Latitude distribution and rate of occurrence, *J. Geophys. Res.*, **89**, 2639, 1984.

Huttunen, E.K., H.E. Koskinen and R. Schwenn, Variability of magnetospheric storms driven by different solar wind perturbations, *J. Geophys. Res.*, **107**, 10.1029/2001JA900171, 2002.

Huttunen, E.K., J. Slavin, M. Collier, H.E.J. Koskinen, A. Szabo, E. Tanskanen, A. Balogh, E. Lucek and H. Rème, Cluster observations of sudden impulses in the magnetotail caused by interplanetary shocks and pressure increases, *Ann. Geophys.*, **23**, 1–16, 2005.

Kivelson, M.G. and C.T. Russell (eds.), *Introduction to Space Physics*, Cambridge Univ. Press, ISBN 0521457149, 1995.

Klassen, A., H. Aurass, G. Mann *et al.*, *AASS*, **141**, 357, 2000.

Klassen, A., V. Bothmer, G. Mann, M. J. Reiner, S. Krucker, A. Vourlidas and H. Kunow, Solar energetic electron events and coronal shocks, *Astron. and Astrophys.*, **385**, 1078–1088, 2002.

Klein, L.W. and Burlaga, L.F., Interplanetary magnetic clouds at 1 AU, *J. Geophys. Res.*, **87**, 613–624, 1982.

Koomen, M., R. Howard, R. Hansen and S. Hansen, The coronal transient of 16 June 1972, *Solar Physics*, **34**, 447, 1974.

Koskinen, H., E. Tanskanen, R. Pirjola, A. Pulkinnen, C. Dyer, D. Rodgers, P. Cannon, J.-C. Mandeville and D. Boscher, *Space Weather Effects Catalogue*, Finish Meteorological Institute, ISSN 0782-6079, 2000.

Lang, K., *The Sun from Space*, Astronomy and Astrophysics Library, Springer-Verlag, 2000.

Lanzerotti, L.J. and T.R. Sanderson, Energetic particles in the heliosphere, in *The Heliosphere Near Solar Minimum: The Ulysses Perspective*, A. Balogh, R.G. Marsden and E.J. Smith (eds.), Springer-Praxis, pp. 259–286, 2001.

Lanzerotti, L.J., D.J. Thomson and C.G. Maclennan, Wireless at high altitudes: Environmental effects on space-based assets, *Bell Labs Technical Journal*, **5**, Summer 1997.

Lanzerotti, L.J., C. Breglia, D.W. Maurer, G.K. Johnson and C.G. Maclennan, Studies of spacecraft charging on a geosynchronous telecommunication satellite, *Adv. Space Res.*, **22**(1), 79, 1998.

Lepping, R.P., J.A. Jones, and L.F. Burlaga, Magnetic field structure of interplanetary magnetic clouds at 1 au, *J. Geophys. Res.*, **95**, 11957–11965, 1991.

Lepping, R.P., D.B. Berdichevsky, L.F. Burlaga, A.J. Lazarus, J. Kasper, M.D. Desch, C.-C. Wu, D.V. Reames, H.J. Singer, C.W. Smith *et al.*, The Bastille Day magnetic cloud and upstrem shocks: Near-earth interplanetary observations, *Solar Physics*, **204**, 287–305, 2001.

Lepping, R.P., D.B. Berdichevsky and C.C. Wu, Sun–Earth electrodynamics: The solar wind connection, *Recent Res. Devel. Astrophys.*, **1**, 139, 2003.

Linker, J.A., Z. Mikic, D.A. Biesecker, R.J. Forsyth, S.E. Gibson, A.J. Lazarus, A. Lecinski, P. Riley, A. Szabo and B.J. Thompson, Magnetohydrodynamic modeling of the solar corona during whole Sun month, *J. Geophys. Res.*, **104**, 9809–9830, 1999.

Linker, J.A., Z. Mikic, R. Lionello and P. Riley, Flux cancellation and coronal mass ejections, *Phys. of Plasmas*, **10**(5), 1971–1978, 2003.

Lites, B.W. and B.C. Low, Flux emergence and prominences: A new scenario for 3-dimensional field geometry based on observations with the advanced stokes polarimeter, *Solar Phys.*, **174**, 91, 1997.

Low, B.C., Coronal mass ejections, magnetic flux ropes, and solar magnetism, *J. Geophys. Res.*, **106**(A11), 25141, 2001.

Low, B.C. and M. Zhang, *Global Magnetic-Field Reversal in the Corona*, AGU Geophys. Monogr. 141, p. 51, 2004.

Luhmann, J.G., D.W. Curtis, R.P. Lin, D. Larson, P. Schroeder, A. Cummings, R.A. Mewaldt, E.C. Stone, A. Davis, T. von Rosenvinge *et al.*, IMPACT: Science goals and firsts with STEREO, *Adv. Space Res.*, **36**, 1534–1543, 2005.

Lyot, B., The study of the solar corona and prominences without eclipses, George Darwin Lecture, *Mon. Not. Roy. Astron. Soc.*, **99**, 580, 1939.

Marsden, R.G., The 3-D heliosphere at solar maximum, *Space Sci. Rev.*, **97**, 1–4, Kluwer Academic, reprint, 2001.

Marubashi, K., Structure of the interplanetary magnetic clouds and their solar origins, *Adv. Space Res.*, **6**, 335–338, 1986.

Marubashi, K., *Interplanetary Magnetic Flux Ropes and Solar Filaments*, AGU Geophys. Monogr. 99, pp. 147–156, 1997.

Marubashi, K., Interplanetary magnetic flux ropes, *J. of the Communications Research Laboratory*, **49**, 3, 2002.

Mayaud, P.N., *Derivation, Meaning, and Use of Geomagnetic Indices*, AGU Geophysical Monogr. 22, 1980.

McComas, D.J., J.T. Gosling, D. Winterhalter and E.J. Smith, Interplanetary magnetic field draping about fast coronal mass ejecta in the outer heliosphere, *J. Geophys. Res.*, **93**, 2519, 1988.

McComas, D.J., S.J. Bame, P. Barker, W.C. Feldman, J.L. Phillips, P. Riley and J.W. Griffee, Solar Wind Electron Proton Alpha Monitor (SWEPAM) for the Advanced Composition Explorer, *Space Sci. Rev.*, **86**, 563–612, 1998.

McPherron, R.L., Magnetospheric substorms, *Rev. Geophys. Space Phys.*, **17**, 657, 1979.

Mikic, Z., J.A. Linker, D.D. Schnack, R. Lionello and A. Tarditi, Magnetohydrodynamic modeling of the global solar corona, *Phys. Plasmas*, **6**, 2217–2224, 1999.

Mulligan, T., C.T. Russell and J.G. Luhmann, Solar cycle evolution of the structure of magnetic clouds in the inner heliosphere, *Geophys. Res. Lett.*, **25**, 2959, 1998.

Mulligan, T. and C.T. Russell, Multispaceraft modeling of the flux rope structure of interplanetary coronal mass ejections: Cylindrically symmetric versus nonsymmetric topologies, *J. Geophys. Res.*, **106**, 10581, 2001.

Ness, N.F., Interplanetary magnetic field dynamics, in *Space Storms and Space Weather Hazards*, I.A. Daglis (ed.), Kluwer Academic, Vol. 131, 2001.

Neugebauer, M. and C.W. Snyder, Mariner 2 observations of the solar wind, 1: Average properties, *J. Geophys. Res.*, **71**, 1966.

Nicolson, I., *Die Sonne*, Herder-Verlag, Freiburg, Germany, 1982.

O'Brien, T.P. and R.L. McPherron, An empirical pase space analysis of ring current dynamics: Solar wind control of injection and decay, *J. Geophys. Res.*, **105**, 7707–7719, 2000.

Odstrcil, D., Modeling 3D solar wind structure, *Adv. Space Res.*, **32**(4), 497, 2003.

Odstrcil, D., J.A. Linker, R. Lionello, Z. Mikic, P. Riley, V.J. Pizzo and J.G. Luhmann, Merging of coronal and heliospheric numerical two-dimensional MHD models, *J. Geophys. Res.*, **107**, doi:10.1029/2002JA009334, 2002.

Panasenco, O., I.S. Veselovsky, A.V. Dmitriev, A.N. Zhukov, O.S. Yakovchouk, I.A. Zhitnik, A.P. Ignat'ev, S.V. Kuzin, A.A. Pertsov, V.A. Slemzin *et al.*, Solar origins of intense geomagnetic storms in 2002 as seen by the CORONAS-F satellite, *Adv. Space Res.*, **36**(8), 1595–1603, 2005.

Parker, E.N., Extension of the solar corona into interplanetary space, *J. Geophys. Res.*, **64**, 1675, 1959.

Parks, G.K., *Physics of Space Plasmas: An Introduction*, ISBN 0813341299, Westview Press, 2003.

Phillips, J.L., J.T. Gosling and D.J. McComas, Coronal mass ejections and geomagnetic storms: Seasonal variations, *Proc. Solar–Terrestrial Predictions Workshop, Ottawa*, 1992.

Priester, W. and O. Cattani, On semiannual variation of geomagnetic activity and its relation to solar corpuscular radiation, *J. Atm. Sci.*, **19**, 121–126, 1962.

Prölss, G., *Physik des erdnahen Weltraums: Eine Einführung*, Springer-Verlag, Berlin, 2004.

Reames, D.V., Particle acceleration by CME-driven shock waves, Highlight Paper, *26th Int. Cosmic Ray Conf.*, 1999.

Richardson, I.G., The formation of CIRs at stream–stream interfaces and resultant geomagnetic activity, in *Recurrent Magnetic Storms: Corotating Solar Wind Streams*, B. Tsurutani, R. McPherron, W. Gonzalez, G. Lu, J.H.A. Sobral and N. Gopalswamy (eds.), AGU Geophys. Monogr. 167, in press, 2006.

Richardson, I.G., E.W. Cliver and H.V. Cane, Sources of geomagnetic storms for solar minimum and maximum conditions during 1972–2000, *Geophys. Res. Lett.*, **28**, 2569–2572, 2001.

Richardson, I.G., D.F. Webb, J. Zhang, D.B. Berdichevsky, D.A. Biesecker, J.C. Kasper, R. Kataoka, J.T. Steinberg, B.J. Thompson, C.-C. Wu *et al.*, Major geomagnetic storms (Dst ≤ -100 nT) generated by corotating interaction regions, *J. Geophys. Res.*, **111**, A07S09, doi:10.1029/2005JA011476, 2006.

Richardson, J.D., Y. Liu, C. Wang and L.F. Burlaga, ICMEs at very large distances, *Adv. Space Res.*, in press, 2006.

Rosenberg, R.L. and P.J. Coleman, Solar cycle-dependent north–south field configurations observed in solar wind interaction regions, *J. Geophys. Res.*, **85**, 3021–3032, 1980.

Rostoker, G., Geomagnetic indices, *Reviews of Geophysics and Space Physics*, **10**, 935, 1972.

Russell, C.T., Geophysical coordinate transformations, *Cosmic Electrodynamics*, **2**, 184, 1971.

Russell, C.T. and R.L. McPherron, The magnetotail and substorms, *Space Sci. Rev.*, **15**, 205, 1973a.

Russell, C.T. and R.L. McPherron, Semiannual variation of geomagnetic activity, *J. Geophys. Res.*, **78**, 92, 1973b.

Russell, C.T., T. Mulligan and B.J. Anderson, Radial variation of magnetic flux ropes: Case studies with ACE and NEAR, *Solar Wind Ten: Proc. of the 10th Int. Solar Wind Conf.*, M. Velli, R. Bruno and F. Malara (eds.), American Institute of Physics, p. 121, 2003.

Scherrer, P.H., R.S. Bogart, R.I. Bush, J.T. Hoeksema, A.G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T.D. Tarbell, A. Title *et al.*, The solar oscillations investigation: Michelson Doppler Imager. *Solar Physics*, **162**, 129–188, 1995.

Schlegel, K., Wenn die Sonne verrückt spielt, *Physik in unserer Zeit*, **31**(5), 222, 2000.

Schreiber, H., On the periodic variations of geomagnetic activity indices Ap and ap, *Ann. Geophys.*, **16**, 510, 1998.

Schrijver, C.J., Catastrophic cooling and high-speed downflow in quiescent solar coronal loops observed with TRACE, *Solar Physics*, **198**, 325, 2001.

Schrijver, C.J. and M.L. DeRosa, Photospheric and heliospheric magnetic fields, *Solar Physics*, **212**, 165, 2003.

Schwenn, R., Transport of energy and mass to the outer boundary of the Earth system, in *Solar–Terrestrial Energy Program: Major Scientific Problems, Proceedings of a SCOSTEP symposium held during the XXVII COSPAR Plenary Meeting, Helsinki, University of Technology, Espoo, Finland*, pp. 13–30, 1988.

Schwenn, R., Large-scale structure of the interplanetary medium, *Physics of the Inner Heliosphere, 1: Large-Scale Phenomena*, Springer-Verlag, p. 99, 1990.

Schwenn, R., Variability of magnetospheric storms driven by different solar wind perturbations, *J. Geophys. Res.*, **107**, 10.1029/2001JA900171, 2002.

Schwenn, R., Space weather: The solar perspective, *Living Rev. Solar Physics*, **3**, 2, *http://www.livingreviews.org/lrsp-2006-2*, 2006.

Schwenn, R. and E. Marsch (eds.), *Physics of the Inner Heliosphere, 1: Large-Scale Phenomena*, Springer-Verlag, 1990.

Schwenn, R. and E. Marsch (eds.), *Physics of the Inner Heliosphere. 2: Particles, Waves and Turbulence*, Springer-Verlag, 1991.

Schwenn, R., A. dal Lago, E. Huttunen and W.D. Gonzalez, The association of coronal mass ejections with the effects of their counterparts near the Earth, *Ann. Geophys.*, **23**, 1033–1059, 2005.

Sheeley Jr., N.R., R.A. Howard, M.J. Koomen, D.J. Michels, R. Schwenn, K.-H. Mühlhäuser and H. Rosenbauer, Coronal mass ejections and interplanetary shocks, *J. Geophys. Res.*, **90**, 163–175, 1985.

Siebert, M., Maßzahlen der erdmagnetischen Aktivität, *Handbuch Physik*, Vol. 49/3 (Geophysik 3/3), pp. 206–275, Springer-Verlag, 1971.

Smith, C.W., J. L'Heureux, N.F. Ness, M.H. Acuna, L.F. Burlaga and J., Scheifele, The ACE magnetic field experiment, *Space Sci. Rev.*, **86**, 613, 1998.

Sonett, C.P. and I.J. Abrahms, The distant geomagnetic field, 3: Disorder and shocks in the magnetopause, *J. Geophys. Res.*, **68**, 1233, 1963.

St. Cyr, O.C., J.T. Burkepile, A.J. Hundhausen and A.R. Lecinski, A comparison of ground-based and spacecraft observations of coronal mass ejections from 1980–1989, *J. Geophys. Res.*, **104**, 12493, 1999.

St. Cyr, O., 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.*, Properties of coronal mass ejections: SOHO, LASCO observations from January 1996 to June 1998, *J. Geophys. Res.*, **105**, 18169, 2000.

Stix, M., *The Sun:An Introduction*, 2nd edition, Astronomy and Astrophysics Library, Springer-Verlag, 2004.

Subramanian, P. and K.P. Dere, Source regions of coronal mass ejections, *Astrophys. J.*, **561**, 372, 2001.

Tang, F., B.T. Tsurutani, W.D. Gonzalez, S.I. Akasofu and E.J. Smith, Solar sources of interplanetary southward Bz events responsible for major geomagnetic storms (1978–1979), *J. Geophys. Res.*, **94**, 3535, 1989.

Thompson, B.J., Moreton waves, in *Encyclopedia of Astronomy and Astrophysics*, P. Murdin (ed.), article 2575, Institute of Physics, Bristol, UK, *http://eaa.iop.org/abstract/0333750888/2575*, 2000.

Tsurutani, B.T., Solar/interplanetary plasma phenomena causing geomagnetic activity at Earth, *Proceedings of the International School of Physics 'Enrico Fermi'*, B. Coppi, A. Ferrari and E. Sindoni (eds.), IOS Press, Amsterdam, pp. 273–286, 2000.

Tsurutani, B.T., The interplanetary causes of magnetic storms, substorms and geomagnetic quiet, in *Space Storms and Space Weather Hazards*, I.A. Daglis (ed.), Kluwer Academic, p. 103, 2001.

Tsurutani, B.T. and W.D. Gonzalez, The cause of high intensity long-duration continuous AE activity (HILDCAAs): Interplanetary Alfvén wave trains, *Planet. Space Sci.*, **35**, 405, 1987.

Tsurutani, B.T. and W.D. Gonzalez, The interplanetary causes of magnetic storms: A review, in *Magnetic Storms*, B.T. Tsurutani, W.D. Gonzalez, Y. Kamide and J.K. Arballo (eds.), AGU Geophys. Monogr. 98, pp. 77–89, 1997.

Tsurutani, B.T., W.D. Gonzalez, F. Tang and Y.T. Lee, Great magnetic storms, *Geophys. Res. Lett.*, **19**, 73, 1992.

Tsurutani, B.T., W.D. Gonzalez, G.S. Lakhina and S. Alex, *The Extreme Magnetic Storm of September 1–2, 1859*.

Tsurutani, B.T., W.D. Gonzalez, X.-Y. Zhou, R.P. Lepping and V. Bothmer, Properties of slow magnetic clouds, *J. of Atm. and Sol.–Terr. Phys.*, **66**, 147–151, 2004.

Tsurutani, B.T., W.D. Gonzalez, A.L.C. Gonzalez, F.L. Guarnieri, N. Gopalswamy, M. Grande, Y. Kamide, Y. Kasahara, G. Lu, I. Mann *et al.*, Corotating solar wind streams and recurrent geomagnetic activity: A review, *J. Geophys. Res.*, **111**, A07S01, doi:10.1029/2005JA011273, 2006.

Tylka, A.J., The historic solar event of January 20, 2005: A challenge for current theories on the origin of solar energetic particles?, *SHINE 2006 Workshop, 31 July–4 August, Zermatt, Utah*, *http://creme96.nrl.navy.mil/20Jan05/*, 2006.

Uchida, Y. *et al.* (eds.), *X-ray Solar Physics from Yohkoh*. University Academy Press, Tokyo, 1994.

Veselovsky, I.S., M.I. Panasyuk, S.I. Avdyushin, G.A. Bazilevskaya, A.V. Belov, S.A. Bogachev, V.M. Bogod, A.V. Bogomolov, V. Bothmer, K.A. Boyarchuk *et al.*, Solar and heliospheric phenomena in October/November 2003: Causes and effects, *Kosmicheskie Issledovaniya*, **42**(5), 453–508, 2004; translated in *Cosmic Research*, **42**(5), 435–488, 2004.

Veselovsky, I.S., V. Bothmer, P. Cargill, A.V. Dmitriev, K.G. Ivanov, E.R. Romashets, A.N. Zhukov and O.S. Yakovchouk, Magnetic storm cessation during northward IMF, *Adv. Space Res.*, **36**, 2460, 2005.

Vourlidas, A., D. Buzasi, R.A. Howard and E. Esfandiari, Mass and energy properties of LASCO CMEs, in *Solar Variability: From Core to Outer Frontiers*, ESA SP-506, p. 9194, 2002.

Wang, C.B., J.K. Chao and C.-H. Lin, Influence of the solar wind dynamic pressure on the decay and injection of the ring current, *J. Geophys. Res.*, **108**, doi:10.1029/2003JA009851, SMP 5-1/5-10, 2003.

Wang, Y.M., *Astrophys. J.*, **543**, L89, 2000.

Wang, Y.M. and N.R. Sheeley Jr., Average properties of bipolar magnetic regions during sunspot cycle 21, *Solar Phys.*, **124**, 81, 1989.

Wang, Y.M. and N.R. Sheeley Jr., On potential-field models of the solar corona, *Astrophys. J.*, **392**, 310, 1992.

Wang, Y.M. and N.R. Sheeley Jr., Filament eruptions near emerging bipoles, *Astrophys. J.*, **510**, L157, 1999.

Wang, Y.M., A.G. Nash and N.R. Sheeley Jr., Evolution of the sun's polar fields during sunspot cycle 21: Poleward surges and long-term behavior, *Astrophys. J.*, **347**, 529, 1989.

Wang, Y.M., S.H. Hawley and N.R. Sheeley Jr., The magnetic nature of coronal holes, *Science*, **271**, 464, 1996.

Wang, Y.M., N.R. Sheeley, J.H. Walters, G.E. Brueckner, R.A. Howard, D.J. Michels, P.L. Lamy, R. Schwenn and G.M. Simnett, Origin of streamer material in the outer corona, *Astrophys. J.*, **498**, L165, 1998.

Wang, Y.M., N.R. Sheeley Jr. and M.D. Andrews, Polarity reversal of the solar magnetic field during cycle 23, *J. Geophys. Res.*, **107**, A12, 1465, doi:10.1029/2002JA009463, 2002.

Webb, D.F. and A.J. Hundhausen, Activity associated with the solar origin of coronal mass ejections, *Sol. Phys.*, **108**, 383–401, 1987.

Webb., D., E.W. Cliver, N.U. Crooker, O.C. St. Cyr and B.J. Thompson, Relationship of halo coronal mass ejections, magnetic clouds, and magnetic storms, *J. Geophys. Res.*, **105**, 7491, 2000.

Wilson, R.M. and E. Hildner, On the association of magnetic clouds with disappearing filaments, *J. Geophys. Res.*, **91**, 5867–5872 , 1986.

Yashiro, S., N. Gopalswamy, G. Michalek, O.C. St. Cyr, S.P. Plunkett, N.B. Rich and R.A. Howard, A catalog of white light coronal mass ejections observed by the SOHO spacecraft, *J. Geophys. Res.*, **7105**, 2004.

Yurchyshyn, V.B., H. Wang, P.R. Goode and Y. Deng, Orientation of the magnetic fields in interplanetary flux ropes and solar filaments, *Astrophys. J.*, **563**, 381, 2001.

Zhang, G. and L.F. Burlaga, Magnetic clouds, geomagnetic disturbances and cosmic ray decreases, *J. Geophys. Res.*, **93**, 2511, 1988.

Zhang, J., K.P. Dere, R.A. Howard, M.R. Kundu and S.M. White, On the temporal relationship between coronal mass ejections and flares, *Astrophys. J.*, **559**, 452, 2001.

Zhang, J., K.P. Dere, R.A. Howard and V. Bothmer, Identification of solar sources of major geomagnetic storms between 1996 and 2000, *Astrophys. J.*, **582**, 520, 2003.

Zhao, X.P. and D.F. Webb, Source regions and storm effectiveness of frontside full halo coronal mass ejections, *J. Geophys. Res.*, **108**, 1234, doi:10.1029/2002JA009606, 2003.

Zhukov, A.N., Solar sources of geoeffective CMEs: A SOHO/EIT view, in *Coronal and Stellar Mass Ejections, Proceedings IAU Symposium No. 226*, K.P. Dere, J. Wang and Y. Yan (eds.), Cambridge University Press, pp. 437–447, 2005.

Zhukov, A.N. and F. Auchere, On the nature of EIT waves, EUV dimmings and their link to CMEs, *Astron. and Astrophys.*, **427**, 705, 2004.

Zhukov, A.N., I.S. Veselovsky, F. Clette, J.-F. Hochedez, A.V. Dmitriev, E.P. Romashets, V. Bothmer and P. Cargill, Solar wind disturbances and their sources in the EUV solar corona, in *Solar Wind Ten: Proceedings of the Tenth International Solar Wind Conference, Pisa, Italy, 17–21 June 2002*, M. Velli, R. Bruno and F. Malara, AIP Conference Proceedings 679, Issue 1, p. 711, 2003.

4

The coupling of the solar wind to the Earth's magnetosphere

Christopher T. Russell

The supersonic solar wind interacts with the terrestrial magnetic field by forming a standing bow shock that slows, deflects and heats the plasma. In the post-shock subsonic flow, pressure gradients develop that further deflect the flow around the magnetic cavity or magnetosphere. While stresses applied by the solar wind along the normal to the magnetospheric boundary determine its size and shape, tangential stress determines how much energy flows into the magnetosphere. The temporal behaviour of the interplanetary magnetic field controls the temporal variation of the reconnection rate, and this in turn creates the magnetospheric event called a substorm. When the interplanetary magnetic field becomes steadily southward, such as during an event known as an interplanetary coronal mass ejection, the plasma in the magnetosphere can be energized to quite high energies, in what is known as a magnetospheric storm. The stresses in the outer magnetosphere are coupled to the ionosphere and atmosphere via field-aligned current systems that apply drag to the ionospheric plasma via the Lorentz or $\mathbf{J} \times \mathbf{B}$ force as the currents pass across auroral magnetic field lines in the resistive ionosphere. Thus the magnetosphere acts both as a shield against the solar wind and as a net that gathers solar wind momentum flux and stirs the magnetospheric plasma.

4.1 INTRODUCTION

As our technological society has advanced, we have made greater and greater use of the Earth's magnetosphere as the host for increasingly sophisticated and critical components of our terrestrial monitoring programme, communication systems, and global positioning satellites. The magnetosphere is a fairly benign host, but it can be energized greatly by the solar wind, and change into a quite inhospitable environment, affecting not just systems in space, but even those in the lower atmosphere and on the surface of the Earth. Fortunately, we have learned much about the magnetosphere and its control by the solar wind since the earliest days when Van Allen and his coworkers discovered that space was 'radioactive'. The solar wind can control both the size of the magnetosphere and the energy flow into the magnetosphere. If we know, or can accurately predict, the plasma conditions in the solar wind ahead of the Earth, we can now predict at least the statistical behaviour of the magnetosphere.

The Earth's magnetosphere is, to a zeroth-order approximation, a giant magnetic cavity that shields the Earth from the flowing solar wind. The size and shape of this cavity is controlled by the three components of the solar wind pressure: the dynamic pressure or momentum flux of the cold stream of ions flowing radially from the Sun; the kinetic or thermal pressure of the plasma that is measured in the frame of the solar wind flow; and the pressure of the interplanetary magnetic field. The dynamic pressure is ρu^2, where ρ is the mass density (m^{-3}) dominated by the ions in the solar wind and u is the bulk velocity (m/s). The thermal pressure has two components – one perpendicular and one parallel to the magnetic field. The perpendicular pressure is $nkT_\perp$ summed over each species where n is the number density (m^{-3}) and $T_\perp$ is the temperature of the particles determined from their motion perpendicular to the magnetic field. Here $kT_\perp$ equals $\frac{1}{2}mv_\perp^2$, and $v_\perp$ is the root mean square (rms) velocity of a maxwellian distribution of gyro velocities. The parallel pressure is $nkT_\parallel$, where $kT_\parallel$ equals $\frac{1}{2}mv_\parallel^2$ and $v_\parallel$ is the rms velocity of a maxwellian distribution of parallel velocities. The magnetic field does not exert a force along its length, so we need consider only the pressure perpendicular to the magnetic field, $B^2/2\mu_o$, where B is measured in Teslas and μ_o is $4\pi \times 10^{-7}$. These pressures act along the normal to the magnetopause.

These three components of pressure have quite different magnitudes. If we take the ratio of the dynamic to thermal pressure we obtain the square of the sonic Mach number times the ratio of specific heats – a number that may be of the order of 100 in the solar wind at the orbit of Earth, 1 AU from the Sun. If we calculate the ratio of dynamic pressure to magnetic pressure we obtain twice the square of the Alfvénic Mach number – also a number that may be of the order of 100. Thus, when considering the forces that determine the zeroth order size of the magnetosphere – that is, the location of the subsolar magnetopause – we need only consider the force of the dynamic pressure along the normal.

In addition to acting along the normal to the magnetopause, forces can act tangential to the surface. These tangential stresses can transfer momentum and energy across the magnetopause. There are several ways in which these tangential stresses could be exerted. Advocacy for the primacy of each of these mechanisms has waxed and waned over the years. Since the particles flowing by the magnetopause have most of their momentum in the direction radially outward from the Sun, one way to transfer momentum would be to move solar wind particles to the inside of the magnetosphere across the magnetopause. Since the gyro radius of the ions (the particles with most of the momentum) is small on the scale of the interaction, the scattering of gyrating ions close to the magnetopause is an obvious candidate for this transfer. However, very little scattering takes place at the magnetopause.

Boundary layers once thought to be caused by diffusion are now attributed to the process, known as reconnection.

Waves can also cause momentum transfer across the magnetopause if the magnetosphere is dissipative. Consider first surface waves on the boundary of a magnetized plasma with no dissipation. Eddies will be set up in the interior of the plasma. In a non-dissipitative medium the eddies would not drop off in amplitude with distance from the oscillating boundary, but in a dissipative medium the velocities in the eddies become smaller with distance from the boundary. As a result the average velocity, inside and near the boundary, flows in the same direction as the flow outside the boundary that caused the wave, transferring momentum from the flow outside the boundary to inside the boundary. This process appears to play some role in momentum transfer in the magnetosphere, but observations indicate that a third process is yet more important in this regard.

This third process – reconnection – involves the linkage of magnetic field lines across the magnetopause. The interplanetary magnetic field and the magnetospheric magnetic field have quite different sources and, in the presence of a perfectly collisionless plasma, one would not expect these magnetic fields to become linked; but indeed they do. The rate at which they become linked depends in part on the relative directions of the magnetic field on the two sides of the boundary between them. When they are antiparallel the rate of reconnection is highest. Since the magnetic field of the Earth produces a three-dimensional array of magnetic field lines that point in all directions, the interaction with the magnetic field carried by the solar wind always produces at least a small region of antiparallel magnetic fields. How large a region of reconnection is produced around the antiparallel point is still an area of active study.

In summary: in order to predict the strength of the coupling of the solar wind to the magnetosphere we need to monitor the solar wind mass density and bulk velocity, and its magnetic field strength and direction. As noted above, the solar wind dynamic pressure far exceeds the thermal and magnetic pressures, and, as a result, the relevant solar wind Mach numbers, the flow velocity of the solar wind relative to the Earth divided by the sound and Alfvén velocities, is much greater than unity. Since the solar wind is supersonic and superalfvénic, a strong shock must form for the solar wind to flow around the magnetosphere. Qualitatively the formation of a shock does not alter the above dependencies. However, quantitative changes may occur because the shocked plasma that flows by the magnetosphere can be quite different than the plasma upstream of the shock. The differences between pre- and post-shock solar wind are controlled by the Mach number of the shock and the direction of the interplanetary magnetic field relative to the shock normal. Thus it may be important to introduce the Mach number of the shock and the shock normal angle of the field into space weather predictions. To date, this added sophistication is not generally included.

In this chapter we discuss the formation of the shock and how it changes the properties of the solar wind. Then we examine the effect of dynamic pressure on the magnetosphere and the effect of changing magnetic fields. Finally we will discuss how the energy is stored in the magnetosphere and released, where the energy goes after release, and how the magnetosphere couples to the ionosphere and atmosphere.

4.2 THE BOW SHOCK AND THE MAGNETOSHEATH

As noted above, the flow of the solar wind is supersonic and superalfvénic. As the solar wind passes the Earth it is moving faster than the speed of thermal ions, and the speed of any wave that might act on the solar wind flow to divert it around the Earth's magnetic obstacle. Any such diversion will be affected by the underlying kinetic nature of the plasma. While the magnetic and electric fields provide collective effects leading to fluid-like behaviour, underlying the fluid-like behaviour of large systems are the motions of the ions and electrons that orbit field lines at the gyro frequency and oscillate at the plasma frequency. Thus the gyro radius (dependent on the particle energy) and the plasma frequency (dependent on the number density) play an important role. This is particularly true at the bow shock, that we discuss herein, and in the magnetopause and tail current sheets, that we discuss in later sections. For very weak magnetic fields, as measured by the distance at which the dipole magnetic field strength is strong enough to stand off the solar wind, such that this 'stand-off' distance is much smaller than the ion inertial length or the gyro radius of the particles, no shock is formed. In fact, the pressure balance distance has to be much larger than the ion inertial length for the bow shock to arise. The shock exists through collective effects, such as the compressed magnetic field, and the electric field along the shock normal, and has fluid-like behaviour, even though underlying the shock processes are significant kinetic processes.

We can learn much about the solar wind interaction with the Earth's magnetic field using the fluid approach because the stand-off distance of the magnetopause is close to 600 times greater than the ion inertial length, which in turn is close to the ion gyro radius in solar wind conditions at 1 AU. In other words the scale lengths in the plasma, such as the radius of curvature of the shock and magnetopause surfaces, are much larger than the kinetic length scales of the particles that constitute the plasma. Along a stream tube the sum of the static pressure and the dynamic pressure are constant. The top panel of Figure 4.1 shows this in a case where there is an obstacle on the right-hand side of the page that reduces the bulk velocity of the gas to zero. The gradient in static pressure acts to slow the flow and to deflect it in a two- or three-dimensional situation. In this case the flow is subsonic. The static pressure exceeds the dynamic pressure and a sufficient gradient can exist in the static pressure to slow the flow. However, if the flow is supersonic, as in the case of the interaction of the solar wind with the Earth's magnetic field, this cannot occur, and we find ourselves in the situation sketched in the bottom panel of Figure 4.1, where the static pressure is much less than the dynamic pressure on the left-hand side of the box. The plasma solves this conundrum by forming a shock that converts dynamic pressure to static pressure by slowing and heating the gas. Then, downstream of the shock the static pressure gradient needed for the remaining slowdown and deflection can form.

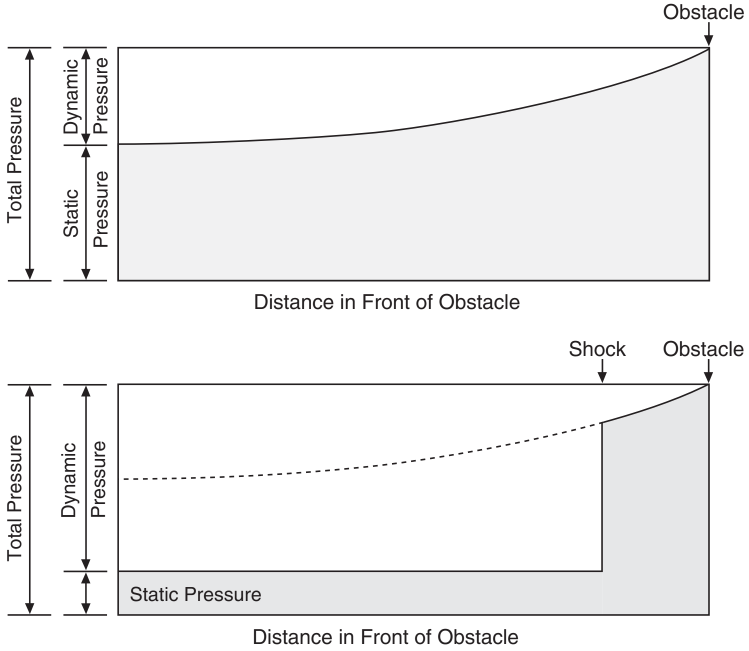


Figure 4.1. (Top) In subsonic flow the static pressure exceeds the dynamic pressure, allowing for the establishment of a pressure gradient that deflects the flow. (Bottom) In supersonic flow the static pressure is too weak to deflect the flow and a shock must form to heat slow and deflect the flow. Downstream from the shock the pressure gradient can deflect the flow.

The fluid or gas dynamic approach can be very instructive for understanding the solar wind interaction with the magnetosphere. This approach has been used extensively and most notably by Spreiter *et al.* (1966) who used a fixed obstacle, shaped like the magnetosphere, and computed numerically the flow around this three-dimensional (but cylindrically symmetric) obstacle for supersonic gas flows. In this calculation the magnetic field exerted no force and was treated as if it consisted of massless threads carried by the flow. Figure 4.2 shows the flow and field lines for such a simulation. The dashed flow lines are straight until they hit the shock, and then bend smoothly around the magnetosphere. The solution is cylindrically symmetric for the flow lines. Symmetry is broken by the magnetic field, but it has no influence on the flow in the gas-dynamic treatment. The thin solid lines, denoting the magnetic field lines, are also straight until they hit the shock, and then drape smoothly over the obstacle. This magnetic pattern is not cylindrically symmetric. The region shaded in Figure 4.2 is not well described by gas dynamics. Here ions and electrons are observed by our spacecraft to be streaming back from the shock, reflected by the potential barrier at the shock or leaking from the heated magnetosheath plasma behind the shock. We cannot begin to treat such phenomena with a gas dynamic approach.

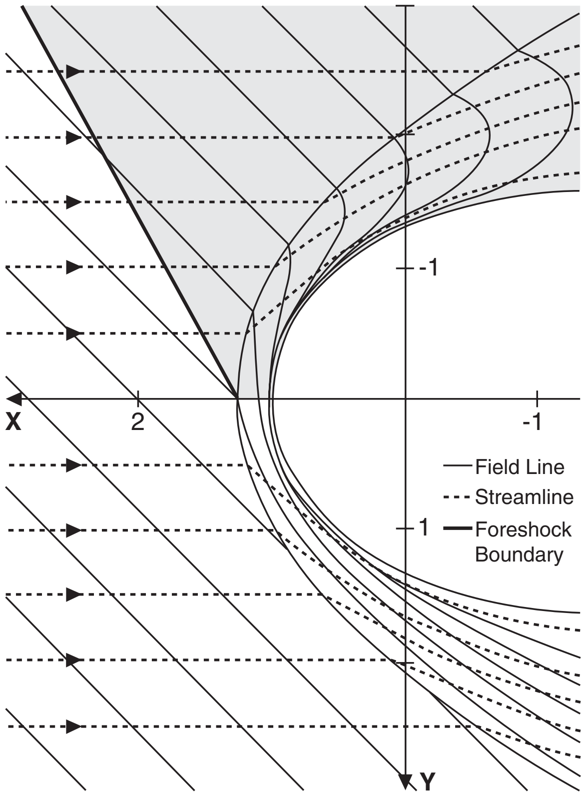


Figure 4.2. Dashed lines show the streamlines of the flow in Spreiter *et al.* (1966) convected field gas-dynamic simulation of the interaction of a fluid with a hard magnetospheric-like obstacle. The thin solid lines show the magnetic field that is carried along with the flow in this approximation. The heavy line denotes the foreshock boundary that cannot be found as a product of this simulation because ion reflection at the shock is a kinetic effect.

The distance from the magnetopause at which the shock forms is not arbitrary. The shocked plasma has to flow around the obstacle between the shock and the magnetopause, so the stand-off distance depends on how compressed the plasma can become, which in turn depends on the strength of the shock – its Mach number – and on the ratio of specific heats of the gas – the polytropic index. It also depends on the shape of the obstacle. A sharply pointed obstacle allows the shock to touch the obstacle. A blunt obstacle causes the shock to stand off. The shock is sensing the radius of curvature of the obstacle, and if the obstacle were an infinite flat plane, the shock would of course have to move away from the obstacle, and, in steady state, would move to infinity. If the obstacle were to be a good absorber and a poor deflector, such as the Moon, the shock would move into the obstacle, perhaps forming only near the limbs of the Moon where its weak magnetic field is sufficient to divert the flow.

Returning to the gas dynamic solution of Spreiter *et al.* (1966), we show, in Figure 4.3 the velocity and temperature contours (top) and the number density contours (bottom). The flow along subsolar streamline stops at the obstacle where the density builds up to 4.23 times that in the upstream solar wind. Away from the subsolar region the density decreases and the flow speed increases, most markedly along the magnetopause. Everywhere along the shock there is a significant jump in density, but along the flanks the enhancement drops quickly behind the shock. We note that in this treatment the downstream-to-upstream temperature ratio maps precisely to the downstream-to-upstream velocity ratio. While the temperature increase is very high in the subsolar region, it is much smaller along the flanks of the magnetosheath. In summary, the shock has slowed and deflected the solar wind so that the dynamic pressure in the solar wind that was so dominant has been replaced by static or thermal pressure along the magnetopause. This complication is not so important for determining the size and shape of the magnetopause, but it is very important for determining the rates of plasma processes at the magnetopause as the plasma conditions have become much different than they were in the solar wind, and this difference is controlled by the Mach number of the shock.

In high-Mach-number flow interactions the magnetic forces can often be ignored; but this is not a good approximation in the magnetosheath, especially near the magnetopause, so the convected-field gas-dynamic simulation can be misleading. Today we can treat the fluid interaction while including the forces due to the magnetic field, albeit still with somewhat less resolution than achieved by Spreiter *et al.* (1966) in their gas-dynamic treatment (in part because they were able to two-dimensionalize it). Figure 4.4 (see colour section) shows a modern MHD simulation (Wang *et al.*, 2004) of the solar wind interaction with now a compressible magnetosphere. Again the (white) streamlines are straight in the solar wind and bend around the obstacle post shock. Here we show the density as colour contours in each of the three panels. We immediately see one important difference from the gas-dynamic solution. The density at the magnetopause is reduced everywhere over that in the gas-dynamic solution. The arrows in the three panels show why this has occurred. In the gas-dynamic solution the pressure gradient that slowed and deflected the plasma could only be provided by the plasma, so the static pressure in the plasma had to maximize at the subsolar magnetopause. Here the magnetic force (middle panel) slows the flow and deflects it at the shock and throughout the magnetosheath. In contrast, the gradient in the static pressure works in the same direction at the shock, but is almost antiparallel to the magnetic force at the magnetopause. Summing the two forces we obtain the bottom panel, where the net forces are seen to be guiding the flow around the obstacle, slowing it as expected, but also lowering the plasma density along the magnetopause. We note that in this simulation the interplanetary magnetic field was northward. This minimizes, but does not eliminate, the effects of reconnection in the simulation.

In order to better appreciate the role of the shock in controlling conditions in the magnetosheath, we switch to an examination of the change of the plasma across a planar shock, the fluid (magnetohydrodynamic) solution of which is analytic. Thus we can examine, in Figure 4.5, how the magnetic field and density vary with Mach

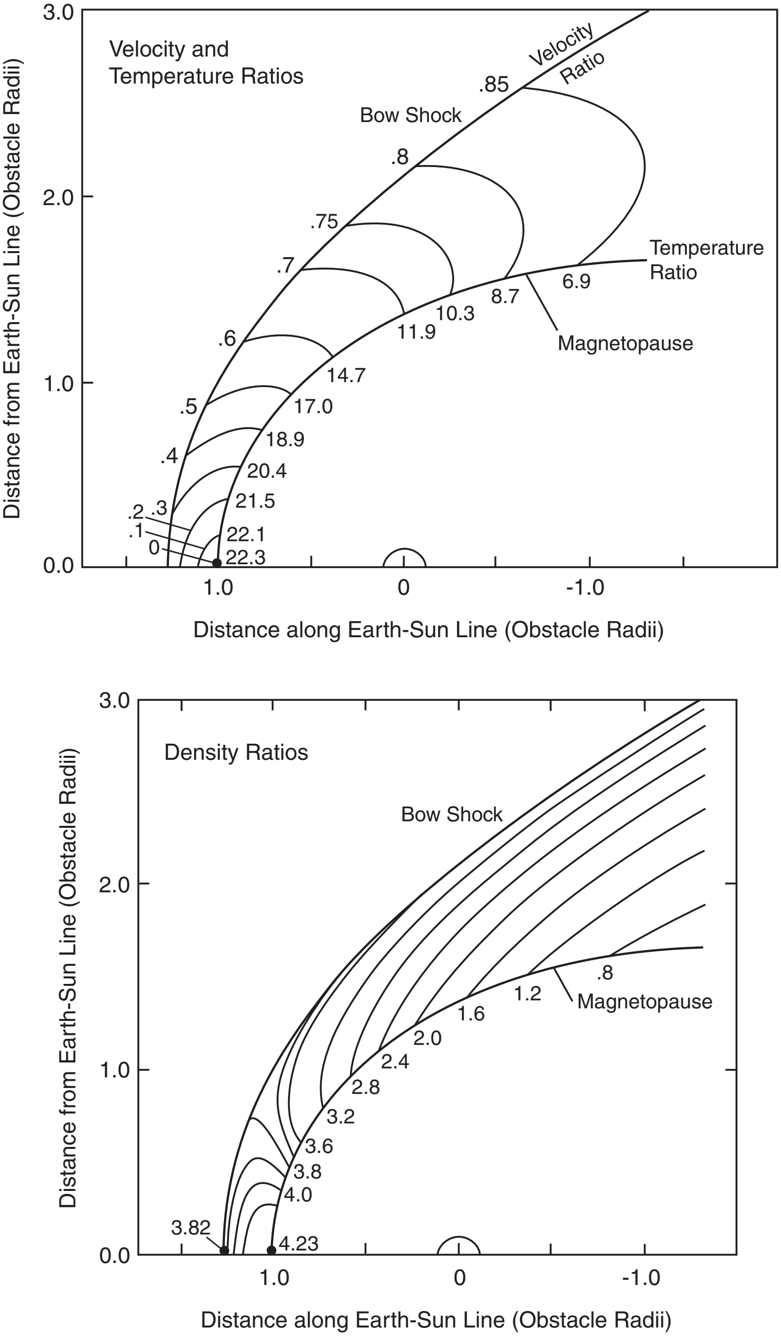


Figure 4.3. (Top) The velocity and temperature contours for the simulation shown in Figure 4.2. (Bottom) The density contours for the simulation shown in Figure 4.2.

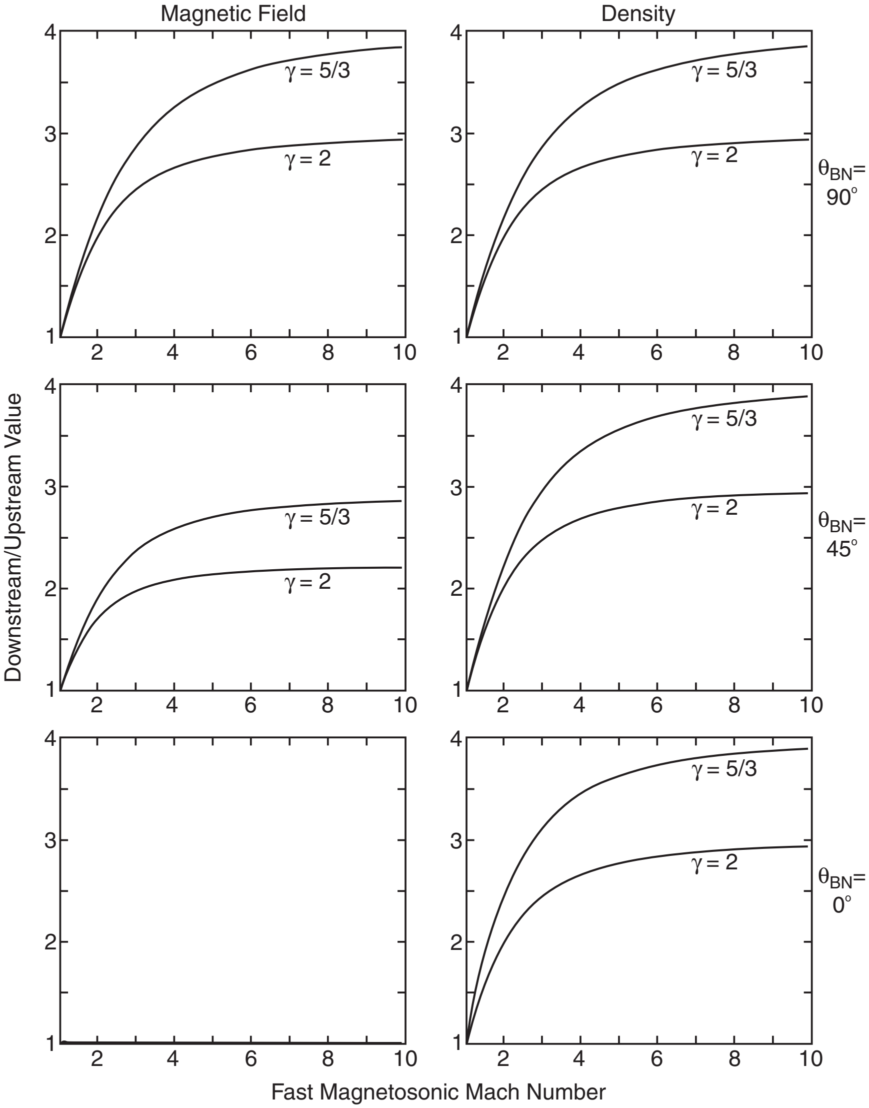


Figure 4.5. Rankine–Hugoniot solutions for the variation with Mach number of the ratio of the magnetic field downstream of the shock to that upstream (left panels) and the ratio of the density downstream to that upstream (right panels) for two ratios of specific heat (or polytropic indices) and at three different angles of the upstream magnetic field to the shock normal.

number and polytropic index for those directions of the upstream magnetic field to the shock normal (top, 90°; middle, 45°; bottom, 0°). The magnetic field jump is very sensitive to the direction of the magnetic field, but this is not true for density. We note that for a polytropic index of 5/3 (appropriate for an ideal gas with three degrees of freedom) the compression of the gas is greater than for a polytropic index of 2 (two degrees of freedom). The asymptotic limit for the compression is 4 and 3 for the density and for magnetic fields at 90° to the shock normal. In short, the shock controls plasma conditions at the magnetopause. To the extent that the coupling of the solar wind to the magnetopause depends on these conditions, we should consider the Mach number of the bow shock when studying this coupling.

The kinetic processes occurring at the bow shock are also extremely rich in interesting plasma physics leading to the production of plasma waves and energetic particles – notably those going back upstream into the solar wind. For the purpose of this chapter we will mention only one peculiarity of the shock kinetics that has some importance for the magnetosphere. In the solar wind the electron temperature is generally higher than the ion temperature because of the greater thermal conductivity of the electrons. (The electrons move faster than the solar wind bulk velocity). However, behind the shock the ions are hotter – generally about seven times hotter. The dynamic pressure drop has resulted in a major increase in the static pressure of the ions but not the electrons. This temperature ratio then stays roughly constant throughout the magnetosheath, plasmasheet and outer magnetosphere, providing a good marker for the source of these plasmas.

4.3 THE SIZE AND SHAPE OF THE MAGNETOSPHERE

The size of the magnetosphere is determined by the pressure applied to it by the solar wind, but, as we see above, the supersonic nature of the solar wind flow relative to the Earth requires the formation of a bow shock standing in front of the magnetosphere that alters the properties of the incoming flow. Fortunately there is a simplifying relation that enables us to ignore most of the complications imposed by the bow shock. This simplification is Bernoulli's law that states that momentum in a stream tube of varying cross section is conserved. Thus:

$$(\rho u^2 + nkT + B^2/2\mu_o)\,S = \text{constant}$$

where S is the cross-section of the stream tube, which in this case expands slightly as the flow moves from the shock to the magnetopause. At the subsolar point in a high Mach number this increase in cross-section results in a decrease in pressure of about 10%. The balancing force is the pressure of the terrestrial magnetic field. We know the magnetic field of a dipole at any distance from it, but the magnetic field of the compressed magnetosphere is stronger than the simple dipole field. This factor is dependent on the shape or radius of curvature of the magnetopause. If the magnetopause is a plane the enhancement factor would be 2, but for the observed shape the enhancement factor is about 2.4. Equating the two pressures we obtain the stand-off distance at the subsolar point:

$$L_{mp} = 107.4(n_{sw}\,u_{sw}^2)^{-1/6}$$

where we have expressed L_{mp} in Earth radii, n_{sw} in protons per cm^3 and u_{sw} in km/s.

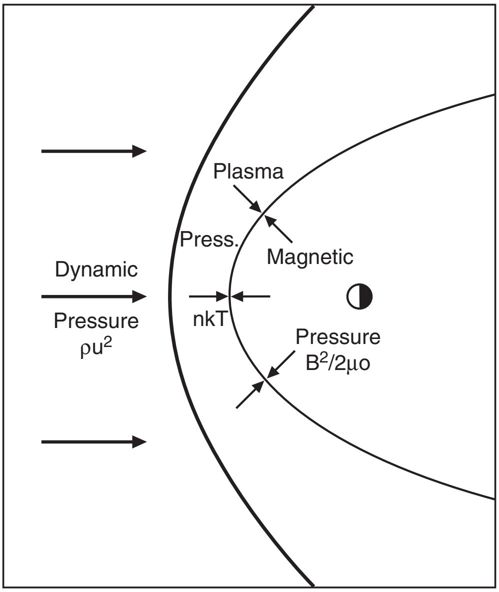


Figure 4.6. Schematic illustration of how solar wind dynamic pressure is converted to pressure along the magnetopause normal to balance the magnetic pressure of the magnetosphere.

Thus a solar wind density of 9.6 protons per cm^3 flowing at 400 km/s will produce a magnetopause distance of 10 Earth radii at the subsolar point.

If we move away from the subsolar point as illustrated in Figure 4.6, the direction of the normal varies, and since the size of the magnetosphere is determined by the normal stress, the velocity used in the momentum flux or dynamic pressure must be that resolved along the normal. If the angle between the normal to the magnetopause and the solar wind flow (upstream of the bow shock) is φ, then the pressure is $\rho_{sw}(u\cos\varphi)^2$. Eventually the normal to the magnetopause turns well away from the solar wind flow, so that despite the supersonic nature of the flow the thermal and magnetic pressures become comparable to or greater than the dynamic pressure. If we let the sum of their pressures be R, ignoring the fact that the magnetic field does not exert a force along its length, we can approximate a second component of the force along the magnetopause normal. This additional component is $R\sin^2\varphi$ (Petrinec and Russell, 1997). It provides the force that establishes the asymptotic width of the magnetotail.

Figure 4.7 illustrates how the plasma pressure in the magnetosheath balances the magnetic pressure in the magnetosphere and forms a current layer, the magnetopause, which has finite thickness. Within the magnetopause, the magnetic field pressure drops while the plasma pressure increases. The sum of the two pressures remains constant and the pressure gradients exert balancing forces. In the region of the gradient in the thermal pressure the net current is in the direction to change the magnetic field as required, and the integrated Lorentz, or $\mathbf{J}\times\mathbf{B}$ force, is that required to hold back the pressure force of the plasma.

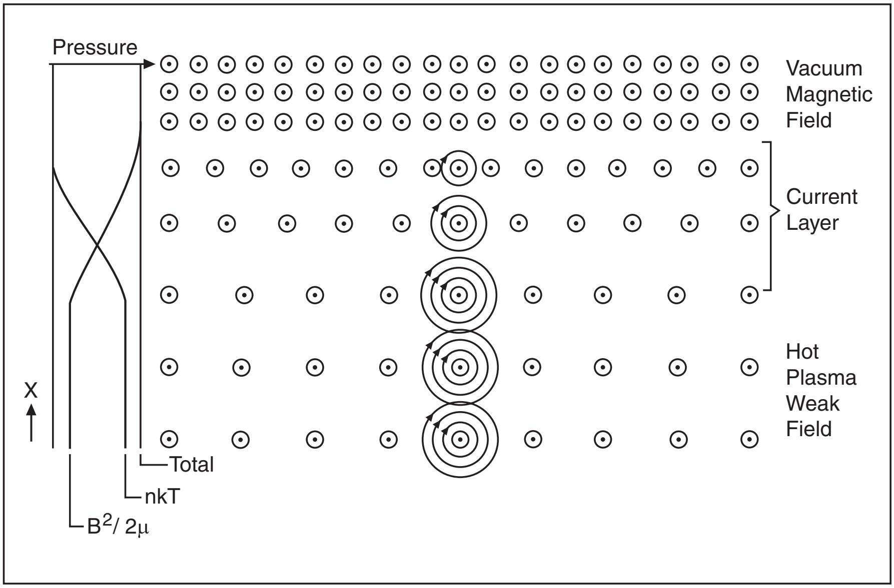


Figure 4.7. Schematic illustration of the balance of pressure between a hot weakly magnetized plasma (bottom) and a vacuum magnetic field (top). The pressure gradient produces a current that both acts to reduce the field strength and balances the magnetic pressure with a $\mathbf{J} \times \mathbf{B}$ or Lorentz force.

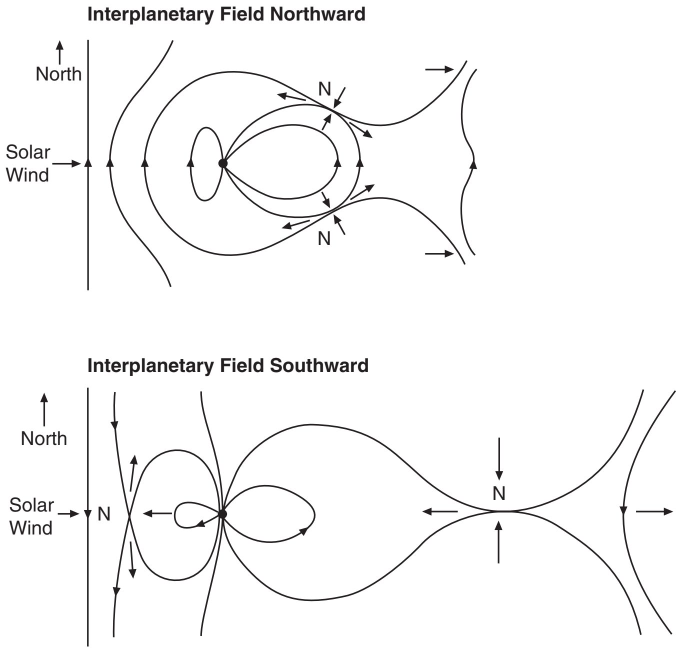
Before discussing, in the next section, the tangential stress applied by the reconnection process that carries out the lion's share of the energization of the magnetosphere, we should say a few words about the energization of the magnetosphere through normal stresses. If the magnetosphere did not dissipate energy, then when it is compressed by an increasing solar wind pressure, doing work on the magnetosphere, the added energy in the magnetosphere would remain until the time at which the solar wind pressure returned to its original state, at which time the expanding magnetosphere would do work on the solar wind. However, if the particles energized by the solar wind compression of the magnetosphere were to be lost to the atmosphere – for example, by pitch angle diffusion or charge exchange – then the magnetosphere could not expand as much when the solar wind pressure subsided, and not as much work would be done on the solar wind. Thus solar wind compressions do work on the magnetosphere, and large sudden compressions of the magnetosphere can be important in the overall energy balance. Nevertheless, reconnection is the major contributor to reconnection because it can tap the mechanical energy flux of the solar wind quite directly.

4.4 RECONNECTION

In the interaction of the solar wind with a planetary magnetosphere there are three possible magnetic topologies, or connectivities, between the interplanetary field and the planet. A magnetic field line might not intersect the Earth at all. We refer to such field lines as interplanetary, whether or not they may penetrate part of the magnetosphere. A field line might intersect the surface of the Earth twice – for instance by leaving the south polar region and moving across the equatorial and crossing the Earth's surface in the north polar region. We call such field lines 'closed'. Field lines leaving the Earth's surface could also link to the magnetic field lines in the solar wind. These are called 'open'. If we begin with only two topological classes of field lines – one that may or may not be connected to the Sun, but certainly are not connected to the Earth, and the other consisting of terrestrial, closed field lines that connect to the surface of the Earth twice – we do not have any members of the topological class that have one 'foot' on the Earth. For such field lines to arise, some mechanism that changes the connection of magnetic field lines, to reconnect them, would be needed. We will discuss the present understanding of reconnection below, but we first examine the consequences of the existence of a reconnection mechanism on the solar wind interaction, following the insight of Dungey (1961; 1963).

Figure 4.8 shows magnetic field lines in the noon–midnight meridian for two extreme directions of the interplanetary magnetic field. The top shows the situation when the interplanetary magnetic field (IMF) is due northward – when the magnetic field lines of the two different topologies (interplanetary and closed) are parallel at the subsolar point. The interplanetary field lines drape over the magnetosphere that has closed field lines of quite varied directions defining its surface. At one spot behind the terminator the field lines are antiparallel. At this point they can reconnect, and since the force applied by a bent field line is in the sense to accelerate the plasma in the strongly curved part of the field and straighten the field line, plasma flows into the two X points, labelled N, from above and below and outward from the X points, toward the Sun and away from the Sun. This process adds magnetic flux to the dayside magnetosphere and erodes it from the back. It is a steady-state process, and magnetic flux tubes must convect from the dayside to replace those removed from the night side. This reconnection stirs the magnetospheric plasma and transfers momentum from the solar wind to the magnetosphere. The rate at which this process proceeds will depend on the size of the reconnecting region, the geometry of the X line, and the plasma conditions in the neighbourhood of the reconnection. At the post terminator reconnection point in the top panel, the region of near antiparallel field lines is small, the X is squashed so the ejection region is small, and the magnetic field strength is relatively weak. Thus northward reconnection, while important, does not provide strong coupling between the magnetosphere and solar wind.

When the interplanetary magnetic field is southward – as it is in the bottom panel of Figure 4.8 – reconnection is much more rapid. For antiparallel fields at the subsolar point the size of the reconnection region could stretch across the entire dayside magnetopause. The curvature of the dayside magnetopause provides a more open X geometry, and the magnetic field is much stronger at the subsolar point thus post-terminator. Southward IMF conditions therefore provide much stronger entry of mass, momentum and energy than do northward conditions. The


Interplanetary Field Northward

Interplanetary Field Southward

Figure 4.8. Dungey's two models of the reconnecting magnetosphere (after Dungey, 1961; 1963). The top panel shows the noon–midnight meridian for northward interplanetary magnetic fields and the bottom panel the southward. The points labeled 'N' are neutral points where the magnetic field goes to zero. These are X-points in modern terminology. Straight arrows show the direction of accelerated flows.

circulation pattern of the magnetospheric plasma also is very different. The plasma flows from the subsolar point over the polar cap and into the tail, where reconnection can take place and return magnetic flux and plasma to the closed portion of the magnetopause and to solar wind in the antisolar direction. In some senses the conditions in the lower panel produce just a stronger version of the circulation pattern produced in the top panel, but in the opposite direction.

Having noted the similarity between Dungey's northward and southward reconnection models, we should also point out a subtle but very important difference that causes these two situations to behave in very different manners despite their topological resemblances. In both panels the X points are labelled N for neutral point – a place where the magnetic field goes to zero, as it does in this plane. In the top panel they are symmetrically placed, and should in this geometry behave similarly and react simultaneously. In the bottom panel the conditions at the two neutral points are quite different. The pressure is much higher at the dayside X point. The X-geometry is quite open, and the region of antiparallel fields may be quite large. We expect reconnection will be fast and quite responsive to changing IMF and solar wind conditions. The conditions on the nightside are quite different. The magnetic field is weak, the X-geometry is flattened, and the region of antiparallel fields may be quite small as the two lobes of the Earth's tail can respond independently to the solar wind stresses. The two lobes may be sheared with respect to one another. Moreover, the two neutral points are well separated, so there is little likelihood that they can respond simultaneously. In this situation the magnetosphere's response to time-varying boundary conditions becomes very important. Energy can be stored in the magnetosphere for long periods before it is released. This storage and release produces the phenomenon known as substorms, in which the magnetosphere stores energy in a growth phase, releases it during an expansion phase, and then relaxes back to the ground state during a recovery phase. We discuss this process in greater length below. Finally we note that there is one other very important difference between the two neutral points in the way that a region enters the reconnecting state. On the dayside the initial boundary condition has no magnetic field along the normal to the current sheet (magnetopause). Thus reconnection should start where the fields are most antiparallel. In the tail the important reconnection event, as we will see, is initiated on field lines that are closed. The onset of instability must involve the reduction of the component of the field normal to the current sheet to sufficiently small values. Thus the development of sufficient stretching should control when and where reconnection occurs, and not the shifting of the relative alignment of field lines in the two lobes. The required antiparallelism is maintained in the closed field region by the connection of the field across the current sheet.

4.5 DAYSIDE RECONNECTION

Above we stated that we expected that southward IMF would produce a much stronger coupling to the solar wind than northward IMF, even though both conditions should result in reconnection. Figure 4.9 provides evidence for this conjecture. Here we have measured the rate of energy flow into the magnetosphere by examining the Dst index that is a measure of the energy content of the magnetosphere (Burton *et al.*, 1975a, b). This provides the ordinate – the injection rate per hour. The abscissa, $Ey(mv - m^{-1})$, is the rate at which southward magnetic flux is convected to the magnetosphere by the solar wind $(-Vx \times Bz)$ in GSM coordinates, where Vx is the solar wind velocity in ms^{-1}, and Bz is the north–south magnetic field in Teslas. When the magnetic field is northward and the electric field of the solar wind is from dusk to dawn (negative in GSM coordinates), then there is not significant change in the energy content of the magnetosphere. When the IMF is southward and the electric field is from dawn to dusk, the energization is proportional to the rate at which southward magnetic flux is brought to the magnetosphere. This half-wave rectification has an important implication on how the reconnection rate must be controlled by the IMF direction and where reconnection occurs.

Figure 4.10 shows a plot of the regions on the magnetopause where the fields of the magnetosheath and magnetosphere have some degree of antiparallelism (Luhmann *et al.*, 1984) The dark shading shows where they are close to antiparallel. When the IMF is due south (upper left) reconnection occurs behind the terminator where the field lines are expected to be open or at least very elongated. As the field rotates to nearly horizontal orientation (upper right) the antiparallel region moves

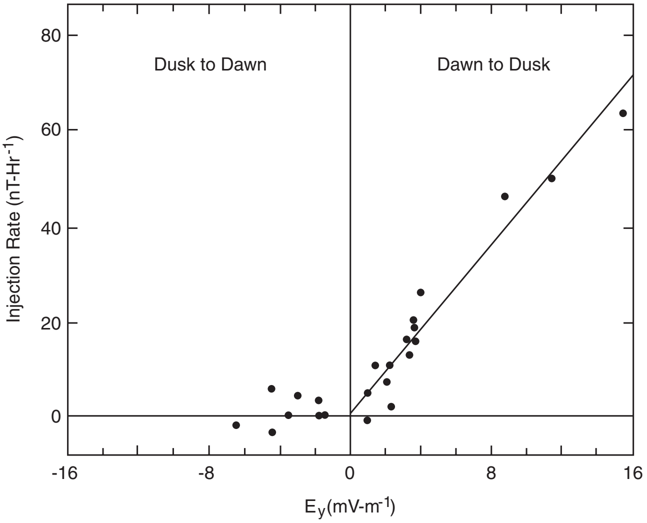


Figure 4.9. The rate of energy injected into the ring current as measured by the change per hour in the depression of the magnetic field in the equatorial regions on the surface of the Earth as a function of the southward magnetic field convected to the Earth by the solar wind in $mV - m^{-1}$.

toward the region on the magnetopause where the magnetic field is closed. When the field turns southward (lower right) the antiparallel region is solidly rooted in closed magnetospheric field lines. When reconnection starts it will transport magnetic flux into the tail and store energy there. When the IMF turns due south (lower left) the antiparallel region becomes much bigger, and we would expect very strong reconnection to ensue.

The function displayed in Figure 4.9 is the simplest coupling function between the solar wind and the Earth magnetosphere, called the half wave rectifier. It is controlled by the component of the interplanetary field perpendicular to the solar wind flow and its clock angle around the flow. For angles from -90° to 90° (horizontally west, through northward, and back to horizontally east) there is no coupling. For angles from 90° to 180° and back to -90°, the coupling is proportional to $\cos(\varphi)$. This coupling function approximates the rate of transport of reconnected closed magnetic flux from the dayside magnetosphere to the tail that, as illustrated by Figure 4.10, should depend on φ. This function depends on the geometry and topology of the magnetosphere and has no simple analytic expression. Some however have chosen to use $\cos^2(\varphi/2)$ and $\cos^4(\varphi/2)$ to approximate it.

The calculation displayed in Figure 4.6 was performed using the Spreiter *et al.* (1966) gas dynamic model for the magnetosheath field and an empirical model for the magnetospheric field. It is now possible to use MHD models for such calculations. This allows one to also tilt the dipole axis. This calculation reveals that the size of the nearly antiparallel fields is very sensitive to dipole tilt (Russell *et al.*, 2003). This dependence expressed as the length of the neutral or reconnection line is shown in Figure 4.11 for these tilts of the dipole as the interplanetary magnetic field rotates

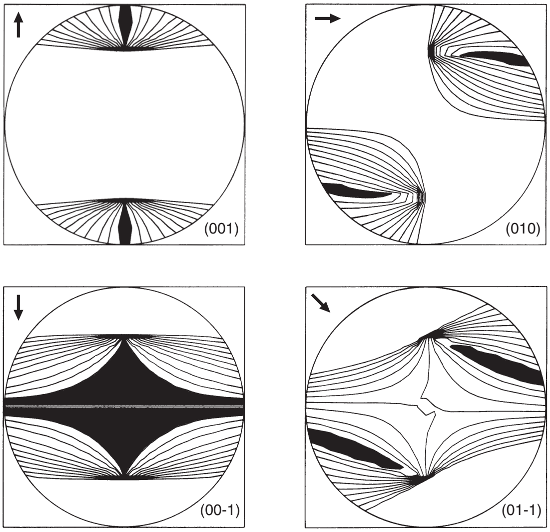


Figure 4.10. Locations of antiparallel magnetic fields on the magnetopause using four different directions of the interplanetary magnetic field. Dipole has zero degree tilt (perpendicular to solar wind flow). Magnetospheric field from analytic model and magnetosheath field from gas dynamic simulation (after Luhmann *et al.*, 1983).

from due south to 40° south. Notice that each curve has a maximum at a clock angle that, when added to the tilt, equals 180°. This inferred dependence on the tilt of the dipole may be important for the seasonal variation of geomagnetic activity as the tilt of the dipole varies over the course of the year, oscillating around 0° at the equinoxes and reaching as high as about 34°N and S at the solstices. The behaviour of the length of the neutral line versus tilt angle shown in Figure 4.11 would enhance the semi-annual variation of geomagnetic activity above that proposed by Russell and McPherron (1973a), who assumed no dipole tilt dependence of the reconnection rate.

While the hypothesis that field lines must be nearly antiparallel to reconnect has much to recommend it, the alternate hypothesis, that the fields need only a small amount of antiparallelism, has also strong advocacy, especially among MHD modellers. MHD models exhibit little dependence on the strength of so-called guide field, out of the plane of the X fields. Kinetic simulations show evidence of a delayed onset of reconnection with a guide field (Scholer *et al.*, 2003). Thus the results from the modeling community are not unambiguous or unanimous. The study of what factors control the rate of reconnection promises to be a very active field for the next several years.

4.6 SUBSTORMS

A characteristic response of the magnetosphere to the inputs of energy from the solar wind is the substorm in which the aurorae intensify and move equatorward and

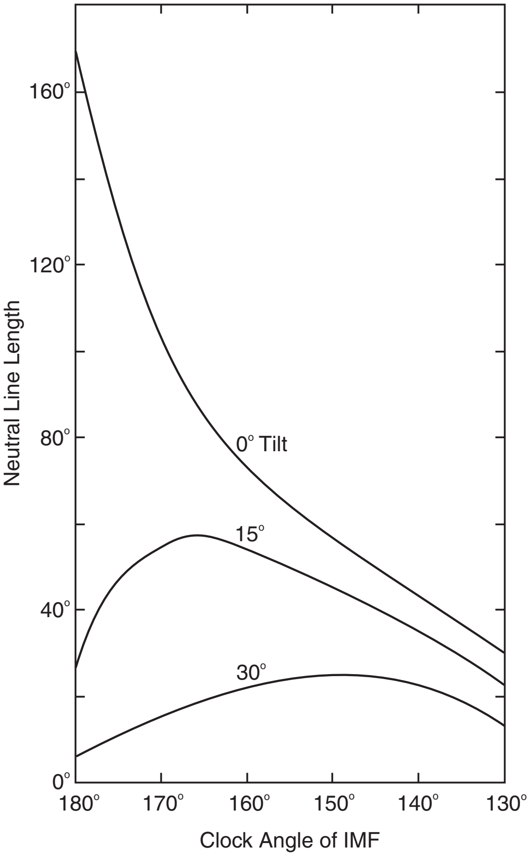


Figure 4.11. Length of the neutral line in degrees of longitude for varying tilt angle and clock angle of the interplanetary magnetic field. Magnetospheric and magnetosheath magnetic fields derived from MHD simulations (Russell *et al.*, 2003).

poleward in repeatable sequences. Our modern understanding of this phenomenon is that energy is stored in the tail for later release into the night magnetosphere, and that the process that stores and releases energy affects the size of the polar cap (open field line region) and the size of the active auroral oval (McPherron *et al.*, 1973; Russell and McPherron, 1973b). Understanding the cause of substorms is difficult, because both dayside and nightside reconnection are involved, and these two regions can act quite independently.

We understand the growth phase quite well. The best way to study this phase of a substorm is to examine periods of southward IMF turning when the IMF has been northward for some time. When the IMF turns southward, as in the top panel of Figure 4.12, reconnection begins on the dayside, erodes the magnetopause so it moves inward, and transports magnetic flux to the tail over the polar caps. The erosion of the dayside and the addition of flux to the tail changes the shape of the

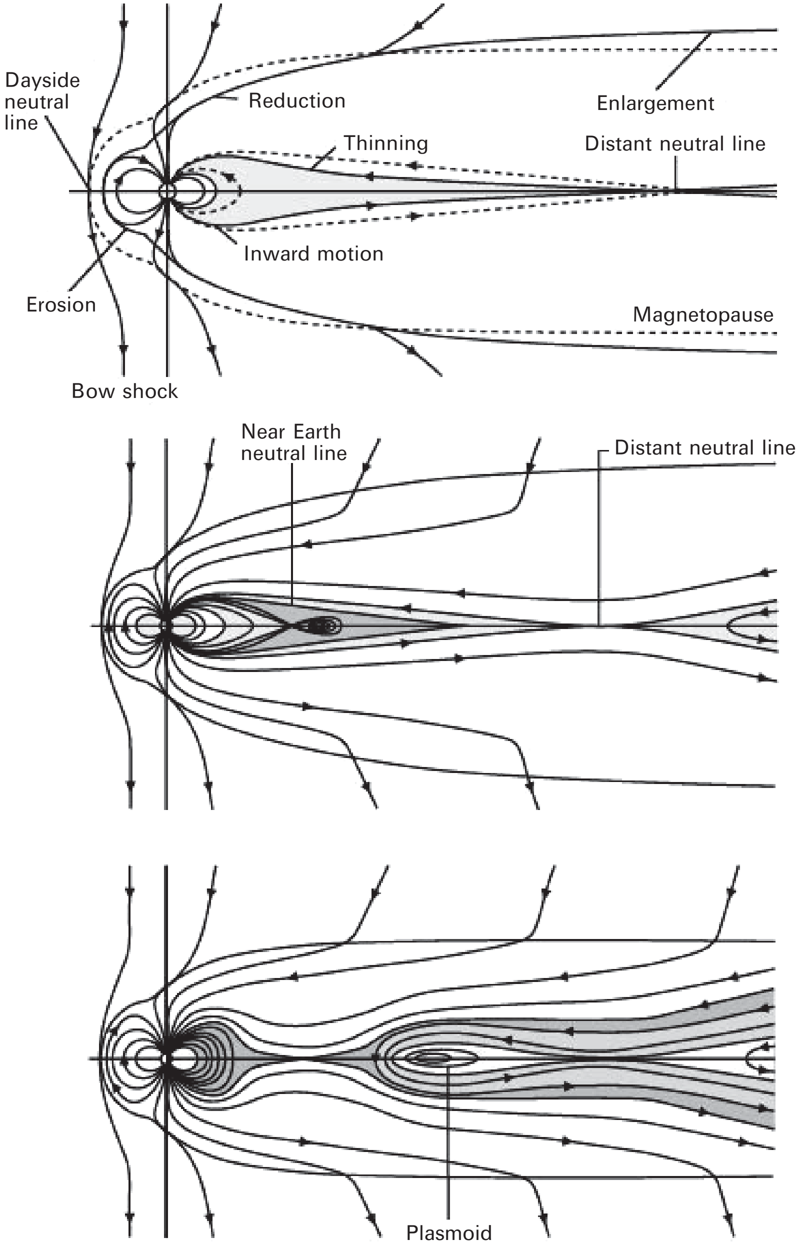


Figure 4.12. Sketch of the evolution of the magnetic field and plasma in the noon–midnight meridian during the period leading up to a substorm onset. In the top panel, magnetic flux is transported to the tail, reducing the size of the dayside magnetopause and increasing the width downtail. Simultaneously, the plasma sheet thins. Later (middle panel) the near-Earth plasma sheet forms a neutral point. Still later, the reconnection process reaches the tail lobes and the reconnection rate increases and a plasmoid is ejected.

magnetosphere so that the solar wind dynamic pressure pushes at a more normal angle of incidence on the tail boundary and the magnetic pressure in the tail lobes increases. The tail current sheet thins close to the Earth for two reasons. First, the current sheet is compressed by the additional pressure in the lobes; and second, the plasma sheet is on closed field lines that are being pulled sunward to replace flux eroded by reconnection. Eventually the near-Earth tail current sheet is thin enough and sufficiently stretched that reconnection begins near the Earth, as shown in the middle panel. This reconnection need not be explosive. In fact, since the field strength is low in this region and the plasma density is high, the Alfvén velocity of the plasma, that controls the velocity of the reconnected plasma accelerated from the X-lines, is also very low. So a plasmoid or magnetic island grows, but does not become large and is not ejected down the tail. Eventually (as discussed below) the reconnection rate increases as it cuts through the plasma sheet into the lobes where the density is low, the magnetic field strength is high and the Alfvén velocity is high. Once reconnection severs the last closed field line it releases the plasmoid, and reconnection proceeds rapidly until a new equilibrium state is reached. Because the newly reconnected flux piles up on the night magnetosphere, the X-point moves down the tail away from the Earth toward where the distant X-point had been. Since magnetic flux is removed from the tail the slope of the tail changes, the solar wind pressure along the normal drops, and the field strength in the lobe decreases.

We can put this qualitative picture on a more quantitative footing by examining the expected inventories of magnetic flux in the different regions as the reconnection rates vary on the dayside magnetopause and in the tail and the transport rate back to the dayside. First, we examine a two-neutral-point model: dayside magnetopause and tail neutral sheet so that a plasmoid is not formed. Then we examine a three-neutral-point model with a neutral point forming on closed field lines, and producing a plasmoid. As above, we start our discussion after a period of northward IMF so that the magnetosphere is in its ground state.

Figure 4.13 shows, in the top panel, the configuration of the magnetosphere just after the southward IMF turning. The flux tube that reconnects at the nose shortens as it moves over the top and bottom of the magnetosphere. Here the magnetic field is putting energy into the plasma. Once the reconnected field line is parallel to the normal to the magnetopause, the acceleration of the plasma stops because the field lines are becoming longer and the curvature of the field line reverses so that it now slows down the solar wind. In this post-cusp region, mechanical energy is extracted from the solar wind and converted to magnetic energy in the stretched-out magnetic flux being added to the tail lobes. Electromagnetic flux (Poynting vector) flows across the magnetopause, building up the magnetic energy stored in the tail. We can measure this energy flow at least qualitatively by tracking the merging rate of the interplanetary and planetary magnetic fields on the dayside. This is shown by the curve M in the middle panel. This process also removes flux from dayside region (Φ_{Day}, in the bottom panel). This allows the magnetopause to move inward under constant solar wind pressure. Since reconnection in the tail is not immediately triggered by this increase in reconnection on the front of the magnetosphere, the

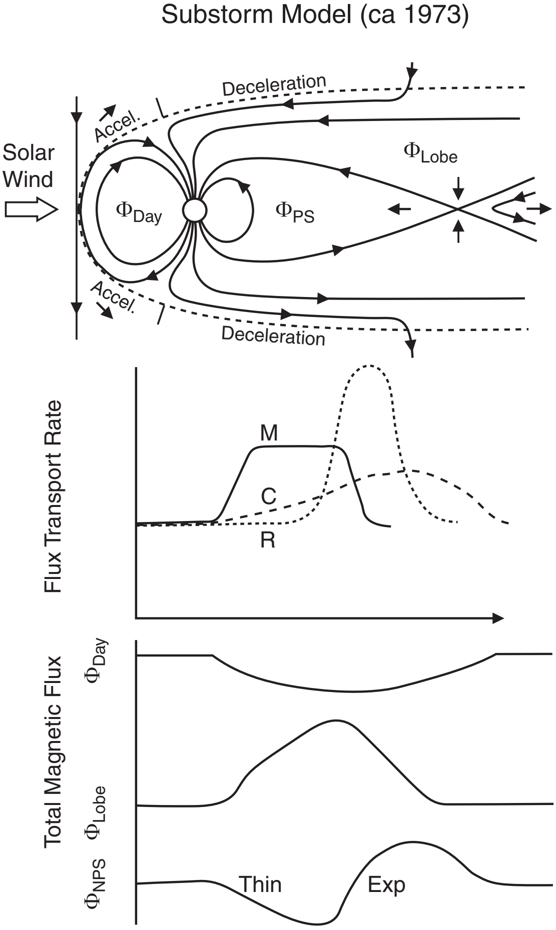


Figure 4.13. Schematic illustration of the flux transport process. The varying rate of dayside reconnection, M, tail reconnection, R, and convection from night to day, C, leads to variations in the flux content of the dayside magnetosphere, Φ_{Day}, the tail lobe, Φ_{Lobe}, and the plasma sheet, Φ_{NPS}. Solar wind is both accelerated (dayside) and decelerated (nightside) in this process.

flux in the lobes Φ_{Lobe} increases, as shown in the bottom panel. Eventually (for reasons discussed later) reconnection begins in the tail and the added magnetic flux is removed from the tail and returned to the dayside, as seen in the profiles of Φ_{Day} and Φ_{Lobe} when the reconnection rate, R, is enhanced.

This explanation of the substorm cycle is incomplete, as it does not address some key observations. First, as drawn it does not explain, at least explicitly, the occurrence of plasmoids – field structures within the tail that appear to close on themselves. It does not explain how a northward turning of the IMF can lead to substorm triggering, nor why the substorm onset is so abrupt.

We can capture some of this richness in substorm phenomenology if we incorporate a distant neutral point in the tail and then associate the substorm onset with the activation of a new, closer reconnection site on previously closed field lines. This produces the more complex diagram in Figure 4.14, where an inventory is also maintained of the flux in the distant plasma sheet, Φ_{DPS}, and the plasmoid, Φ_{PM}. This model can control the time of the onset of the expansion phase by keeping the near-Earth neutral point surrounded by hot plasma produced by the distant neutral point. As long as the magnetic field is low and the plasma is dense at the near-Earth neutral point, the reconnection rate will be low and the formation of a plasmoid will be slow. But if the near-Earth neutral point starts to reconnect in lobe field strengths and plasma densities, the reconnection rate will increase rapidly. A northward turning can lead to substorm onset by turning off production of plasma sheet plasma. In reality there may be more than two neutral points at any time, and multiple onsets may occur as a result. The magnetotail is a complex region, and our two-dimensional sketches do not do justice to this complexity. Hopefully, future multispacecraft missions will enable us to sort out the key processes that control substorm behaviour.

4.7 STORMS

The normal hour-to-hour variability of the IMF leads to substorms as the IMF switches from northward to southward and back to northward, but on occasion the solar wind contains magnetic and plasma structures that apply a coherent stress to the magnetosphere over many hours, and the response of the magnetosphere is quite different under such conditions than it is to a substorm. Figure 4.15 shows the variation of the horizontal component of the Earth's magnetic field over a 10-day period during what we call a magnetic storm. Like a substorm there are three phases. First, there is the sudden commencement where the magnetic field is suddenly compressed, which occurs when the solar wind dynamic pressure increases. The magnetic field then decreases rapidly in what is called the main phase. Eventually the field strength starts to return to normal in the recovery phase. The second two phases have a simple explanation in terms of the energization of plasma inside the magnetosphere. Energetic plasma is found in the near-tail region and in the main part of the magnetosphere circulating in a ring current. The field depression occurs because of the currents generated as the particles gradient drift around the magnetosphere and also diamagnetic currents due to the gyration of the particles. These currents produce a depression in the surface field proportional to the energy of the plasma stored therein (Dessler and Parker, 1959; Sckopke, 1966).

A very simple formula has been developed that can predict the magnetospheric energization and the magnetic field profile observed. This formula uses the solar wind dynamic pressure and the convected southward magnetic field to calculate the compression of the magnetosphere and the rate of energy added to the magnetosphere (Burton *et al.*, 1975a, b). For most purposes this suffices, but some have

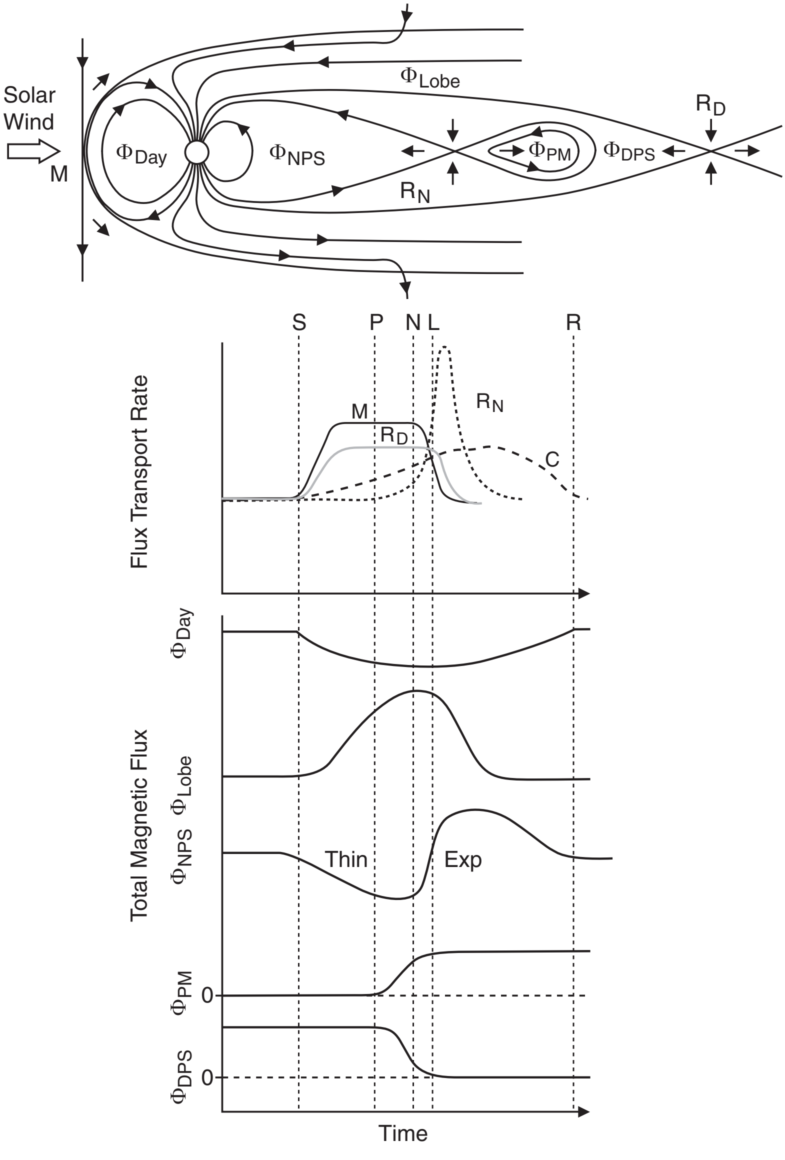


Figure 4.14. Schematic illustration of the interplay between a distant neutral point and a near-Earth neutral point in modulating the transport of plasma and magnetic flux in response to changing interplanetary magnetic conditions. Near-Earth reconnection forms a plasmoid or magnetic island that is released down the tail when the near-Earth neutral point finds itself on open magnetic field lines where the Alfvén velocity is high and the reconnection rate rapid. The lower panel shows how the flux in the distance plasma sheet, Φ_{DPS}, and the plasmoid, Φ_{PM}, vary with time in addition to those described in Figure 4.13. The reconnection rate at the distant neutral point, R_D, may be quite responsible to the dayside rate, M, but the rate at the near-Earth neutral point, R_N, experiences explosive growth.

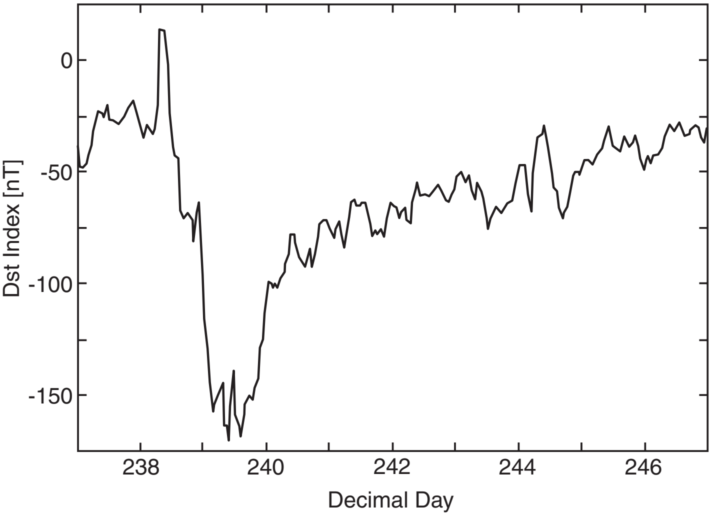


Figure 4.15. D_{st} index for a typical geomagnetic storm, created by averaging the horizontal component of the magnetic field at the surface of the Earth over all local times. The sudden increase in D_{st} on day 238 occurs when the solar wind dynamic pressure suddenly jumps. The subsequent drop in Dst occurs when the interplanetary magnetic field turns southward.

attempted to develop more sophisticated predictors at the price of a more complex set of equations.

4.8 FIELD-ALIGNED CURRENTS

Whenever the magnetic field is sheared or twisted so that it is not 'curl-free' in the mathematical sense, there is a current. Figure 4.16 shows a cut-away diagram of the magnetosphere with these magnetospheric currents illustrated. Many of these current systems flow in essentially dissipation-free regions. The current on the magnetopause closes in the magnetopause and serves principally to change the direction of the magnetic field on the outside to the inside of the magnetosphere. However, the field-aligned currents close in a resistive ionosphere across the magnetic field lines, and the $\mathbf{J} \times \mathbf{B}$ force, associated with these currents, pulls on the plasma and overcomes the drag of the neutrals on the ionosphere. In the magnetosphere there is also a $\mathbf{J} \times \mathbf{B}$ force – one that drives the currents through the ionosphere. The currents cross the field line on pressure gradients, much as illustrated for the magnetopause in Figure 4.7. Thus, through field-aligned currents stresses can be transmitted from the outer magnetosphere to the ionosphere, and an entire flux tube can be moved around the magnetosphere.

The two major regions of connection are over the polar cap, and the return flow at lower, auroral, latitudes. We recall that northward reconnection can make a weak flow from midnight to noon across the polar cap. However, the flow is predomi-

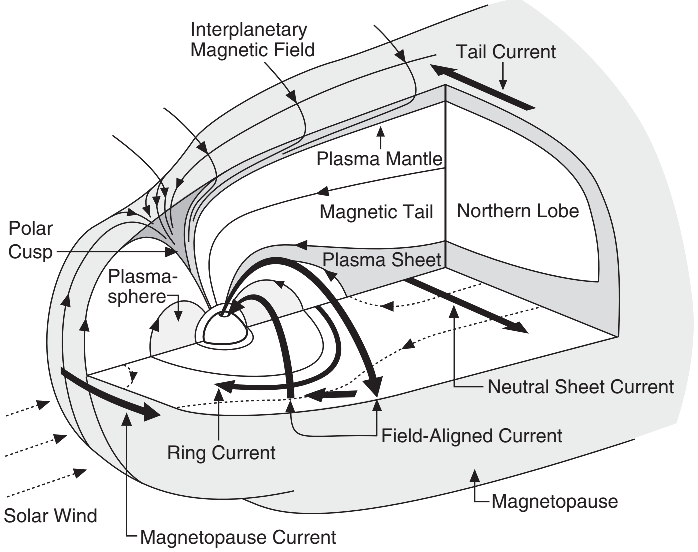


Figure 4.16. A cut-away diagram showing the magnetic and plasma structure of the magnetosphere together with some of the more important current systems.

nantly from noon to midnight over the polar cap, and we attribute this to the observation that reconnection is a much more rapid process for southward IMF. The field-aligned current systems that enable this transport are illustrated in Figures 4.17 and 4.18. Figure 4.17 is a view of the north polar cap from above, showing the polarity of the current as it enters the polar ionosphere. The pattern is that of a large spiral circling the polar cap, with opposite senses of currents on the dawn and dusk sides. Figure 4.18 shows the view looking back to the Sun, with currents coming down (in the terminator plane) on the poleward edge of the auroral zone on the dawn side and going upward at dusk. On the lower edges the currents are reversed. The currents were originally named after their locations – region 1 and region 2 – but they are better understood in terms of what they do and where they are created. Loop A is connected to the stresses associated with dayside reconnection, and loop B is connected to the pressure gradients set up to push the plasma from the nightside (reconnection) region to the dayside reconnection region. The strength of these currents is very dependent on the reconnection rate and also on the illumination of the ionosphere where the ionospheric drag occurs. The higher the electrical conductivity of the ionosphere the more 'frozen-in' will be the field lines and the stronger need be the field-aligned currents or equivalently the bending of the magnetic field to overcome the ionospheric drag.

Figure 4.19 illustrates in more detail how the drag is applied to the ionosphere. At high altitudes magnetospheric stresses pull a set of field lines, in this case, into the page. At low altitudes the ionosphere resists this pull, the plasma moves but slowly relative to the plasma at high altitudes, and a shear in the field line direction develops

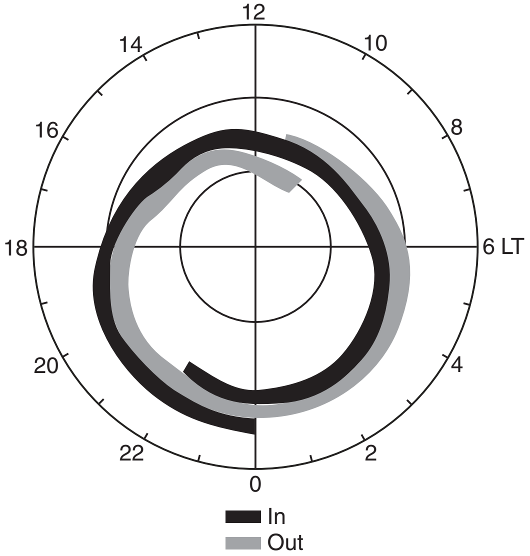


Figure 4.17. A schematic view of the north polar cap showing the region in which electrical current flows down into the auroral oval (dark shading) and up out of the auroral oval (light grey shading).

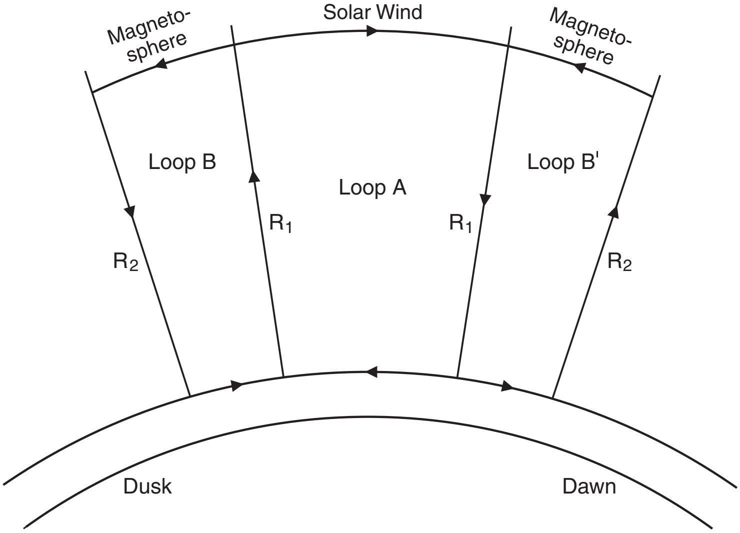


Figure 4.18. View toward the Sun over the north polar cap of the field-aligned current system in the dawn–dusk plane. Region 1 currents, labelled R_1, flow down on the dawn side and up on the dusk side at the polarward edge of the auroral oval. They have two components: one that links to solar wind induced stresses, and one that links to magnetospheric stresses. The lower latitude region 2 currents respond only to magnetospheric stresses.

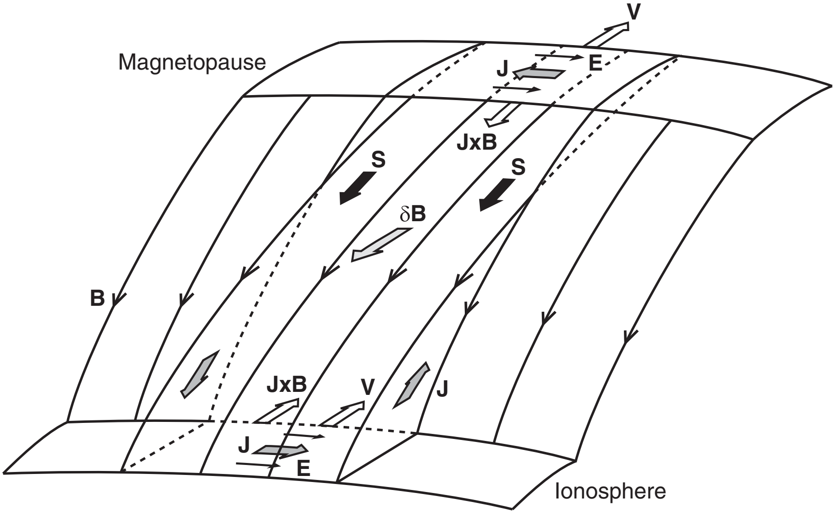


Figure 4.19. Three-dimensional sketch of the linkage of stresses at high altitudes in the magnetosphere to the ionosphere and the consequent drag on the ionospheric plasma. At high altitudes the currents cross magnetic field lines along pressure gradients as shown in Figure 4.7. Field lines are pulled backwards in the region of applied stress. Where the field lines are sheared, currents flow along the magnetic field, as shown. If the plasma is static there is no electric field, but if there is flow in the direction labelled V, an electric field and a Poynting vector, S, arise. Energy is deposited where there is a divergence of this Poynting vector.

at the edges of the block of stressed magnetic flux. These magnetic sheer layers are the field-aligned currents.

4.9 SUMMARY

The magnetosphere is a very complex system. Here we have presented some qualitative insight into the coupling of parts of the system with emphasis on those controlled by the solar wind. We still continue to improve our understanding of the magnetosphere with the study of existing databases, with the development of various types of numerical models, and with the development of new and more sophisticated instruments and missions. We have learned much about the magnetosphere in the last half century since the beginning of the International Geophysical Year in 1957. It has been important to do so because our increasingly technological society has moved critical systems into space for monitoring weather, for communication, and for global monitoring of the health of the planet and the activities of one's neighbours. It is clear that if we are to use space as much as our present plans foretell, we must continue to improve our understanding of this environment and how it works, and be able to predict its behaviour given knowledge of interplanetary conditions.

4.10 ACKNOWLEDGEMENTS

The preparation of this review was enabled by research grant ATM04-02213 from the National Science Foundation.

4.11 REFERENCES

Burton, R. K., R. L. McPherron, and C. T. Russell, An empirical relationship between interplanetary conditions and Dst., *J. Geophys. Res.*, **80**(31), 4204–4214, 1975.

Burton, R. K., R. L. McPherron, and C. T. Russell, The terrestrial magnetosphere: A half-wave rectifier of the interplanetary electric field, *Science*, **189**(4204), 717–718, 1975.

Dessler, A. J., and E. N. Parker, Hydromagnetic theory of magnetic storms, *J. Geophys. Res.*, **64**, 2239–2259, 1959.

Dungey, J. W., Interplanetary magnetic field and the auroral zones, *Phys. Rev. Lett.*, **6**, 47–48, 1961.

Dungey, J. W., The structure of the exosphere or adventures in velocity space, *Geophysics: The Earth's Environment*, C. Dewitt, J. Hieblot and A. Lebeau, editors, New York, Gordon and Breach: 505–550, 1963.

McPherron, R. L., C. T. Russell, and M. P. Aubry, Satellite studies of magnetospheric substorms on August 15, 1968. IX. Phenomenological model for substorms, *J. Geophys. Res.*, **78**, 3131–3149, 1973.

Petrinec, S. M. and C. T. Russell, Investigations of hydrodynamic and magneto-hydrodynamic equations across the bow shock and along the outer edge of planetary obstacles, *Adv. Space Res.*, **20**, 743–746, 1997.

Russell, C. T. and R. L. McPherron, The magnetotail and substorms, *Space Sci. Rev.*, **15**, 205–266, 1973.

Russell, C. T. and R. L. McPherron, Semiannual variation of geomagnetic activity, *J. Geophys. Res.*, **78**, 92–108, 1973.

Russell, C. T., Y. L. Wang, and J. Raeder, Possible dipole tilt dependence of dayside magnetopause reconnection, *Geophys. Res. Lett.*, **30**(18), 1937, doi:10.1029/2003GL017725, 2003.

Scholer, M., I. Sidorenko, C. H. Jaroschek, R. A. Treumann, and A. Zeiler, Onset of collisionless magnetic reconnection in thin current sheets: three-dimensional particle simulations, *Physics of Plasmas*, **10**, 3521–3527, 2003.

Sckopke, N., A general relation between the energy of trapped particles and the disturbance field near the Earth, *J. Geophys. Res.*, **71**, 3125, 1966.

Spreiter, J. R., A. L. Summers, and A. Y. Alksne, Hydromagnetic flow around the magnetosphere, *Planet. Space Sci.*, **14**, 223–253, 1966.

Wang, Y. L., J. Raeder, and C. T. Russell, Plasma depletion layer: The role of the slow mode waves, *Ann. Geophys.*, **22**, 4259–4272, 2004.

5

Major radiation environments in the heliosphere and their implications for interplanetary travel

Norma B. Crosby

Highly energetic particles fill our Solar System, and are a significant hazard for current space missions as well as for any future interplanetary travel. This chapter introduces the various populations of particles in the energy range from eV to more than 10^{21} eV that are present in our Solar System, and covers galactic and anomalous cosmic rays, solar energetic particle events (especially solar proton events), particles accelerated by corotating interaction regions, interplanetary shocks, and planetary bow shocks, as well as geomagnetically trapped particles occuring in some planets' atmospheres. These populations – especially the constant flux of galactic cosmic rays along with the sporadic solar energetic particle events – provide challenges for scientists that work on the prediction of these phenomena, as well as for engineers who design mitigation strategies for spacecraft. The last part of the chapter concerns the implications of these radiation environments on interplanetary travel, presenting interplanetary space weather forecasting parameters. As a case study, a manned mission to Mars is considered.

5.1 INTRODUCTION

In the beginning of the twenty-first century, the impetus for human space travel is being driven by unprecedented opportunities that exist for tomorrow (colonies on planets and moons, asteroid mining, space tourism and hotels, transportation technology, commercial opportunities on the International Space Station, solar power, and so on). In the next decades, human missions to the Moon on a routine basis will no doubt be the stepping stone for interplanetary travel such as a human mission to Mars. Only several days away, the Moon will become our first Solar System colony and offer us new opportunities for studying the space environment outside the protection of our magnetosphere. For space flight beyond Earth's magnetosphere, both crew and spacecraft equipment face significant hazards from the natural ionizing environment found in our Solar System.

Interplanetary space, better known as the heliosphere, is the ‘space’ between the planets in our Solar System. ‘Helios’ is the ancient Greek word for the Sun, and the prefix is used to describe the space environment between the planets, which is under the influence of the Sun. The heliosphere can be thought of as a vast magnetic bubble containing the Solar System, the solar wind and the interplanetary magnetic field (IMF), as well as numerous cosmic-ray populations and dust. This section introduces the heliosphere as well as the major radiation environments that it includes. (The atmospheres of planets, including phenomena such as dust storms, are outside the scope of this chapter.)

5.1.1 The heliosphere

The solar wind, which expands radially from the Sun at supersonic speed, blows a huge cavity – the heliosphere – into the surrounding interstellar cloud, filling it with solar material and magnetic field. The plasma component of the interstellar gas is kept outside the heliosphere. Since their launches in 1977 the twin Voyager 1 and 2 spacecraft continue exploring where nothing from Earth has flown before.

Both Voyagers are headed towards the outer boundary of the Solar System in search of the heliopause – the region where the Sun’s influence wanes and the solar wind encounters the interstellar plasma and magnetic field. Scientists have long claimed that the first sign of such an encounter would be a ‘termination shock’ where the solar wind abruptly slows down, to be followed by the heliopause where it achieves pressure balance with the interstellar medium. During 15–17 December 2004, the Voyager 1 dual magnetometers observed that the IMF intensity rose from approximately 0.05 nT to $\sim$0.15 nT (Figure 5.1, upper panel), and since then the magnetic field has remained at these high levels (Burlaga *et al.*, 2005). An increase in magnetic field strength is expected when the solar wind slows down, and the general consensus is that the termination shock was crossed at 94.01 AU in mid-December 2004 (Stone *et al.*, 2005). These pioneering spacecraft are still functioning, and have practically encountered all the major radiation environments that are present in our Solar System (such as trapped particle environments around the outer giant planets (Section 5.8) and galactic cosmic rays in the heliopause region (Section 5.2)).

5.1.2 Cosmic rays

At the beginning of the twentieth century scientists were puzzled by the fact that air in electroscopes (instruments for detecting electrical charges) became electrically charged (ionized) no matter how well the containers were insulated. In 1900 C. T. R. Wilson had discovered ‘continuous atmospheric radiation’. It was believed for many years that radioactivity from ground minerals was responsible for this effect. Reading about these earlier experiments, the German scientist Victor Hess speculated as to whether the source of ionization could be located in the sky

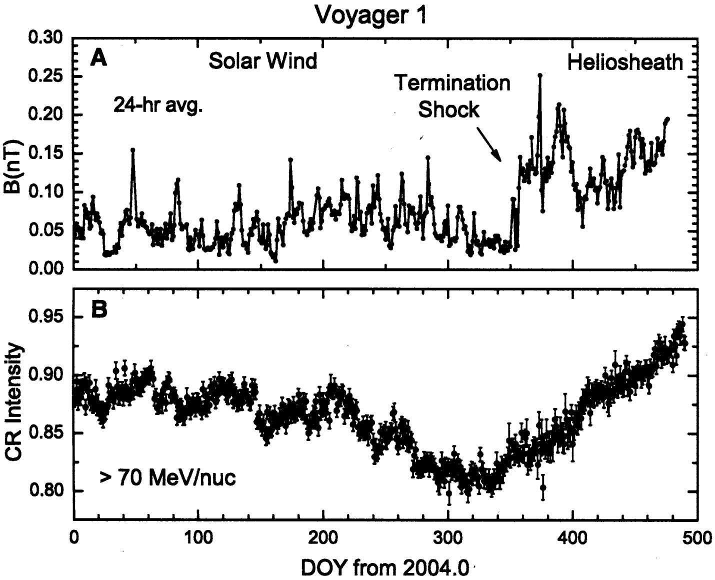


Figure 5.1. The relationship between daily averages of the magnetic field strength, B (panel A) and the intensity of cosmic rays, CR (panel B) in the heliosheath is different from that in the 'solar wind'. Error bars show means ±SD. DOY is day of year. (From Burlaga *et al.*, 2005.)

rather than on the ground. Thus he determined the height at which ground radiation would stop producing ionization (about 500 metres) and designed instruments (ion chambers) that would not be damaged by temperature and pressure changes. During 1911–1913 he made ten balloon ascents and found that an electroscope discharged more rapidly as he ascended. He had discovered that the ionization at a height of several miles was many times greater than at the Earth's surface (Hess, 1911, 1912).

Hess's theory about 'rays from space' did not receive general acceptance at the time he proposed it, but increased research almost a decade later supported it. First named for Hess, the newly discovered radiation was dubbed 'cosmic' by Robert A. Millikan in 1925. For some time it was believed that the radiation was electromagnetic in nature (hence the name cosmic 'rays'). However, during the 1930s it was found that cosmic rays must be electrically charged because they are affected by the Earth's magnetic field. Pfotzer (1936) showed that the flux did not continue to increase but reached a peak at about 15 km, after which it diminished rapidly. For his discovery of cosmic radiation Hess was named co-recipient of the Nobel Prize in physics in 1936. (For more information about the work of Hess and his contemporary colleagues, see Federmann (2003).)

In 1938 Pierre Auger positioned several particle detectors high in the Alps and noticed that two detectors located many metres apart both detected the arrival of particles exactly at the same time (Auger and Grivet (1939); Auger *et al.* (1939)). He had discovered what is known today as 'extensive air showers' or 'cosmic ray showers'. By changing the distance between the detectors he could observe cosmic rays with energies up to 10^{15} eV.

Technically, cosmic rays from space are termed 'primary particles', and any particles (pions, muons, neutrons, etc.) created in the atmosphere from collisions with nuclei of atmospheric gases (mostly nitrogen and oxygen) are termed 'secondary particles'. A bit of energy is transferred to each new secondary particle. Secondary cosmic rays spread out and continue to hit other particles and air molecules, creating a nucleonic–electromagnetic cascade towards the ground. Only muons and neutrinos penetrate to significant depths underground. The muons produce tertiary fluxes of photons, electrons, and hadrons. The type of cosmic rays that produce 'showers' are the ones with the highest energies ($>10^{14}$ eV).

The term 'cosmic rays' usually refers to galactic cosmic rays, although this term has also come to include anomalous cosmic rays and solar cosmic rays:

- Galactic cosmic rays originate from outside our Solar System.
- Anomalous cosmic rays originate from interstellar space beyond the heliopause.
- Solar cosmic rays, also called solar energetic particles, originate from solar flares and shock-associated coronal mass ejections (CMEs). Shocks in the interplanetary medium (interplanetary shocks) also produce energetic particles.

Most cosmic rays are ionized atoms, ranging from protons, helium (α-particles), up to the iron nucleus and even beyond to heavier nuclei (such as uranium). Cosmic rays also include high-energy electrons, positrons, and other subatomic particles. The energy range of these combined sources extends over fifteen orders of magnitude, from the super-thermal to more than 10^{21} eV.

5.1.3 Other particle populations

Spacecraft in near-Earth orbit and people (astronauts and cosmonauts) in low-Earth orbit (such as on the International Space Station) are susceptible to cosmic rays, as well as other types of particles (this applies to other planets too). Planets and their interaction with the interplanetary medium can also be a rich source of energetic particles, especially those planets with magnetic fields. In general there are three distinct regions where energetic particles can be found:

- Planetary bow shocks can act as acceleration sites.
- Inner magnetospheres, where particles can be stably trapped in well-defined radiation belts.
- The dynamic regions of the outer magnetospheres, magnetosheaths and magnetotails that are buffeted by the external force of the varying solar wind and the magnetospheric structure it contains.

Using our Moon as a lunar base will mean traversing our locally geomagnetically trapped particle environment – Earth's radiation belts – on a regular basis. Such belts also exist on other planets with magnetic fields, especially on the outer planets of our Solar System (Jupiter, Saturn, Uranus and Neptune).

5.1.4 Summary

There are many different types of particle environments in our heliosphere, each type with their own specific characteristics. The particle populations originate in different sources, all having their individual energy spectrum, temporal development, and spatial extent. Table 5.1 lists the various particle populations that make up our Solar System as a function of energy, temporal and spatial range. Populations are listed in order of energy (highest to lowest), and the section number where they are discussed in this chapter is included (right column). A general review about particle acceleration at the Sun and in the heliosphere can be found in Reames (1999).

This chapter presents the individual particle populations found in our heliosphere (listed in Table 5.1) and discusses their implications for interplanetary travel. Galactic cosmic rays are discussed in Section 5.2, followed by anomalous cosmic rays in Section 5.3. Solar energetic particle events are presented in Section 5.4, with emphasis on solar proton events. Energetic storm particles are discussed in Section 5.5. The two phenomena (corotating interaction regions and planetary bow-shocks) are discussed as mechanisms for particle acceleration in the next two Sections. Section 5.8 gives an introduction to Earth's radiation belts as well as trapped particle populations found around other planets in our Solar System. The implications of particle populations on interplanetary travel (engineering and scientific issues), for both spacecraft and humans, is discussed in Section 5.9. A mission to Mars scenario is given as a case study example. The chapter finishes with a summary and some final words.

Table 5.1. Particle populations that make up our Solar System with corresponding energy, temporal and spatial range. The Section number (where they are presented in this chapter) is listed in the right-hand column.

Particle populations	Energy range	Temporal range	Spatial range	Section number
Galactic cosmic rays	GeV–TeV	Continuous	Global (isotropic)	5.2
Anomalous cosmic rays	<100 MeV	Continuous	Global (isotropic)	5.3
Solar energetic particles	keV–GeV	Sporadic (minutes to days)	Directional (anisotropic)	5.4
Energetic storm particles	keV–(>10 MeV)	Days	Extended	5.5
Corotating interaction regions	keV–MeV	27 days	Large-scale	5.6
Particles accelerated at planetary bow shocks	keV–MeV	Continuous	Continuous	5.7
Trapped particle populations	eV–MeV	Variations 'minutes–years'	Variations 'height–width'	5.8

5.2 GALACTIC COSMIC RAYS

The most energetic particles (energies up to 10^{21} eV) found in our Solar System are those that have origins far outside our Solar System. They are known as galactic cosmic rays (GCRs). Their composition is mostly hydrogen nuclei (protons), $\sim$7–10% helium and $\sim$1% heavier elements. All GCRs are fully ionized; that is, they consist of nuclei only. This is caused by processes in space that accelerate charged particles, stripping off the electrons from atoms, resulting in isolated nuclei and electrons. Electrons constitute about 1% of GCRs. It is not known why electrons are apparently less efficiently accelerated than nuclei. As a first approximation, the flux of GCRs in near-Earth space can be considered to be isotropic. See Diehl *et al.* (2001) for a short review concerning the astrophysics of GCRs. For more information about GCRs the following books are recommended: Clay and Dawson (1999), Dorman (2004), Friedlander (2000), Grieder (2001).

Due to randomly oriented magnetic fields in space, charged GCRs are subject to deflection when they propagate through space. During their travels they can gain or lose energy; collide with the interstellar matter, producing lighter nuclear fragments; and/or get lost in collisions. As a result, their directions as observed from Earth yield no information about the location of their original source. One must therefore use indirect methods to determine their sources and the way they propagate. Determining the chemical composition of GCRs is one such method.

The composition of GCRs is important because it is a direct sample of matter from outside the solar system and contains elements that are much too rare to be seen in spectroscopic lines from other stars. GCRs provide important information on the chemical evolution of the Universe. Comparing the elemental composition measured for GCRs with the composition of the Solar System shows some large differences (see Figure 5.2). As mentioned above, GCRs are nuclei and include essentially all of the elements in the periodic table (about 89% hydrogen, 10% helium, and about 1% heavier elements). As seen in Figure 5.2 the common heavier elements (such as carbon, oxygen, magnesium, silicon and iron) are present in about the same relative abundances as in the Solar System. But there are important differences in elemental and isotopic composition that provide information on the origin and history of GCRs. For example, there is a significant overabundance of the rare elements (lithium, beryllium and boron) produced when heavier GCRs such as carbon, nitrogen and oxygen fragment into lighter nuclei during collisions with the interstellar gas. There are also more GCR elements between silicon and iron than in the Solar System, but there is less hydrogen and helium.

Not only are GCRs deflected by the magnetic fields in interstellar space, but they are also affected by the IMF embedded in the solar wind, and therefore have difficulty reaching the inner Solar System. Since December 2004 the Voyager 1 spacecraft has observed that the intensity of GCRs increases with distance from the Sun (lower panel of Figure 5.1) (Burlaga *et al.* (2005)). This suggests that the termination shock of the heliosphere plays an important role in their exclusion.

However, some GCRs do reach the inner Solar System, and it is known that the

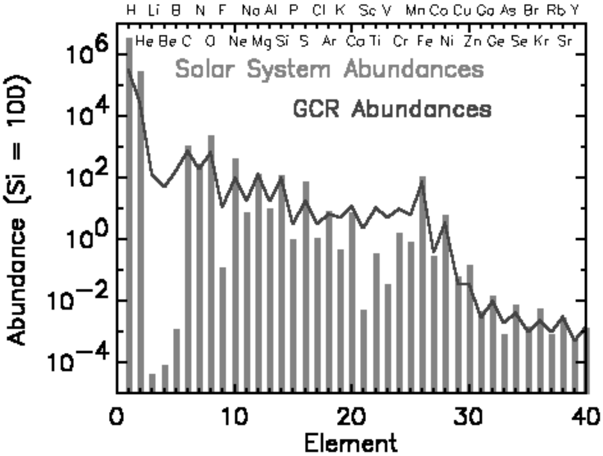


Figure 5.2. The elemental composition of cosmic rays (full line) compared with that of the solar system (individual columns). Courtesy of NASA/GSFC, see *http://imagine.gsfc.nasa.gov/docs/science/know_l2/cosmic_rays.html.*

GCR flux in the Solar System is modulated by solar activity. As solar activity varies over the 11-year solar cycle, the intensity of GCRs at Earth also varies, in anti-correlation with sunspot number. During solar maximum (when the sunspot number is at its highest) the increase in the IMF strength provides enhanced shielding of the heliosphere against penetrating GCR particles. Therefore, the GCR population is most intense during solar minimum. Solar radiation is the reason that the spectrum of primary GCR particles below some 10 GeV/nucleon cannot be directly measured (Section 5.2.1). Figure 5.3 (Svensmark, 1998) shows a measure of cosmic-ray flux, based on ion chambers which measure mainly the muon flux, covering the period 1937–1994 (middle curve). These data represent part of the high-energy GCR spectrum. Also plotted (upper curve) is data from the Climax neutron monitor (1953–1995) in Colorado, which measures the low-energy nucleonic part of the GCR spectrum. For comparison the relative sunspot number is plotted (bottom curve), which closely follows the solar 10.7-cm flux.

5.2.1 The energy spectrum

The differential GCR energy spectrum summed over all species (see Figure 5.4) extends from 1 GeV to somewhat above 10^{20} eV, and is based on measurements ('direct measurements' $< \sim 10^{13}$ eV $<$ 'indirect measurements') from different instruments (Swordy, 2001). Over this energy range the GCR flux (the number of GCR particles passing through a unit area surface in a unit time from a unit space angle) at different energies decreases by 24 orders of magnitude.

Within the energy range 10^{10}–10^{15} eV the energy spectrum is seen to be a simple power law with the spectral index 2.7:

$$dN/dE \sim E^{-2.7}\,\mathrm{m}^{-2}\,\mathrm{s}^{-1}\,\mathrm{sr}^{-1}\,\mathrm{GeV}^{-1}$$

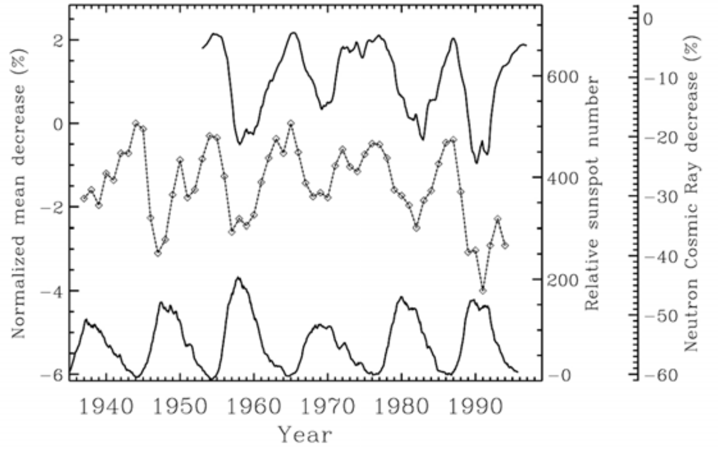


Figure 5.3. The top curve shows the cosmic ray flux from the neutron monitor in Climax, Colorado (1953–1996). The middle curve is the annual mean variation in cosmic ray flux as measured by ionization chambers (1937–1994). The neutron data has been normalized to May 1965, and the ionization chamber data has been normalized to 1965. The relative sunspot number is shown in the bottom curve. (From Svensmark, 1998.)

This is actually quite well understood in the framework of the conventional acceleration mechanisms for charged particles. Solar modulation (solar cycle effect shown in Figure 5.3) can account for the turn-over observed in the lower end of the energy spectrum (0.3–1 GeV/nucleon) (Figure 5.4).

Above 10^{15} eV there exist at least three irregularities in this otherwise simple form of the energy spectrum: one around 10^{16} eV, called 'the knee'; another around a few 10^{18}, called 'the ankle'; and one around a few tens of 10^{18} eV (not visible on Figure 5.4), called 'the toe' (Boratav, 2002). The two first structures are not totally understood, but reasonable hypotheses exist as to what is their cause (Erlykin, 2001). The last one is not understood at all, and it is widely agreed that the answers brought to the many open questions raised by 'toe physics' will no doubt open new windows in the fields of astrophysics, cosmology and/or fundamental interactions (Watson, 2001). (Section 5.2.2 describes in more detail these different groups of GCR; see also Gaisser (2000), Cronin (1999).)

5.2.2 Origin and acceleration mechanisms

Those GCRs below the 'ankle' on the energy spectrum (Figure 5.4) are generally thought to be mainly produced in our Galaxy and originate with supernovae explosions. Atomic nuclei crossing the supernova front will pick up energy from the turbulent magnetic fields embedded in the shock. The particle may be deflected in such a way that it crosses the boundary of the shock multiple times, picking up more energy on each passage, until it escapes as a GCR. A compelling reason for the 'supernova theory' is that the power required to maintain the observed supply of

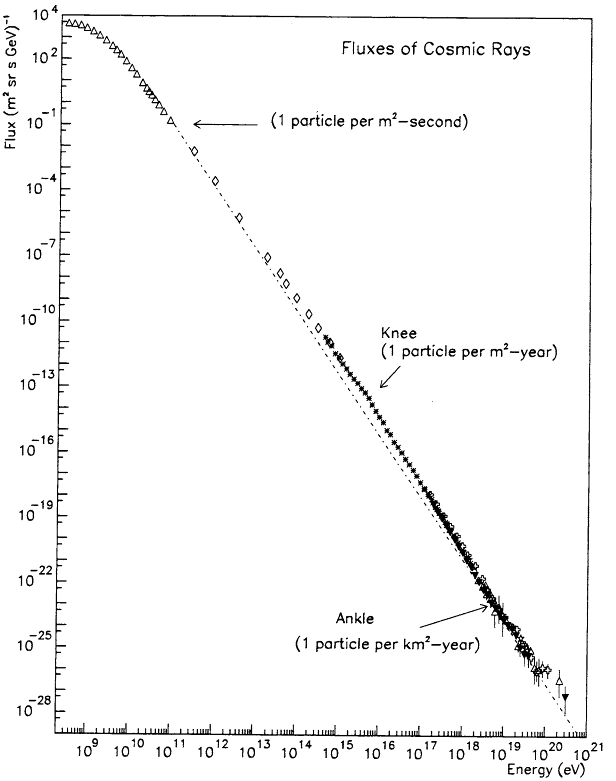


Figure 5.4. The differential energy spectrum of galactic cosmic rays. The arrows and values between parentheses indicate the integrated flux above the corresponding energies. (From Swordy, 2001.)

'cosmic-ray' nuclei in our Milky Way galaxy is only slightly less than the average kinetic energy delivered to the galactic medium by the three supernova explosions that occur every century (Cronin, Gaisser and Swordy, 1997). There are few, if any, other sources of this amount of power in our galaxy. Shocks associated with supernova remnants are expected to accelerate nuclei from the material they pass through, turning them into GCRs (Figure 5.5, see color section).

However, as mentioned earlier the composition of GCRs and the solar system is not always the same, suggesting that other acceleration mechanisms must also be at work. Cassé and Paul (1982) proposed that a significant fraction of heavy cosmic rays originate in Wolf-Rayet (W-R) stars, massive stars that are well-suited to accelerate cosmic rays through the stellar wind mechanism (Cassé and Paul, 1980). They suggested that W-R stellar wind material might be pre-accelerated by processes analogous to those in corotating interaction regions in the heliosphere and then further accelerated at the termination shock surrounding a W-R star. For more information about current cosmic ray theories and models see Mewaldt and Mason (2005).

Another method to determine the source of GCRs is by observing GCRs that have no electrical charge and so arrive at Earth undeflected by the galactic magnetic field. A small fraction (0.1%) of GCRs are photons in the form of gamma-rays, which are important when trying to find the origin of GCRs. By looking at the synchrotron radiation sometimes associated with supernova remnants, researchers have found more direct evidence that supernovae can act as accelerators. Synchrotron radiation is characteristic of high-energy electrons moving in an intense magnetic field of the kind that might act as a 'cosmic-ray' accelerator, and the presence of synchrotron X-rays in some supernova remnants suggests particularly high energies (Cronin, Gaisser and Swordy, 1997). Recently, researchers using a telescope array in Namibia found a stream of gamma radiation coming from a supernova remnant called RX J1713.7-3946. They suggest that the radiation is probably generated by the supernova remnant 'accelerating' cosmic ray electrons (Aharonian *et al.*, 2004).

In 1949 the physicist Enrico Fermi first provided an explanation for the acceleration of GCRs. The simplest acceleration mechanism is the stochastic and repetitive scattering by magnetic fields (second-order Fermi acceleration), by which plasma clouds roughly play the role of a magnetic mirror. A more efficient and faster process is acceleration by crossing shock fronts generated in explosive phenomena (first-order Fermi mechanism) such as supernovae. Fermi's theory, however, cannot account for the highest-energy cosmic rays (Boratav, 2002) that are discussed in Section 5.2.2.1.

5.2.2.1 Ultrahigh-energy cosmic rays

In the past decade researchers have found that the highest energy cosmic rays require a dramatically new explanation and possibly new physics. The cosmic ray observations above 10^{17} eV reported by the high-resolution (HiRes) Fly's Eye cosmic ray observatory (Abbasi *et al.*, 2005) show a break in the spectrum at $\sim 5 \times 10^{18}$ eV. The HiRes Prototype results showed a composition change from heavy to light in the 10^{17}–10^{18} eV range. This change in composition may reflect a change from a dominant Galactic cosmic ray flux to an extragalactic flux that dominates near 10^{19} eV. Observation of anisotropy from the Galactic plane would support this picture.

These particles are called ultrahigh-energy cosmic rays (UHECRs), and have energies exceeding 10^{20} eV. This is enough to accelerate the single proton to 99.99% of the speed of light. They hit the Earth at a rate of one per square kilometre per century (Figure 5.4).

Calculations show that supernovae cannot generate UHECRs (Semeniuk, 2003). Cosmic ray protons with energies above 5×10^{19} eV are likely to collide with the cosmic microwave background photons while traversing the Universe. The odds favour a collision every 20 million light years, with each collision costing the cosmic ray proton 20% of its energy. Consequently, no GCR can travel much further than roughly 100 million light-years before dropping below the 5×10^{19} eV energy threshold – the so-called 'Greisen-Zatsepin-Kuz'min (GZK) cut-off'. (For details about this value see Greisen (1966), Zatsepin and Kuz'min, (1966).) Therefore, UHECRs cannot be a result of quasars, gamma-ray bursts, or anything else at cosmological distances (Semeniuk, 2003). It is suspected that they are extragalactic in origin – perhaps generated in the cores of active galactic nuclei, in powerful radio galaxies, or by cosmic strings. These sources are known to have the tremendous amounts of energy needed to accelerate particles to such high energies, though a direct correlation has yet to be found. (For an introduction to UHECR physics, see Sokolosky (2004).)

5.2.3 Summary

There are still questions to be discussed and answered regarding GCRs, especially UHECRs, even though they have been studied for almost a century:

- Where do they originate?
- How are they accelerated to such high velocities?
- What role do they play in the dynamics of the Galaxy and the Universe?
- What does their composition tell us about matter from outside the Solar System?

New experiments, both on the ground and in space, will help us answer these questions. The impending Orbiting Wide-angle Light-collectors (OWL) is an Earth-orbiting system to study air showers initiated by $>10^{19}$-eV particles (Stecker *et al.*, 2004).

In Argentina the Pierre Auger Cosmic Ray Observatory is in the final stages of construction and has begun to collect data. It is a 'hybrid detector', employing two independent methods to detect and study high-energy cosmic rays. One technique is ground-based and detects high-energy particles through their interaction with water, while the other technique tracks the development of air showers by observing ultraviolet light emitted high in the Earth's atmosphere. (More information about the Auger Observatory can be found in Castelvecchi (2005).)

5.3 ANOMALOUS COSMIC RAYS

Anomalous cosmic rays (ACRs) are thought to represent a sample of the very local interstellar medium. They are not thought to have experienced such violent processes as GCRs, and they have a lower speed and energy. ACRs include large quantities of helium, nitrogen, oxygen, neon, argon and other elements with high ionization potentials; that is, they require a great deal of energy to ionize. These elements are therefore difficult to ionize, and the term 'anomalous cosmic rays' was given to these particles because of their unusual composition. They were first discovered in 1973 as a bump in the spectrum of certain elements (helium, nitrogen, oxygen and neon) at energies of about 10 MeV/nucleon (Garcia-Munoz, Mason and Simpson, 1973), and later in hydrogen (Christian, Cummings and Stone, 1988) and sulphur (Takashima *et al.*, 1997). They were observed to be modulated in phase with the GCRs during the solar cycle.

The outward flow of the solar wind excludes from the heliosphere thermal interstellar particles that are ionized, and permits only neutral particles to enter. Traditionally, ACRs are defined as those particles in the energy spectra of cosmic rays that originate as interstellar neutral gas atoms flowing into the heliosphere.

Fisk, Kozolovsky and Ramaty (1974) suggested that in interstellar material, elements with first ionization potentials (FIPs) below that of hydrogen would be ionized, while those with high FIP (helium, nitrogen, oxygen and neon) would remain neutral. Therefore, neutral atoms easily enter the heliospheric cavity, while ions are effectively excluded by magnetic fields.

When closer to the Sun, some of the penetrating neutral atoms become ionized (lose one electron) either by the loss of one electron in photoionization (the electron is knocked off by a solar ultraviolet photon) or by charge exchange (which involves giving up an electron to an ionized solar wind atom). Once these particles are charged, the Sun's magnetic field picks them up and carries them outward to the solar wind termination shock where they are eventually accelerated (Pesses, Jokipii and Eichler, 1981). As a result, ACRs at low energies are mostly singly-charged particles (Klecker *et al.*, 1995). Because ACRs are less than fully ionized, they are not as effectively deflected by the Earth's magnetic field as GCRs at the same energies.

The differential energy spectra for nitrogen, oxygen and neon, observed by the Geotail spacecraft, are shown in Figure 5.6. Note the typical GCR component above ~50 MeV/nucleon. Below this value the fluxes are steeply increasing, and this

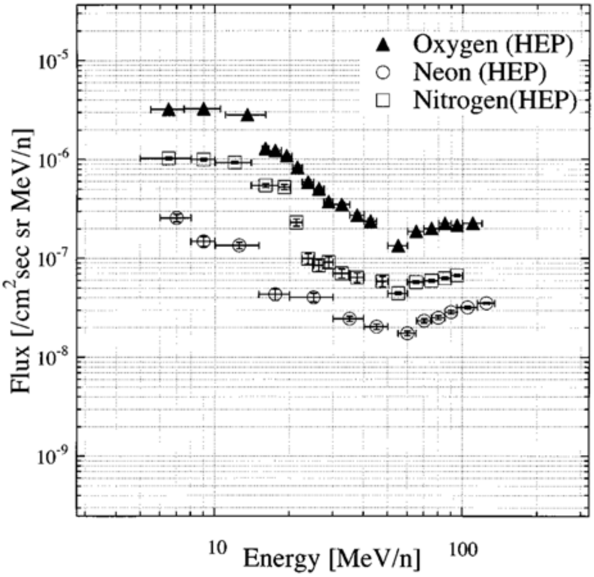


Figure 5.6. Differential energy spectra for N, O and Ne during quiet times from September 1992 to August 1995, observed by Geotail (from Takashima *et al.*, 1997).

enhancement in the flux is the anomalous component. (See Cummings, Stone and Steenberg (2002) for more information regarding the composition of ACRs.)

5.4 SOLAR ENERGETIC PARTICLES

The Sun is a sporadic source of cosmic-ray nuclei and electrons that are accelerated by shock waves travelling through the solar corona, and by magnetic energy released in solar flares. During such occurrences the intensity of energetic particles in space can increase by a factor of 10^2–10^6 for hours to days. Such solar energetic particle (SEP) events are much more frequent during the active phase of the solar cycle. The maximum energy reached in SEP events is typically 10–100 MeV, although CME-driven shocks can accelerate particles up to 20 GeV (Kahler, 1994). SEPs can be used to measure the elemental and isotopic composition of the Sun, thereby complementing spectroscopic studies of solar material.

SEPs are mainly protons, electrons and α-particles, with small admixtures of ^{3}He nuclei and heavier ions up to iron. Protons and ions can be accelerated up to some tens or hundreds of MeV/nucleon. The electron acceleration is limited to energies of some megaelectronvolts.

For several decades it was assumed that solar flares were the drivers of SEP events, and the 'flare myth' was born (Gosling, 1993). The earliest clear evidence that two distinct processes of particle acceleration contribute to SEP events came from radio observations (Wild *et al.*, 1963). Coronagraphs would later allow us to discover the world of coronal mass ejections (CMEs) – huge ejections of plasma in the Sun's outer atmosphere. The shock wave driven by a CME has been found to be an excellent particle accelerator. Hence, particles constituting an SEP can derive either from a solar flare or from the shock wave driven by a CME. (See Aschwanden (2004) and Chapter 3 for more information about solar flares and CMEs.)

Solar proton events (SPEs) constitute a sub group of SEPs that are very worrisome for interplanetary travel (see Chapter 11). The proton event threshold for which particles penetrate a space suit is 10 protons/(cm^2 s sr) at ≥ 10 MeV, and is also the threshold value that forecasters at the Space Environment Center (SEC) at the National Oceanic and Atmospheric Administration in Boulder, Colorado, USA, monitor to watch for the onset of an SPE. The October–November 2003 period was one of the largest outbreaks of solar activity in recent history. On 28 October 2003 a majestic SPE was recorded (see Figure 5.7) following the eruption of an intense solar flare and an associated CME.

5.4.1 Impulsive and gradual events

The terms 'gradual' and 'impulsive' SEPs originally came from the time scales of the associated X-ray events, each with distinct signatures and broad characteristics (Cane, McGuire and von Rosenvinge, 1986; Reames, 1995; Reames *et al.*, 1997). The characteristics of these two types of SEP are listed in Table 5.2.

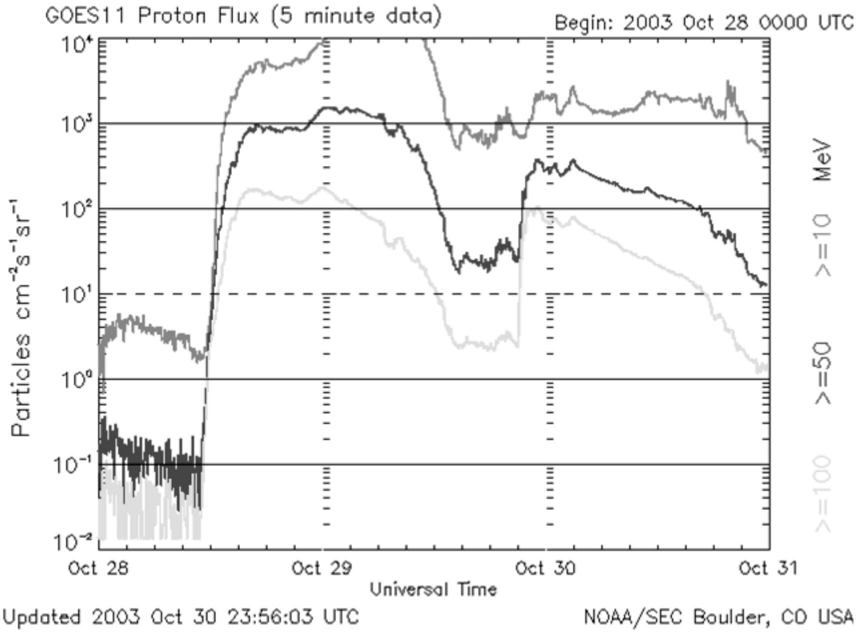


Figure 5.7. The solar proton event that began on 28 October 2003, reaching its maximum on 29 October 2003, observed at energies greater than 10, 50 and 100 MeV (highest to lowest proton flux) at geostationary orbit (courtesy of NOAA/SEC, see *http://www.sec.noaa.gov/Data/alldata.html*).

Table 5.2. Properties of impulsive and gradual events based on X-ray signature (Reames, 1995).

Properties	Impulsive events	Gradual events
Particles	Electron-rich	Proton-rich
^{3}He/^{4}He	~1	~0.0005
Fe/O	~1	~0.1
H/He	~10	~100
Q_{Fe} (mean ionization state)	~20	~14
Duration	Hours	Days
Longitude cone	<30°	~180°
Radio type	III, V (II)	II, IV
Coronagraph	–	CME (96%)
Solar wind	–	Interplanetary shock
Flares/year	~1,000	~10

Gradual events have durations of several days, they are proton-rich, and they have, on average, the same element composition and ionization states as those in the low-density ambient plasma of the high solar corona or solar wind. They are associated with gradual X-ray flares, type II and type IV radio emission, and are produced by the shock associated with fast CMEs (Kahler, 1994; Cane, McGuire and von Rosenvinge, 1986; Reames *et al.*, 1997). Such events are observed over a broad range of heliolongitudes – in some case over 180°.

Impulsive short-duration events, lasting several hours, are only observed at magnetically well-connected locations on the Sun. They are generally limited to within a 30-degree longitude band about the footprint of the nominal field line connected to the active region. They are electron-rich, and have a strong association with impulsive H_α and X-ray flares, and type III radio bursts. The high ion charge state indicates that their origin is in plasma heated by solar flares.

However, there are exceptions to this classification, and recent observations challenge this strict separation of all SEP events into these two types. For example, there are large gradual events with abundances more like those of impulsive events. It also seems that abundance variations are organized by the heliolongitude of the parent solar activity (solar flare or CME).

Recently, Smart and Shea (2003) have suggested an alternative classification of events based on other distinctions (near-Sun injection and interplanetary shock dominated):

- Near-Sun injection events are those that are the result of solar activity on the western hemisphere of the Sun near the 'favourable propagation position' for field lines connecting to an observer at Earth. This class covers both the 'impulsive' flare associated events and the western heliolongitude fast CME associated events.
- Interplanetary events are those associated with solar activity near the central meridian of the Sun (presumably the result of a very fast CME) and the resulting powerful fast interplanetary shock directed toward the observer that continues to accelerate particles at the shock front that propagates along the pre-existing interplanetary magnetic field lines toward the observer. These types of events may further be subdivided into two subclasses of (regular shock events and converging shock events).

The characterization of their classification is found in Table 5.3.

5.4.2 Solar proton events (empirical models and forecasting)

For interplanetary travel with humans onboard it is mainly the solar proton events (SPEs) that are the problem. Essentially, four SPE models are available to spacecraft engineers for predicting long-term solar proton fluences: the King model (King, 1974), the Jet Propulsion Laboratory (JPL) model (Feynman *et al.*, 1993), the Emission of Solar Protons (ESP) total fluence and worst-case-event models (Xapsos *et al.*, 1999, 2000), and the Moscow State University (MSU) fluence and peak flux model (available online at Moscow State University). Presently, the models rely only on data collected at 1 AU. Also, no event prediction is obtained by using the current models; only total mission fluence.

The usual method for estimating the energetic proton environment for a Mars mission is to take the solar proton observations at 1 AU (such as modelled by

Table 5.3. Space weather characteristics for the two types of SEPs based on the classification of Smart and Shea (2003).

	Type of SEP	
	Near-Sun injection	Interplanetary shock dominated
Solar origins	Western (front and far side) hemisphere: 'impulsive' flares and fast CMEs	Fast CMEs from: – Western (front and far side), rapid rise and decline – Halo (front sided), well-connected, flat profile – Eastern, poorly connected until shock passes
Time of arrival at 1 AU after the event onset at the Sun	~8–80 mn	~12 hours–2 days or more
Peak intensities	~5 orders of magnitude	~2 orders of magnitude
Duration	Limited (hours)	Several days
Radiation hazard	Low	High
Forecasting goals	Warning (probability) levels	Flux profile predictions

Feynman *et al.*, 1993) and then extrapolate these observations to other distances. It is assumed that the proton flux is confined to a magnetic 'flux tube' the volume of which will behave in the classical manner as the radial distance (R) from the Sun increases. From this purely geometrical argument, the peak flux extrapolations should behave as R^{-3} and the fluence extrapolations should behave as a function of R^{-2} (Smart and Shea, 2003).

Events dominated by a near-Sun injection of particles onto interplanetary magnetic field lines leading to the spacecraft position represent the classical solar particle event associated with solar activity. This class of event scales in radial distance by the classical power law extrapolation. The interplanetary shock dominated event generates a maximum flux as the shock passes the detection location, but the flux does not scale in the classical manner with radial distance (Smart and Shea, 2003). This discrepancy in scaling must be considered when estimating the possible solar proton flux encountered in interplanetary space during, for example, a manned mission to Mars.

It is important to note that perhaps the largest possible SPE event has not yet been measured, and that the largest SPE events used in the above empirical models originate only from the satellite era. McCracken *et al.* (2001a,b) analysed a total of 125 large fluence SPEs identified from the nitrate deposition in ice cores from Greenland for the period 1561–1950. These data have been augmented with ionospheric and satellite data for the period 1950–1994. There were five periods in the vicinity of 1610, 1710, 1790, 1870 and 1950 when large >30 MeV proton events with fluence greater than 2×10^9 cm^{-2} were up to eight times more frequent than in the era of satellite observations. The largest SPE in the nitrate record (associated with the Carrington white light flare event in 1859) had a $>30\,\mathrm{MeV}$ proton fluence that was in the range $18\text{–}36 \times 10^9\,\mathrm{cm}^{-3}$ (McCracken *et al.*, 2001b). This is a factor 4–8 times greater than the value for the August 1972 event, which is frequently regarded as the 'worst case' SPE.

CMEs and solar flares are the first links in the chain of causes that connects eruptions on the Sun to SEPs at 1 AU and beyond. To understand the solar origins of SEPs, it is essential that one understands how CMEs and solar flares are initiated. Because of their limited intensities and duration, impulsive solar flare events do not constitute a significant radiation hazard. However, most of the impulsive SEPs reach Earth nearly as rapidly as any electromagnetic signatures (~8–80 minutes). This leaves insufficient time to make appropriate changes during, for example, extravehicular activity, defined as when work is being done by an astronaut or cosmonaut outside his or her spacecraft. As a result, solar flare monitoring with present and impending spacecraft will provide only short-term warning, at best, of approaching energetic particles from solar flare events.

On the contrary, gradual SEP events are usually associated with CMEs that drive large particle events – especially SPEs, where peak intensities occur at the time of the shock passage. It is important to note that only when CME shock transit speeds exceed 500 km/s do SEP events become likely, while speeds greater than 750 km/s always produce SEPs (Reames, Kahler and Ng, 1997).

Including CME-driven shocks, there exist three available sites for *in situ* observations of shocks and accelerated particles:

- Interplanetary travelling shocks (particles accelerated at interplanetary shocks are called energetic storm particles).
- Corotating interaction regions (when the high-speed solar wind interacts with the low-speed wind, a shock front is created).
- Planetary bow shocks (solar wind-driven).

These acceleration sites, where particles are accelerated, are discussed in the next three sections.

5.5 ENERGETIC STORM PARTICLES

Energetic storm particles (ESPs) were originally thought to be particle enhancements related to the passage of an interplanetary shock. The name was chosen to reflect their association with the magnetic storm observed as the shock hits the Earth's magnetosphere. It is now known that the acceleration of particles at the shock, their escape, and the subsequent propagation through interplanetary space, is a continuous process, lasting for days to weeks until the shock finally stops accelerating particles.

Interplanetary shocks can accelerate particles up to the energy range of tens to hundreds of keV (Tsurutani and Lin, 1985; Wenzel, Reinhard and Sandersen, 1985), as well as the MeV range (Cane, Reames and von Rosenvinge, 1988; Kallenrode,

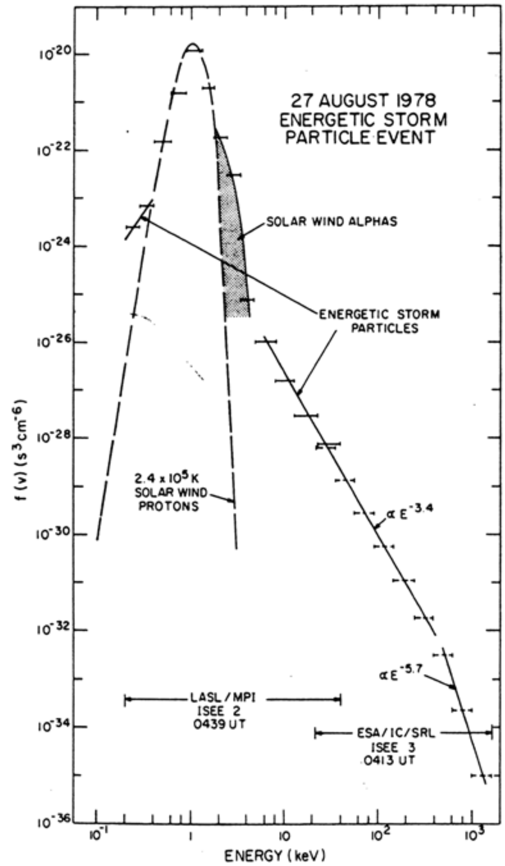


Figure 5.8. Energy spectrum for ions from 200 eV to 1.6 MeV in the spacecraft rest frame just downstream from an interplanetary shock. The double power-law spectrum corresponds to energetic storm particles. (From Gosling *et al.*, 1981.)

1996). With very strong shocks, protons can be accelerated up to about 100 MeV (Reames, 1990).

Figure 5.8 shows the energy spectrum for ions from 200 eV to 1.6 MeV in the spacecraft rest frame just downstream from an interplanetary shock. The double power-law spectrum corresponds to ESPs.

Energetic particles accelerated at interplanetary shocks can be separated into two energy bands:

- Low-energy component
  - Ion energy range (up to a few hundred keV/nucleon)
  - Electron energy range (up to some tens of keV)
- High-energy component
  - Ion energy range (MeV to tens of MeV/nucleon)
  - Electron energy range (hundreds of keV to MeV)

One reason for this classification is the break observed in the ion spectrum (see Figure 5.8). Another reason stems from different particle speeds relative to the shock. Protons in the low-energy range (tens of kiloelectronvolt) on average are only slightly faster than the shock and therefore stick close to the shock front. Thus an observer in interplanetary space at lower energies detects the shock of accelerated particles around the time of the shock passage. On the other hand, protons in the high-energy range (megaelectronvolt) are much faster than the shock and can easily escape from it. In this case the particles can be detected when the shock is still remote from the observer.

In conclusion: acceleration is more efficient if particles stay close to the shock and interact repeatedly, than for particles that escape easily from the vicinity of the shock. The steeper slope for the high-energy power-law spectrum in Figure 5.8 also indicates a less efficient acceleration at these energies.

5.6 COROTATING INTERACTION REGIONS

It has been well demonstrated by the Ulysses spacecraft that there exist two regimes of the solar wind (high speed and low speed). Near solar minimum, activity is focused at low altitudes, high-speed solar wind prevails, and magnetic fields are dipolar; whereas near solar maximum, the solar winds are slower and more chaotic, with fluctuating magnetic fields.

High-speed solar streams originate in coronal holes at the Sun. Coronal holes are associated with ‘open’ magnetic field lines, and are often found at the Sun’s poles, but may also be found at lower latitudes. As the Sun rotates, these various streams also rotate (corotation) and produce a pattern in the solar wind much like that of a rotating lawn sprinkler. The interaction of high speed (>400 to $\sim 800\,\mathrm{km/s}$) and low speed ($<400\,\mathrm{km/s}$) winds lead to three-dimensional corotating interaction regions (CIRs) in the heliosphere. If a slow-moving stream is followed by a fast-moving stream, the faster-moving material will catch up to the slower material and plough into it. This interaction produces shock waves at the front and rear edges called the forward shock and the reverse shock. Both are effective as energetic particle accelerators. A ‘forward shock’ continues in the direction of the overtaking fast wind, while the ‘reverse shock’ propagates inward into the high-speed stream.

A large amount of energy is concentrated in the compressed region close to the shock, and charged particles passing through a shock can acquire part of this energy. Possible sites for the acceleration are the forward and the reverse shocks which bound the CIR, as was suggested by Barnes and Simpson in 1976. At low solar latitudes, within the domain occupied by the wavy current sheet, the interaction of fast and slow solar wind is a common occurrence. Occasionally the shocks form at 1 AU, but they strengthen with distance as the plane of the interface becomes less radial and more azimuthal (Reames, 1999).

The acceleration of charged particles by CIRs leads to a permanent generation of high-energy particles in the heliosphere, even in times when the Sun is quiet (solar minimum), as there are more high-speed winds during this phase of the solar cycle. It is well established that energetic particles (up to approximately 10 MeV) are accelerated in association with CIRs in the solar wind (Barnes and Simpson, 1976; Fisk and Lee, 1980; Zöllich *et al.*, 1981).

Intensity time profiles of He ions over one solar rotation obtained at eight energies ranging from 44 keV/nucleon to 6.00 MeV/nucleon are shown in the lower panel of Figure 5.9 (see colour section). The upper panel shows the magnetic azimuthal angle to define the magnetic sector structure, and the middle panel the solar wind speed. It is clearly seen that at the onset of the high-speed solar wind on both 23 May and 30 May a particle event begins, each associated with the passage of a CIR. (See Richardson (2004) for a review on energetic particles and CIRs.)

5.7 PLANETARY BOW SHOCKS

In interplanetary space there are three different types of obstacle that deflect the solar wind flow:

- Planetary magnetospheres.
- Planetary ionospheres.
- Cometary and planetary atmospheres.

In this section it is planetary magnetospheres that are of interest. As mentioned in the previous section, solar wind speeds vary approximately between 300 and 700 km/s. This greatly exceeds the speed of Alfvén waves that are found in a low-density, magnetized and completely ionized gas, such as a planet's magnetosphere. Thus, a shock is formed upstream of the planetary magnetosphere that is imposed on the super-Alfvénic solar wind flow. See Russell (1985) for an introduction about planetary bow shocks.

In a planetary magnetosphere, the bow shock is the boundary at which the solar wind abruptly drops as a result of its approach to the magnetopause. The most well-studied planetary bow shock is the one that is created 2–3 Earth radii ahead of the magnetopause when the solar wind encounters the Earth's magnetopause. Like a rock in a flowing river, the Earth's magnetosphere acts, for the solar wind, as an obstacle, and the bow shock is formed. To first approximation the magnetic field of the planet deflects the plasma flow around it, carving out a cavity in the solar wind. The strongly heated and compressed plasma layer of deflected solar wind behind the bow shock is called the magnetosheath and is the boundary of the magnetosphere. The boundary between the magnetosphere and the magnetosheath is called the magnetopause. The solar wind generally pulls out part of the planetary magnetic field into a long cylindrical magnetotail, extending far downstream behind the planet.

The upstream region of the Earth's bow-shock region is characterized by turbulence and energetic particles. The occurrence of upstream events is more probable during high-speed solar wind streams, when the shock speed is highest relative to the upstream solar wind. The foreshock region, where the upstream solar wind is connected to the bow shock by the IMF, is illustrated in Figure 5.10. Energetic

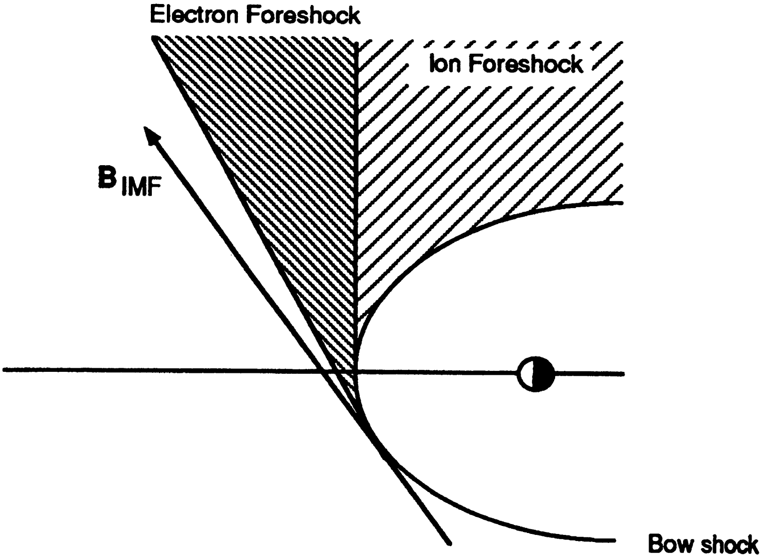


Figure 5.10. Schematic diagram of the foreshock region, where energetic electrons and ions can escape the IMF and populate the foreshock region. Courtesy of Parks (2004).

Table 5.4. Characteristics of ions accelerated at Earth's bow shock.

Local geometry	Ion acceleration mechanism	Local geometry	Accompanying waves
Upstream	Shock drift	Quasi-perpendicular	Low-amplitude (~ 1 Hz)
Dawn side	Diffusive shock	Quasi-parallel	Low frequency with large amplitudes

electrons and ions can escape the shock along the IMF and populate the foreshock regions. At the Earth's position the IMF spiral has an angle of 45° with respect to the Sun–Earth line. Depending on the local geometry, different ion distributions and different types of waves can be observed (see Table 5.4):

- Upstream of the bow shock, a foreshock region develops and is characterized by energetic particles streaming away from the shock front and by waves excited by these particles. Close to the nose of the shock the geometry is quasi-perpendicular, and shock-drift acceleration leads to a particle distribution in the form of a rather narrow beam of reflected ions. The wave field consists of low-amplitude waves with frequencies of about 1 Hz.
- At the dawn side the geometry is more quasi-parallel, and particles are accelerated by diffusive shock acceleration. The waves excited by these ions are again low-frequency waves with large amplitudes, occasionally containing shocklets in the sense of discrete wave packets which are often associated with discrete beams of particles.

Between these two extremes the local geometry is oblique, and both mechanisms contribute to the particle acceleration. The resulting particle distribution is an intermediate distribution in which both the reflected beam and the more diffusive population can be found. The accompanying wave field basically consists of transverse low-frequency waves.

Electrons also form a shock region. Starting close to the nose of the magnetosphere, where the geometry is quasi-perpendicular, the electrons are accelerated by shock-drift acceleration. Since the gyro-radii of the electrons are much smaller than those of the protons, their drift path along the shock front is shorter, leading to an earlier escape and therefore a more extended foreshock region.

Compared with the particle populations observed at interplanetary shocks, the bow-shock particles, although significantly more energetic than the solar wind, are still low-energy particles. Electron acceleration is observed only up to a few keV, and ion acceleration up to some tens of keV (Cornwall, 1986).

Waves and particle upstream of bow shocks can also be found on other planets (Saturn, Uranus, Jupiter, and so on) (Orlowski, Russell and Lepping, 1992; Russell and Hoppe, 1983; Russell, Lepping and Smith, 1990). Jupiter's magnetosphere is the largest in our Solar System, and at its bow-shock electrons are accelerated up to about 10 MeV. With a suitable magnetic connection between Earth and Jupiter, these jovian electrons can be observed even at Earth's orbit (Chenette *et al.*, 1977; Conlon, 1978). Figure 5.11 shows a plot of 48-hour averages of ~1.5–11.5 MeV

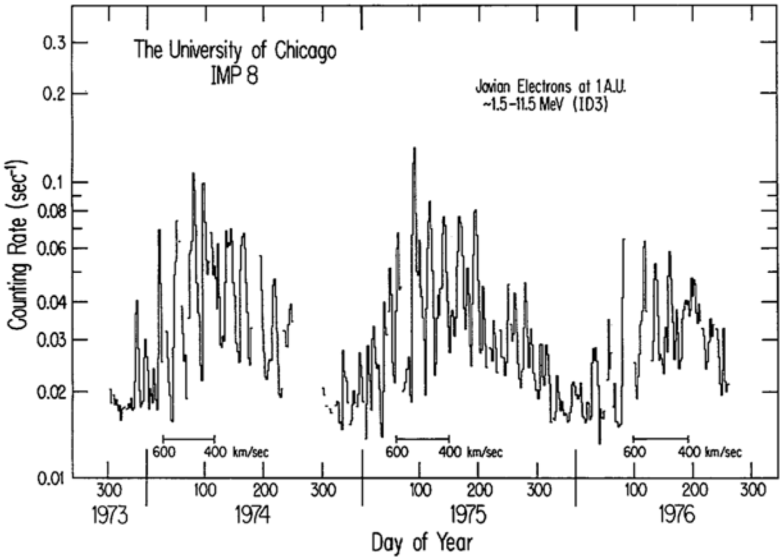


Figure 5.11. Forty-eight-hour averages of the IMP-8 ~1.5–11.5 MeV (ID3) electron counting rate from the launch of the satellite on day 303, 1973 into 1976. The data have been filtered to eliminate the contribution of solar flare and geomagnetospheric electron events (from Chenette *et al.*, 1977).

electron count rates performed by an instrument aboard the IMP-8 spacecraft. Solar and geomagnetospheric electron events have virtually all been removed using a technique presented in Chenette *et al.* (1977). The horizontal bars in Figure 5.11 indicate the intervals over which Earth and Jupiter are connected along the direction of the IMF for a range of solar wind speeds from 600 to 400 km/s.

5.8 GEOMAGNETICALLY TRAPPED PARTICLES

Space missions throughout the Solar System have shown that most of the planets are magnetized. In particular, the giant planets are magnetized much more strongly than Earth (Bagenal, 1992), and their magnetospheres (Russell, 2004) are all much larger than Earth's – in part because of the stronger dipole moments, and also because the solar wind becomes increasingly rarefied far from the Sun (see Table 5.5). Mercury has a weak magnetic moment, about 4×10^{-4} of that of the Earth (Russell and Luhmann, 1997). Combined with a solar wind pressure, about seven times larger than the pressure at Earth, the result is a very small planetary magnetosphere (in both absolute dimensions, and relative to the size of the planet). Venus has no detectable intrinsic magnetic field (Russell *et al.*, 1980).

Unlike Earth, Mars does not have a substantial global magnetic field with which to deflect cosmic rays away from the planet (Aplin, 2006; Molina-Cuberos *et al.*, 2001). However, weak magnetic fields below the ionosphere exist on Mars (≤ 5 nT) (Acuña *et al.*, 2001), arising from magnetization of the martian crust (Connerney *et al.*, 2001).

Earth's magnetic field causes particles (protons and electrons) to become trapped in the radiation belts. Due to the intense magnetic fields of the gas giant planets (Jupiter, Saturn, Uranus and Neptune), these planets all have radiation belts similar to Earth's outer electron belt. The first part of this section provides an introduction to Earth's radiation belts (for more detailed information see Chapter 6 and Walt (1994)), and is followed by a description of radiation belts that are found on other planets in our Solar System.

Table 5.5. Magnetic fields of Earth and the giant planets in our Solar System (Bagenal, 1992).

Planet	Earth	Jupiter	Saturn	Uranus	Neptune
Radius, km	6,378	71,398	60,330	25,559	24,764
Spin period, hrs	24	9.9	10.7	17.2	16.1
Magnetic Moment/M_{Earth}	1	20,000	600	50	25
Mean equatorial field (Gauss) 1 Tesla = 10^4 Gauss	0.31	4.28	0.22	0.23	0.14
Dipole tilt and sense	+11.3°	−9.6°	0.0°	−59°	−47°
Solar Wind density (cm^{-3})	10	0.4	0.1	0.03	0.005
Distance to 'nose' (planet radii)	11	50–100	16–22	18	23–26

5.8.1 Earth's radiation belts

The heliosphere is the first defence in the Earth's three-layer system of shielding against high-energy GCRs, which consists of: 1, the heliosphere with the Sun's magnetic field carried out by the solar wind; 2, the Earth's magnetic field (magnetosphere) (which acts as a shield against the Sun's radiation and magnetic fields); and 3, the Earth's atmosphere. The latter two also protect Earth from the solar wind itself. Earth's magnetosphere exists as Earth is a magnetized planetary object embedded in a flowing plasma. Particles are not able to penetrate to the Earth's surface but are forced by the magnetic field to move around the Earth.

The geomagnetic field of Earth is basically a dipolar field, with the magnetic flux tubes emanating from the Antarctic and entering the Arctic. The magnetosphere is formed as a cavity in the solar wind flowing past it, with dayside field lines being compressed and nightside field lines drawn out into a comet-like tail (Figure 5.12, see colour section). Particles gain entry through the cusps that are shaped like funnels over the polar regions, or gain entry far downstream from the Earth. The particles that enter downstream travel toward the Earth and are accelerated into the high-latitude ionosphere and produce the auroral oval light shows. Other higher energy particle radiation that could pose a danger to life here on Earth is forced to drift around the Earth within two large doughnut-shaped regions called the radiation belts. Invisible magnetic fields are the reason that particle radiation moves in this way.

Earth's radiation belts are principally composed of naturally occurring energetic charged particles trapped in Earth's inner magnetosphere at equatorial distances ranging from approximately 1.2 to 7 Earth radii (Re), and are the main contributor to Earth's natural radiation environment. SEPs and GCRs also contribute to this environment, but on a smaller scale. The two doughnut-shaped rings (inner and outer belt) consist mainly of protons and electrons, although heavy ions (at least up to iron) are also present. It was James Van Allen, who headed the team of scientists investigating these bands of particles, who insisted that a Geiger counter for particle detection be included in the payload on the Earth satellite Explorer 1 in 1958 (Van Allen *et al.*, 1958; Van Allen and Frank, 1959). For this reason, Earth's radiation belts are also better known as the Van Allen belts.

The motion of charged particles in magnetic and electric fields can be calculated from the Lorentz force law. These calculations show that charged particles can be trapped by magnetic fields. Electrons drift eastward around the Earth, while ions drift westward. They bounce between the stronger magnetic fields in the northern and southern hemispheres and gyrate around the local magnetic field. They execute three characteristic types of motion with characteristic time scales: gyro (function of magnetic field strength and the mass of the particle), bounce (function of kinetic energy of the particles), and drift (dependent on the energy of the particles and their pitch angle).

5.8.1.1 The inner belt

Earth's inner proton radiation belt is a by-product of cosmic rays that enter Earth's magnetosphere. The main source for the protons trapped in the inner belt is Cosmic Ray Albedo Neutron Decay (CRAND) (Albert *et al.*, 1998, and references therein). The primaries are cosmic rays that interact with the Earth's atmosphere, creating secondaries (neutrons by spallation) which are partly deflected (albedo = reflectivity) into the radiation belts, where the neutrons decay into protons which then become trapped.

This is a fairly stable population of protons, but is subject to occasional perturbations due to geomagnetic storms, and varies with the 11-year solar cycle. It is the combination of two effects which leads to a solar cycle variation of the proton flux levels in the inner radiation belt, with high fluxes observed during solar minimum and low fluxes during solar maximum. The first concerns the fact that when the Sun is active the IMF is stronger, as is its shielding effect; and as a result, fewer GCRs arrive in the vicinity of Earth (see Section 5.2). The opposite is therefore seen when solar activity is low. The density of the upper atmosphere is enhanced during solar maximum because of the enhanced solar heating, causing the second effect – an enhanced absorption of protons in the upper atmosphere during solar maximum, thus a strong loss.

The South Atlantic Anomaly is a feature of the inner radiation belt occurring where the inner belt reaches a minimum altitude of about 250 km above the Atlantic Ocean off the Brazilian Coast. It is caused by the offset between the Earth's geographical and magnetic axes.

5.8.1.2 The outer belt

The outer electron belt which extends from about 3 Re to geostationary orbit and beyond at active times (7–9 Re) in the equatorial plane, consists mostly of electrons with energies below 10 MeV. Figure 5.13 shows two periods of observation for the electron count rate data ($E > 750\,\mathrm{keV}$) obtained by the Cold Ion Detector (CID) aboard the Space Technology Research Vehicle (STRV-1a), and it can be seen how the boundaries of the electron radiation belt changes over time as it expands and contracts. Indeed, the observations emphasize the dynamical aspect of the electron radiation belts on all spatial and temporal scales. Unlike the proton belt, the electron belt exhibits substantial variation on short time scales related to magnetic storms and high-speed solar winds (minutes to days). Also observed is the long-term variation from days to years, associated with solar rotation, seasons and the solar cycle.

Electrons within Earth's magnetosphere derive from two main sources: the solar wind and the ionosphere. The typical temperature of electrons in the ionosphere is less than 1 eV, and for solar wind electrons it is about 10 eV. Yet Earth's electron belts range in energy from approximately 200 keV to about 15 MeV (Li and Temerin, 2001). There are many unanswered questions regarding the electron radiation belt:

- How do these electrons come to be energized? Acceleration mechanism theories that have been proposed in the past range from electrons from Jupiter,

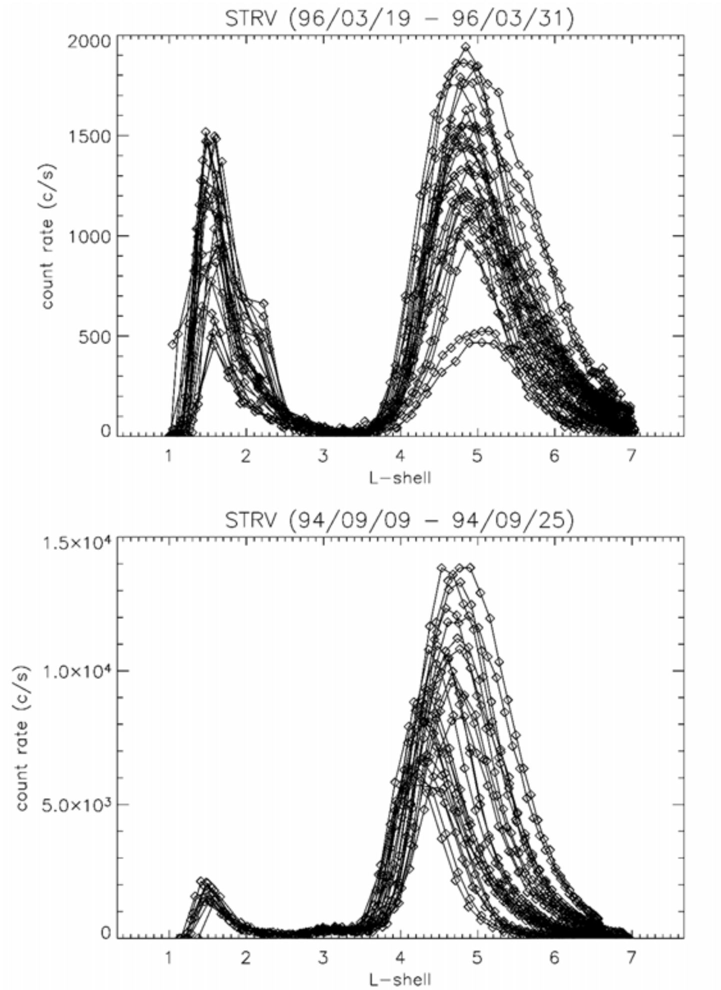


Figure 5.13. CID/STRV electron count rate as a function of L-shell for two periods of observation. (From Crosby *et al.*, 2005.)

shock acceleration, classical radial diffusion, ULF enhanced radial diffusion, wave–particle interactions, recirculation, direct substorm injection, adiabatic effects, and cusp injection.

- What is the dominant loss mechanism? Proposed loss mechanisms include detrapping due to E-fields, Coulomb fields, and wave–particle interactions.
- What is their relation to the solar driver (fast solar wind streams, CMEs and magnetic storms)?
- What are their effects on the atmosphere (chemistry and ozone)?

An important challenge has been to explain how the charged particles within the electron belts are accelerated to very high energies of several million eV. Recently, Horne *et al.* (2005) showed, on the basis of the analysis of a rare event where the outer radiation belt was depleted and then reformed closer to the Earth, that the long-established theory of acceleration by radial diffusion is inadequate. They found that the electrons are accelerated more effectively by electromagnetic waves at frequencies of a few kHz. Wave acceleration can increase the electron flux by more than three orders of magnitude over the observed time scale of 1–2 days – more than sufficient to explain the new radiation belt. The authors add that wave acceleration could also be important for Jupiter, Saturn and other astrophysical objects with magnetic fields (Horne *et al.*, 2005).

5.8.1.3 Empirical modelling of the radiation belts

Empirical modelling has been, and is still, the preferred approach for simulating Earth's radiation belts. For engineering applications the NASA AP-8 and AE-8 models (Vette 1991a, b; Sawyer and Vette, 1976) for protons and electrons, respectively, still are the *de facto* standard. Up to now they are the only models that completely cover the region of the radiation belts, and have a wide energy range for both protons (0.1–400 MeV) and electrons (0.04–7 MeV). Figure 5.14 shows the average structure of the radiation belts, as described by these models. The left panel shows integral flux level contours of > 10 MeV protons, and the right panel, contours of the > 1.0 MeV electrons. The inner and outer zones of the radiation belts have their largest fluxes centred at Re ~ 1.5 and near Re $= 4$ (but higher energies at lower L values), respectively. Whereas protons are restricted to the inner zone, electrons can be found in both regions. The zone between the two regions is known as the slot region.

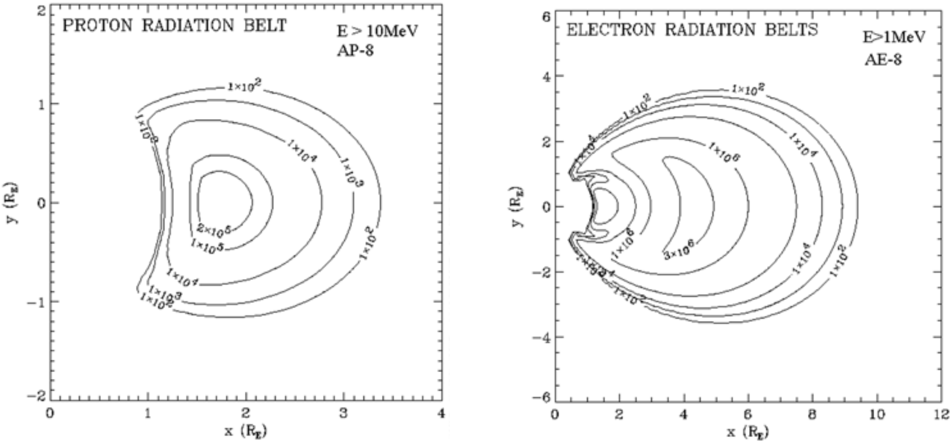


Figure 5.14. Contour plots of the electron and proton radiation belts. Omnidirectional fluxes are for particles > 1 MeV and > 10 MeV respectively. The data are derived from the AE-8 and AP-8 models. Courtesy of ECSS Space Environment Standard (ECSS E-10-04).

It was originally presumed that the radiation belts were stable environments, and therefore only two versions (solar maximum or solar minimum) were necessary for engineering purposes. Thus, apart from separate versions for solar minimum and solar maximum, the NASA AP-8 and AE-8 models that rely on data obtained in the 1960s do not include variations shorter than the traditional 11-year solar cycle.

Better models (empirical and theoretical) are needed both for engineering purposes as well as for forecasting. Existing magnetic field models are not accurate for disturbed times. An accurate electron model needs to be developed to explain the many unanswered questions listed in Section 5.8.1.2. Better satellite measurements are especially needed.

5.8.1.4 New radiation belts

It is also possible to form temporary belts during large SEP events when the associated CME and shock strike the Earth (Blake *et al.*, 1992). The large perturbation in the magnetosphere allows sudden trapping of SEP ions and electrons that have filled the outer magnetosphere (Hudson *et al.*, 1997, 1998). The particles are transferred to the inner shells ($L = 2$), and these new radiation belts can last for a period of months.

Sudden storm commencement compressions of the dayside magnetopause have in the past illustrated the rapid formation of new radiation belts on the particle drift time scale; for example, the 24 March 1991 event (Vampola and Korth, 1992; Blake *et al.*, 1992; Looper *et al.*, 1994).

5.8.2 Radiation belts of other planets

A couple of years before the discovery of the Van Allen belts, Burke and Franklin (1955) discovered powerful radio emissions from Jupiter. By the late 1960s it was clear that Jupiter's radio emissions were being generated by energetic electrons trapped in a strong magnetic field. Unfortunately for future space travellers, and of real concern to the designers of spacecraft that venture into this region of interplanetary space, the environment near Jupiter contains high levels of energetic particles (protons and electrons) trapped by the planet's magnetic field.

The Galilean moons – especially Io, which loads Jupiter's magnetosphere with many ions of sulphur and sodium from its volcanoes – affect not just their local environment but also the jovian ionosphere at the ends of the flux tubes connected to the moons. Moreover, the mass added to the magnetosphere by Io affects much of the rest of the magnetosphere. The magnetosphere is energized by this mass-loading, powering the aurorae, accelerating radiation-belt particles, and generating radio emissions. (See Bagenal *et al.* (2004) for a global overview about Jupiter and Russell (2005) for a general review about the interaction of the Galilean moons with their magnetosphere and ionosphere.)

With the exception of the Jovian magnetosphere, the radiation belts of the outer magnetospheres behave very much like those of the Earth's magnetosphere. Processes such as radial diffusion and pitch-angle diffusion act to transport

Table 5.6. Peak energetic particle fluxes (Kivelsen and Russell, 1995).

	Electrons		Protons	
Planet	Flux ($cm^{-2} s^{-1}$)	Energy (MeV)	Flux ($cm^{-2} s^{-1}$)	Energy (MeV)
Earth	10^5	≥ 3	10^4	≥ 105
Jupiter	10^8	≥ 3	10^7	≥ 80
Saturn	10^5	≥ 3	10^4	≥ 63
Uranus	10^4	≥ 3	< 10	≥ 63

particles across field lines and cause the particles to precipitate into the atmosphere and be lost (Kivelsen and Russell, 1995). As mentioned above, wave acceleration could be important for Jupiter, Saturn and other astrophysical objects with magnetic fields (Horne *et al.*, 2005).

The electron radiation belts of the different belts are similar. They are most intense just above the atmosphere (except at Saturn, where the fluxes maximize just outside the rings). At lowest altitudes the spectrum is harder; that is, the flux decreases less sharply with increasing energy than at high altitudes. However, the peak fluxes differ immensely. In Table 5.6 the peak electron flux at Jupiter is about 1,000 times greater than that at Earth (> 3 MeV), and the peak flux at Uranus is an order of magnitude less than that at Earth.

Similar tendencies are observed in the proton radiation belts. The proton belts of the various outer planets look very similar, but the fluxes (see Table 5.6) reveal an excess of three orders of magnitude at Jupiter and a deficit of three orders of magnitude at Uranus.

5.9 INTERPLANETARY SPACE WEATHER AND THE IMPLICATIONS

In the previous sections we have seen that the primary radiation sources of the interplanetary environment are energetic protons and heavy ions during SEPs, with energies up to a few 100 MeV, and GCRs, which consist of protons and heavy ions with energies in the GeV range. Whereas GCRs are isotropically distributed (Section 5.2), SEPs are directional and sporadic (Section 5.4). For some transportation scenarios the Earth's proton belts may also be a factor. For manned spaceflights, the largest concerns are SPEs with 70–100 MeV energies, capable of penetrating a space suit or vehicle skin and affecting the blood-forming organs (see Wilson *et al.*, 1997 and Chapter 11).

Comparing to terrestrial conditions, and using the well-know clicks of a Geiger counter, it can be said: 'One click per second on Earth, a hundred clicks per second in Deep Space, except in a large flare, when the click rate would be "off-scale".' Written by Andrew Holmes-Siedle. This was adopted as the motto of the European Space Agency's Radiation Exposure and Mission Strategies for Interplanetary Manned Missions (REMSIM) project (Foullon *et al.*, 2005; Cougnet *et al.*, 2005). The quote illustrates that the health risks from cosmic rays in deep space are much more severe than on Earth. REMSIM consisted of various work packages related to current strategies and countermeasures to ensure the protection of astronauts from radiation during interplanetary missions, taking into consideration radiation environment and its variability, radiation effects on the crew, transfer trajectories and associated fluences, vehicle and surface habitat concepts, passive and active shielding concepts, and space weather monitoring and warning systems.

In the context of what has been discussed in the previous sections of this chapter, the contents of REMSIM work-package WP5000 'Radiation Hazard and Space Weather Warning System' dealt with space science and warning issues (Foullon *et al.*, 2004). Part I dealt with the science of solar precursors, especially in regard to CMEs and with existing monitoring and warning systems, including a review of appropriate radiation monitor technology. The latter part concerned hardware details, internal radiation environment models, data handling and management methods, drawn from current mission science and terrestrial science. The team also proposed the measures required to manage space weather and radiation detection issues in deep space missions, especially the Human Mars Mission. Some specific recommendations for warning systems were made. (For overviews see Foullon, Crosby and Heynderickx (2005) and Kumar (2005).)

Relevant considerations for spaceflight beyond the Earth's magnetosphere include the structure of the spacecraft, the materials used to construct the vehicle, extravehicular activity start time and duration, the interplanetary proton and heavy ion flux, and the position in the solar cycle. Adequate radiation protection measures must be conceived for any lengthy interplanetary endeavours. Classical engineering includes the design, construction and implementation of radiation shielding for the interplanetary space vehicle. Storm shelters will be necessary both on the transit spacecraft and on the planet's surface. Interplanetary space weather monitoring and forecasting is of utmost importance for any mission. It encompasses new space environment monitoring scenarios as one will not only be able to rely on classical Lagrangian L1 point solar monitoring, especially for the mitigation of SPE events.

Future interplanetary manned missions will need to consider solar activity (solar flares, CMEs, solar wind, and so on) very carefully due to the obvious detrimental effects of radiation on humans. Very high doses during the transit phase of a mission can result in radiation sickness or even death (see Chapter 11). This is equally true for extended visits to surfaces of other planets (for example, to Mars) and moons lacking a strong magnetic field capable of deflecting solar particles. The risk of developing cancer several years after a mission is somewhat more difficult to quantify, but must also be considered in mission planning.

In August 1972, between the Apollo 16 and 17 manned missions, one of the largest SPEs ever recorded arrived at Earth. Computer simulations of the radiation levels an astronaut inside a spacecraft would have experienced during this event, found that the astronaut would have absorbed lethal doses of radiation within 10 hours after the start of the event (Turner, 1996; Hanslmeier, 2003). However, a number of studies have concluded that a properly shielded astronaut would not have been exposed to radiation levels above the monthly recommended limit

(Wilson *et al.*, 1997, 1999). It was found that a 10 g/cm^2 aluminium shelter would have provided adequate protection. With sufficient spacecraft shielding, SPEs will not prevent humans from going to Mars, but having a reliable SPE prediction capability will have high priority in order to minimize their impact on future interplanetary manned missions. The improved understanding, combined with the observatories and sensors that are or soon will be available, can provide the comprehensive space weather data necessary to implement physics-based SPE risk management (Turner, 2000).

Space agencies in Europe, USA, Russia, China, Japan and India are all considering future mission scenarios to Mars, with the Moon as a first stop. No doubt the next 50 years will see more countries joining in on these efforts, and will emphasize the importance of countries working together to reach these goals.

In the next part of this section a look at the various parameters that must be taken into consideration when considering a Mars trip will be discussed, especially in regard to interplanetary space weather monitoring.

5.9.1 Case study: mission to Mars scenario

Terrestrial space weather warning systems rely on early warning from the L1 point, where detection of SEPs provides a warning of less than one hour. These existing systems could also be used for Moon missions by relaying the L1 alert to the onboard warning system. However, this would not be feasible for an interplanetary mission at locations away from the Sun–Earth line. A plausible Mars space weather warning system relies mainly on four parameters:

- the orbit of Mars;
- telecommunications (signal travel time);
- location of Earth and Mars in relation to forecasting; and
- radial extrapolations of SEPs.

Mars' orbit is slightly elliptical, ranging from about 1.4 AU to about 1.6 AU. The motion of Mars, which is farther from the Sun than Earth, is slower. As Earth keeps racing ahead and Mars falls behind, there are instances when the two planets form a straight line, with the Sun interposed. At such times the planets are said to be in conjunction. Mars disappears from Earth's view behind the disk of the Sun, and is about 400 million km away from Earth. Thus, Earth and Mars move in and out of favourable phasing for transfers to and from each other. It takes approximately 26 months before the phasing is appropriate, and this is the approximate minimum trip time required for a manned mission to Mars.

The Earth–Mars distance varies from about 56 million to 400 million km, when the Earth is on the opposite side of the Sun from Mars – an important factor for telecommunications from Earth. The time required for radio communication from Earth to a spacecraft on Mars varies from 3.1 up to 22.2 minutes. It is important to mention that when Mars is at conjunction, communication is not possible at all. For instance, during August 2002, at the time of conjunction, no data from the Martian Radiation Environment Experiment (MARIE) instrument onboard Mars Odyssey were received for approximately two weeks (Atwell *et al.*, 2003). MARIE's mission was to assess the radiation environment at Mars to determine the radiation risk that astronauts on a Mars mission may encounter, and it stopped working properly on 28 October 2003 during a time of extreme solar activity (see Figure 5.7).

The Mars–Sun–Earth angle is an important parameter for an Earth-based space weather system. Real-time monitoring from Earth is limited to the solar disk and limb activity and to CME observations on the plane of the sky. This is sufficient to give warnings of near-Sun injection events for the period of time when Mars is near opposition or connected to the western solar hemisphere as seen by an observer on Earth. When Mars is magnetically connected to the limb regions as seen from Earth, the predictions may be based on chromospheric and coronal activity observations. Existing alternative techniques based on far-side imaging services and corotation projections may provide additional information for the limb regions and some information on the far side, but risk estimates would be much less reliable when Mars is connected to these longitudes. However, helioseismological methods allow one to detect the appearances of newly emerging magnetic flux on the solar surface at the back side of the Sun which appears a promising tool for advanced warning capabilities. See Bothmer (2006) for more information regarding future solar forecasting scenarios.

The usual method for estimating the energetic proton environment for a Mars mission is to take the solar particle observations at 1 AU and then extrapolate these observations to other radial distances (1.4–1.6 AU). However, the radial gradient associated with the shock acceleration is not well understood. As mentioned in Section 5.4.1, Smart and Shea (2003) suggest that the radial extrapolations expected by a power law geometry only apply to specific types of well-connected solar flare-associated events, but do not apply to the case of general shock accelerated events. Therefore, the method of radial extrapolation is presently not reliable. In order to improve the solar particle fluence models on Mars and provide more reliable estimates for mission planning, models need to be updated with data collected by missions to Mars and interplanetary space.

Comparing Mars' orbital parameters (the Earth–Sun–Mars angle, the Mars–Earth distance and the Sun–Mars distance) with an Earth-based space weather system (Table 5.7) tell us what strategies to adopt for the future development of interplanetary space weather forecasting, and results in the following requirements (bottom of Table 5.7):

- Onboard warning and forecasting system.
- Multi-viewpoint system.
- Test of comprehensive simulations of the interplanetary medium to fit readings from numerous unmanned missions.

These strategies imply having real-time observations from simultaneous missions at complementing positions and orbits with appropriate and complementary instruments, improving the reliability of onboard warning and forecasting systems, and enhancing the development of interplanetary models to be tested against data.

Table 5.7. Mars' orbital parameters versus Earth-based space weather system, and consequences for forecasting strategies. R is the radial distance from the Sun. (From Foullon *et al.*, 2004.)

Mars' orbital parameters	Terrestrial Space Weather System		
	Telecommunications	Observations	Models and predictions
Earth–Mars distance	Radio travel time: 3.1–22.2 light-minutes		
Mars–Sun–Earth angle	Mars at conjunction: $\sim$2 weeks without contact	Limited to solar disk and limb activity Limited to limb-to-limb radio bursts and CME observations on the plane of the sky	Accuracy of warnings for near-Sun injection events compromised for most of the mission duration Flux profile predictions for IP-shock dominated events, most compromised for far-sided events with Mars near conjunction
Sun–Mars distance (1.4–1.6 AU)			Radial extrapolations: solar wind speed (effect on longitudes of IMF lines connected to the observer); no classical gradients for particle flux ($\sim R^{-3.3}$) and fluence ($\sim R^{-2.5}$)
Favourable Mars–Earth phasing			Mission start and duration: predictions with respect to solar cycle; input to proton fluence models
Requirements	Onboard warning and forecasting system	Multi-viewpoint system	Test of comprehensive simulations of the interplanetary medium to fit readings from numerous unmanned missions

5.9.1.1 Interplanetary space weather monitoring and forecasting

In designing radiation protection systems for a manned vehicle on an interplanetary mission, radiation detectors which provide daily awareness of the background rate and incipient surges of radiation are an important subsystem which can save lives. For the study of dosimetry and warning systems, two space environment conditions are considered (Foullon *et al.*, 2004, Chapter 6).

(1) Galactic cosmic-ray background
    - A chronic condition (perpetual).
    - Minimize by ergonomics.
(2) Solar flare/CME – occasional surges of radiation due to solar particle events.
    - An emergency condition (infrequent).
    - Minimize by shelters.

Both conditions rely on reliable radiation monitors that can track and count individual particles or photons, as well as measure the energy absorbed in a material (especially tissue) during exposure to particles and photons. The ability of the radiation to deposit energy per unit mass of the target material is defined as the radiation dose. Total dose refers to the integrated radiation dose that is accrued by satellite electronics over a certain period of time. Although spacecraft components are manufactured to withstand high total doses of radiation (Holmes-Siedle and Adams, 2002), it is important for the spacecraft operator to know the size of the dose which each spacecraft in her/his fleet has endured. This knowledge allows for reasonable replacement strategies in an industry with very long manufacturing lead times. Understanding and mitigating for radiation effects on biological systems is especially of vital importance for human interplanetary travel.

Aside from traditional *in situ* instruments (solar wind analyser, magnetometer and energetic particle detector), remote sensing instruments, including a full disk imager, a Doppler magnetograph and a coronagraph, are also essential for space weather monitoring. In particular, it is important that future Earth-based platforms provide all necessary instruments, not only for terrestrial space weather and missions to the Moon, but also for the development of interplanetary space weather expertise.

Studying the energetic particle environments present in the close martian environment, and their energization in local and external processes, is essential for current and future missions to Mars. Reanalysis of old data, such as observations obtained from the Phobos 2 spacecraft in 1989, can help us understand and predict the energetic particle radiation that can occur in the close martian environment (McKenna-Lawlor *et al.*, 1998; McKenna-Lawlor *et al.*, 2005).

As was mentioned in the Introduction, the atmospheres of planets and their space weather implications are outside the scope of this chapter. However, it should be noted that there is an increasing need for direct measurements of planetary atmospheric electrification – in particular, on Mars – to assess the risk for future unmanned and manned missions. (See Aplin (2005) for a general review about atmospheric electrification in the Solar System, especially in regard to dust storms on Mars (also a future space weather hazard).)

5.10 SUMMARY

This chapter has provided an introduction to the various particle environments that make up our heliosphere: galactic and anomalous cosmic rays, solar energetic particle events, particles accelerated by corotating interaction regions, interplanetary shocks, planetary bow-shocks, and geomagnetically trapped particles. The energy ranges of these particles extend from thermal to GeV, and the characteristics (temporal and spatial) of each population are unique. It is not possible to do justice concerning all the information that is available about these particle populations, and it is to hoped that this chapter will act as inspiration for the reader to consult the literature for more detailed information (see also Chapter 11).

In the final section it was shown that for future interplanetary missions, the understanding of these particle environments, both from an engineering point of view as well as relating to interplanetary space weather forecasting, is essential, and there is still work to be done. This is especially important in regard to the sporadic SEPs that occur and to the constant GCR background. Our ability to protect the crew of future interplanetary missions from these particle environments depends on knowledge gained from past, current and future non-manned missions. With the associated development of real-time models and forecasts, it will be possible to decide on the choice of multi-spacecraft missions and minimum instrument packages necessary to assure continuous monitoring of the Sun and interplanetary space.

Humans are born explorers, and without doubt the next generations will explore our interplanetary neighbourhood. The future of interplanetary space travel looks bright!

5.11 ACKNOWLEDGEMENTS

The author wishes to thank the editors of this book for inviting her to write this very inspiring chapter. She also acknowledges all the interesting papers and books, listed in the References, that were consulted during its preparation.

5.12 REFERENCES

Abbasi R.U., Abu-Zayyad, T., Archbold G., Atkins R., Bellido J., Belov K., Belz J.W., BenZvi S., Bergman D.E., Boyer J., *et al.*, A study of the composition of ultra-high-energy cosmic rays using the High-Resolution Flye's Eye, *Ap. J.*, **622**, 910–926, 2005.

Acuña M.H., Connerney J.E.P., Wasilewski P., Lin R.P., Mitchell D., Anderson K.A., Carlson C.W., McFadden J., Rème H., Mazelle C., *et al.*, Magnetic field of Mars: Summary of results from the aerobraking and mapping orbits, *J. of Geophys. Res.*, **106**, Issue E10, 23403–23418, 2001.

Aharonian F.A., Akhperjanian A.G., Aye K.-M., Bazer-Bachi A.R., Beilicke M., Benbow W., Berge D., Berghaus P., Bernlöhr K., Bolz O., *et al.*, High-energy particle acceleration in the shell of a supernova remnant, *Nature*, **432**, 75–77, 2004.

Albert J.M., Ginet G.P., and Gussenhoven M.S., CRRES observations of radiation belt proton 1; data overview and steady state radial diffusion, *J. Geophys. Res.*, **103**, 9261–9274, 1998.

Aplin K.L., Atmospheric electrification in the Solar System, *Surveys in Geophysics*, **27**, Number 1, 63–108, 2006.

Aschwanden M.J., *Physics of the Solar Corona*, Springer–Praxis, Chichester, UK, 2004.

Atwell W., Zeitlin C., Cucinotta F.A., Cleghorn T., Saganti P., Lee K., Andersen V., Pinsky L., Turner R., Solar Proton Event Observations at Mars with MARIE, (SAE 2003-01-2329), *33rd International Conference on Environmental Systems (ICES), Vancouver, British Columbia, Canada, 7–10 July 2003*. Warrendale, Pennsylvania, USA: Society of Automotive Engineers, 2003.

Auger P. and Grivet T., Analysis of deep rays, *Reviews of Modern Physics*, **11**, Issue 3–4, 232–234, 1939.

Auger P., Ehrenfest P., Maze R., Daudin J. and Fréon R.A., Extensive cosmic-ray showers, *Reviews of Modern Physics*, **11**, Issue 3–4, 288–291, 1939.

Bagenal F., Magnetospheres of the giant planets, *Annu. Rev. Earth Planet. Sci.*, **20**, 289–328, 1992.

Bagenal F., Dowling T.E. and McKinnon W.B. (eds), *Jupiter. The Planet, Satellites and Magnetosphere* (Cambridge planetary science), Cambridge University Press, 2004.

Barnes C.W. and Simpson J.A., Evidence for interplanetary acceleration of nucleons in corotating interaction regions, *Ap. J.*, **210**, L91–L96, 1976.

Blake J.B., Kolasinski W.A., Fillius R.W. and Mullen E.G., Injection of electrons and protons with energies of tens of MeV into L less than 3 on 24 March 1991, *Geophysical Research Letters*, **19**, Number 8, 821–824, 1992.

Boratav M., Probing theories with cosmic rays, *Europhysics News*, **33**, Number 5, 2002.

Bothmer V., The solar atmosphere and space weather, In:Blondel P. and Mason J. (eds), *Solar System Update*, Springer–Praxis, Chichester, UK, pp. 1–53, 2006.

Burke L.F. and Franklin K.L., Observations of a variable radio source associated with the planet Jupiter, *J. Geophys. Res.*, **60**, 213–17, 1955.

Burlaga L.F., Ness N.F., Acua M.H., Lepping R.P., Connerney J.E.P., Stone E.C. and McDonald F.B., Crossing the Termination Shock into the Heliosheath: Magnetic Fields, *Science*, **309**, Issue 5743, 2027–2029, 2005.

Cane H.V., McGuire R.E. and von Rosenvinge T.T., Two classes of solar energetic particle events associated with impulsive and long-duration soft X-ray flares, *Ap. J.*, **301**, 448–459, 1986.

Cane H.V., Reames D.V. and von Rosenvinge T.T., The role of interplanetary shocks in the longitude distribution of energetic particles, *J. Geophys. Rev.*, **93**, 9555–9567, 1988.

Cassé M. and Paul J.A., Local gamma rays and cosmic-ray acceleration by supersonic stellar winds, *Astrophys. J.*, **237**, 236–243, 1980.

Cassé M. and Paul J.A., On the stellar origin of the Ne Eexcess in cosmic rays, *Astrophys. J.*, **258**, 860–863, 1982.

Castelvecchi D., Let it rain, *Symmetry—dimensions of particle physics*, **2**, Issue 1, 13–17, 2005.

Chenette D.L., Conlon K.R., Pyle K.R. and Simpson J.A., Observations of jovian electrons at 1 AU throughout the 13 month jovian synodic year, *Ap. J.*, **215**, L95–L99, 1977.

Christian E.R., Cummings A.C. and Stone A.C., Evidence for anomalous cosmic-ray hydrogen, *Ap. J.*, **334**, L77–L80, 1988.

Clay R. and Dawson B., *Cosmic Bullets*, Allen & Unwin Ptv Ltd, Perseus Publishing, a member of the Perseus Books Group, Australia, 1999.

Conlon T.F., The interplanetary modulation and transport of jovian electrons, *J. Geophys. Res.*, **83**, 541–552, 1978.

Connerney J.E.P., Acuna M.H., Wasilewski P.J., Kletetschka G., Ness N.F., Reme H., Lin R.P. and Mitchell D.L., The global magnetic field of Mars and implications for crustal evolution, *Geophys. Res. Lett.*, **28**, Number 21, 4015–4018, 2001.

Cornwall J.M., Magnetospheric ion acceleration processes, In: Chang T. (ed.), *Ion Acceleration in the Magnetosphere and Ionosphere* (Geophys. Monongr. 38), American Geophys. Union, Washington D.C., 1986.

Cougnet C., Crosby N.B., Eckersley S., Foullon C., Guarnieri V., Guatelli S., Heynderickx D., Holmes-Siedle A., Lobascio C., Masiello S., *et al.*, Radiation Exposure and Mission Strategies for Interplanetary Manned Missions (REMSIM), *Earth, Moon, and Planets*, 2005 ISSN:0167-9295 (Paper) 1573-0794 (Online), DOI:10.1007/s11038-005-9014-1.

Cronin J.W., Gaisser T.K. and Swordy S.P., Cosmic rays at the energy frontier, *Scientific American*, **276**, 32–37, 1997.

Cronin J.W., Cosmic rays: The most energetic particles in the Universe, *Reviews of Modern Physics*, **71**, Number 2, S165–S172, 1999.

Crosby N., Meredith N., Coates A. and Iles R., Modelling the outer radiation belt as a complex system in a self-organized state, *Nonlinear Processes in Geophysics*, **12**, Number 6, 993–1001, 2005.

Cummings A.C., Stone E.C. and Steenberg C.D., Composition of anomalous cosmic rays and other heliospheric ions, *Ap. J.*, **578**, 194–210, 2002.

Diehl R., Kallenbach R., Parizot E. and Von Steiger R., The astrophysics of Galactic Cosmic Rays, *Space Science Reviews*, **99**, 3–11, 2001.

Dorman L.I., Cosmic rays in the Earth's atmosphere and underground, Astrophysics and Space Science Library Series, Volume 303, 2004.

Erlykin A.D. and Wolfendale A.W., The origin of cosmic rays, *Europhysics News*, **32**, Number 6, 246, 2001.

Federmann G., Viktor Hess und die Entdeckung der Kosmischen Strahlung, diploma thesis (in German) , Institut für Radiumforschung und Kernphysik, Wien (in German), 2003. *http://www.mpi-hd.mpg.de/hfm/HESS/public/hessbio.html*

Feynman J., Spitale G. Wang J. and Gabriel S., Interplanetary Proton Fluence Model: JPL 1991, *J. Geophys. Res.*, **98**, 13281–13294, 1993.

Fisk L.A., Kozlovsky B. and Ramaty R., An interpretation of the observed oxygen and nitrogen enhancements in low-energy cosmic rays, *Ap. J.*, **190**, L35–L37, 1974.

Fisk L.A. and Lee M.A., Shock acceleration of energetic particles in corotating interaction regions in the solar wind, *Ap. J.*, **237**, 620–626, 1980.

Foullon, C., Holmes-Siedle A., Crosby N.B. and Heynderickx D., Radiation Hazard and Space Weather Warning System, work-package in ESA study on *Radiation Exposure and Mission Strategies for Interplanetary Manned Missions (REMSIM)*, ESTEC, Noordwijk, Netherlands, 2004, On-line (under REMSIM) at: *http://space-env.esa.int/R_and_D/PN-Radiation.html*

Foullon C., Crosby N.B. and Heynderickx D., Towards interplanetary space weather: Strategies for manned missions to Mars, *Space Weather Journal*, **3**, Number 7, S07004, DOI:10.1029/2004SW000134, 2005.

Friedlander M., *A Thin Cosmic Rain: Particles from Outer Space*, Harvard University Press, USA, 2000.

Gaisser T.K., Origin of cosmic radiation, in: Mewaldt R.A, Jokipii J.R., Lee M.A., Moebius E. and Zurbuchen T.H. (eds), *Acceleration and Transport of Energetic Particles in the Heliosphere*, AIP Conference Proceedings 528, pp. 293–332, 2000.

Garci-Munoz M., Mason G.M. and Simpson J.A., A new test for solar modulation theory: The 1972 May–June low-energy galactic cosmic-ray proton and helium spectra, *Astrophys. J.*, **182**, L81–L84, 1973.

Gosling J., Asbridge J.R., Bame S.J., Feldman W.C., Zwickl R.D., Paschmann G., Sckopke N. and Hynds R.J., Interplanetary ions during an energetic storm particle event: The distribution function from solar wind thermal energies to 1.6 MeV, *J. Geophys. Res.*, **86**, Number A2, 547–554, 1981.

Gosling J.T., The solar flare myth, *J. Geophys. Res.*, **98**, 18937–18950, 1993.

Greisen K., End to the cosmic-ray spectrum?, *Phys. Rev. Lett.*, **16**, Issue 17, 748–750, 1966.

Grieder P.K.F., *Cosmic Rays at Earth*, Elsevier Science, Amsterdam, 2001.

Hanslmeier A., Space weather: Effects of radiation on manned space missions, *Hvar Obs. Bull.*, **27**, 159–170, 2003.

Hess V.F., Über die absorption der γ-strahlung in der atmosphre, *Physikalische Zeitschrift*, Number 22|23, 12. Jahrgang, 998–1001, 1911.

Hess V.F., Über beobachtungen der durchdringenden strahlung bei sieben freiballonfahrten, *Physikalische Zeitschrift*, Number 21/22, 13. Jahrgang, 1084–1091, 1912.

Holmes-Siedle A. and Adams L., *Handbook of Radiation effects*, Oxford University Press, New York, 2002.

Horne R.B., Thorne R.M., Shprits Y.Y., Meredith N.P., Glauert S.A., Smith A.J., Kanekal S.G., Baker D.N., Engebretson M.J., Posch J.L., Spasojevic M., Inan U.S., Pickett J.S. and Decreau P.M.E., Wave acceleration of electrons in the Van Allen radiation belts, Nature, Vol. 437, Issue 7056, pp. 227-230, 2005.

Hudson M.K., Elkington S.R., Lyon J.G., Marchenko V.A., Roth I., Temerin M., Blake J.B., Gussenhoven M.S. and Wygant J.R., Simulations of radiation belt formation during storm sudden commencements, *J. Geophys. Res.*, **102**, Issue A7, 14087–14102, 1997.

Hudson M.K., Marchenko V.A., Roth I., Temerin M., Blake J.B. and Gussenhoven M.S., Radiation belt formation during storm sudden commencements and loss during main phase, *Advances in Space Research*, **21**, Issue 4, 597–607, 1998.

Kahler S.W., Injection profiles of solar energetic particles as functions of coronal mass ejection heights, *Astrophys. J.*, **428**, 837–842, 1994.

Kallenrode M.B., A statistical survey of 5 MeV proton events at transient interplanetary shocks, *J. Geophys. Res.*, **101**, 24393–24409, 1996.

King, J. H., Solar proton fluences for 1977–1983 space missions, *J. Spacecraft Rockets*, **11**, Number 401, 1974.

Kivelson, M.G. and Russell C.T., *Introduction to Space Physics*, Cambridge University Press, Cambridge, New York, Melbourne, 1995.

Klecker B., McNab M.C., Blake J.B., Hamilton D.C., Hovestadt D., Kaestle H., Looper M.D., Mason G.M., Mazur J.E. and Scholer M., Charge state of anomalous cosmic-ray nitrogen, oxygen, and neon: SAMPEX observations, *Astrophys. J.*, **442**, L69–L72, 1995.

Kumar, M., The mitigation of radiation hazards on missions to deep space, *Space Weather*, **3**, Number 12, 2005.

Li X. and Temerin M.A., The Electron Radiation Belt, *Space Science Reviews*, **95**, 569–580, 2001.

Looper M.D., Blake J.B., Mewaldt R.A., Cummings J.R. and Baker D.N, Observations of the remnants of the ultrarelativistic electrons injected by the strong SSC of 24 March 1991, *GRL*, **21**, Number 19, 2079–2082, 1994.

McCracken K.G., Dreschhoff G.A.M., Zeller E.J., Smart D.F. and Shea M.A., Solar cosmic ray events for the period 1561–1994, Part 1, Identification in polar ice, 1561–1950, *J. of Geophys. Res.*, **106**, 21585–21598, 2001a.

McCracken K.G., Dreschhoff G.A.M., Smart D.F. and Shea M.A., Solar cosmic ray events for the period 1561–1994, Part 2, The Gleissberg periodicity, *J. of Geophys. Res.*, **106**, 21599–21609, 2001b.

McKenna-Lawlor S.M.P., Afonin V.V., Kirsch E., Schwingenschuh K., Slavin J.A. and Trotignon J.G., An overview of energetic particles (from 55 keV to >30 MeV) recorded in the close Martian environment, and their energization in local and external processes, *Planet Space Sci.*, **46**, Number 1, 83–102, 1998.

McKenna-Lawlor S.M.P., Dryer M., Fry C.D., Sun W., Lario D., Deehr C.S., Sanahuja B., Afonin V.A., Verigin M.I. a,d Kotova G.A., Predictions of energetic particle radiation in the close Martian environment, *J. of Geophys. Res.*, **110**, Issue A3, 2005.

Mewaldt R.A. and Mason G.M., Cosmic rays in the galaxy and the heliosphere, in: Geiss J. and Hultqvist B. (eds), *The Solar System and Beyond: Ten Years of ISSI*, ISSI Scientific report, SR-003, ISSI/ESA, pp. 69–81, 2005.

Molina-Cuberos G.J., Stumptner W., Lammer H., Kömle N.I. and O'Brien K.:, Cosmic ray and UV radiation models on the ancient Martian surface, *Icarus*, **154**, 216–222, 2001.

Orlowski D.S., Russell C.T. and Lepping R.P., Wave phenomena in the upstream region of Saturn, *J. Geophys. Res.*, **97**, 19187, 1992.

Parks G.E., *Physics of Space Plasmas*, Westview Press, a member of the Perseus Books Group, Boulder, Oxford, 2004.

Pesses M.E., Jokipii J.R. and Eichler D., Cosmic ray drift, shock wave acceleration, and the anomalous component of cosmic rays, *Astrophysical Journal* (Part 2 - Letters to the Editor), **246**, L85–L88, 1981.

Pfotzer G., Dreifachkoinzidenzen der Ultrastrahlung aus vertikaler Richtung in der Stratosphäre II. Analyse der gemessenen Kurve, *Z. Phys.*, **102**, Number 23, 1936.

Reames D.V., Acceleration of energetic particles by shock waves from large solar flares, *Astrophys. J. Lett.*, **358**, L63, 1990.

Reames D.V., Coronal abundances determined from energetic particles, *Advances in Space Research*, **15**, Number 7, (7)41–(7)51, 1995.

Reames D.V., Barbier L.M., von Rosenvinge T.T., Mason G.M., Mazur J.E. and Dwyer J.R., Energy spectra of ions accelerated in impulsive and gradual solar events, *Ap. J.*, **483**, 515, 1997.

Reames D.V., Kahler S.W. and Ng C.K, Spatial and temporal invariance in the spectra of energetic particles in gradual solar events, *Astrophysical Journal*, **491**, 414, 1997.

Reames D.V, Particle acceleration at the Sun and on the heliosphere, *Space Sci. Rev.*, **90**, 413–491, 1999.

Richardson I.G, Energetic particles and corotating interaction regions in the solar wind, *Space Science Reviews*, **111**, Issue 3, 267–376, 2004.

Russell C. T., Elphic R.C. and Slavin J.A., Limits on the possible intrinsic magnetic field of Venus, *J. Geophys. Res.*, **85**, 8319–8332, 1980.

Russell C.T and Hoppe M.M., Upstream waves and particles, *Space Sci. Rev.*, **34**, Number 155, 1983.

Russell C.T., Planetary bow shocks. In: Tsurutani and R.G. Stone (eds), *Collisionless Shocks in the Heliosphere: Reviews of Current Research* (Geophysical Monograph 35), AGU, pp. 109–130, 1985.

Russell C.T., Lepping R.P. and Smith C.W., Upstream waves at Uranus, *J. Geophys. Res.*, **95**, Number 2273, 1990.

Russell C.T. and Luhmann J.G., Mercury: Magnetic Field and Magnetosphere. In: J.H. Shirley and R.W. Fainbridge (eds), *Encyclopedia of Planetary Sciences*, Chapman and Hall, New York, pp. 476–478, 1997.

Russell C.T., Outer planet magnetospheres: A tutorial, *Adv. Space Res.*, **33**, Issue 11, 2004–2020, 2004.

Russell C.T., Interaction of the Galilean Moons with their plasma environments, *Planetary and Space Science*, **53**, Issue 5, 473–485, 2005.

Sawyer D.M. and Vette J.I., AP-8 trapped proton environment for solar maximum and solar minimum, National Space Science Data Center, Report 76-06, Greenbelt, Maryland, 1976.

Semeniuk I., Astronomy's phantom foul balls: Ultrahigh-energy cosmic rays, *Sky & Telescope*, March 2003.

Smart D.F. and Shea M.A., Comment on estimating the solar proton environment that may affect Mars missions, *Adv. Space Res.*, **31**, Number 1, 45–50, 2003.

Sokolosky P., *Introduction to Ultrahigh Energy Cosmic Ray Physics*, Westview Press, a member of the Perseus Books Group, Boulder, Oxford, 2004.

Stecker F.W., Krizmanic J.F., Barbier L.M., Loh E., Mitchell J.W., Sokolsky P. and Streitmatter R.E., Observing the ultrahigh energy universe with OWL Eyes, *Nuclear Physics B: Proceedings Supplements*, **136**, 433–438, 2004.

Stone E.C., Cummings A.C., McDonald F.B., Heikkila B.C., Lal N., Webber W.R., Voyager 1 explores the termination shock region and the heliosheath beyond, *Science*, **309**, Issue 5743, 2017–2020, 2005.

Svensmark H., Influence of cosmic rays on Earth's climate, *Physical Review Letters*, **81**, Issue 22, 5027–5030, 1998.

Swordy S.P, The energy spectra and anisotropies of cosmic rays, *Space Science Reviews*, **99**, Issue 1/4, 85–94, 2001.

Takashima T., Doke T., Hayashi T., Kikuchi J., Kobayashi M., Shirai H., Takehana N., Ehara M., Yamada Y., Yanagita S., *et al.*, The first observation of sulfur in anomalous cosmic rays by the Geotail and the Wind spacecrafts, *Ap. J.*, **477**, L111–L113, 1997.

Tsurutani B.T. and Lin R.P, Acceleration of >47 keV ions and >2 keV electrons by interplanetary shocks at 1 AU, *J. Geophys. Res.*, **90**, Number A1, 1–11, 1985.

Turner, R. (ed.), Foundations of Solar Particle Event Risk Management Strategies, in: *Findings of the Risk Management Workshop for Solar Particle Events*, publication NASA Grant NAGW-4166, ANSER, Arlington, VA, USA. Figure available at *http://www.windows.ucar.edu/spaceweather/flare_dose.html*

Turner R., Solar particle events from a risk management perspective, *IEEE Transactions on Plasma Science*, **28**, Number 6, 2103–2113, 2000.

Vampola A.L. and Korth A., Electron drift echoes in the inner magnetosphere, *GRL*, **19**, 625–628, 1992.

Van Allen J.A., Ludwig G.H., Ray E.C. and McIlwain C.E., Observations of high intensity radiation by satellites 1958 Alpha and Gamma, *Jet Propul.*, **28**, 588–592, 1958.

Van Allen J.A. and Frank L.A., Radiation around the Earth to a radial distance of 107,400 km, *Nature*, **183**, Number 430, 1959.

Vette J.I., The AE-8 Trapped Electron Model Environment, *NSSDC/WDC-A-R&S 91-24*, 1991a.

Vette J.I., The NASA/National Space Science Data Center Trapped Radiation Environment Model Program (1964–1991), *NSSDC/WDC-A-R&S 91-29*, 1991b.

Walt M., Introduction to Geomagnetically Trapped Radiation, Cambridge University Press, Cambridge, New York, Melbourne, 1994.

Watson A.A., The highest energy particles in nature, *Europhysics News*, **32**, Number 6, 2001.

Wenzel K.-P., Reinhard R. and Sandersen T.R., Characteristics of energetic particle events associated with interplanetary shocks, *J. Geophys. Res.*, **90**, Number A1, 12–18, 1985.

Wild J.P., Smerd S.F. and Weiss A.A., Solar bursts, *Annual Review of Astronomy and Astrophysics*, **1**, 291, 1963.

Wilson J.W., Shinn J.L., Simonsen L.C., Cucinotta F.A., Dubey R.R., Jordan W.R., Jone T.D., Chang C.K and Kim M.Y., *Exposures to Solar particle Events in Deep Space Missions* (NASA Technical Paper 3668), 1997. *http://techreports.larc.nasa.gov/ltrs/PDF/1997/tp/NASA-97-tp3668.pdf*

Wilson J.W, Cucinotta F.A., Shinn J.L, Simonsen L.C., Dubey R.R, Jordan W.R., Jones T.D., Chang C.K. and Kim M.Y, Shielding from solar particle event exposure in deep space, *Radiation Measurements*, **30**, 361–382, 1999.

Xapsos, M.A., Summers G.P., Barth J.L., Stassinopoulos E.G. and Burke E.A, Probabillity model for worst case solar proton event fluences, *IEEE Trans. Nucl. Sci.*, **46**, 1481–1485, 1999.

Xapsos, M.A., Summers G.P., Barth J.L., Stassinopoulos E.G. and Burke E.A, Probabillity model for cumulative solar proton event fluences, *IEEE Trans. Nucl. Sci.*, **47**, 486–490, 2000.

Zatsepin G.T. and Kuz'min V.A., *Pis'ma Zh.Eksp. Teor. Fiz.*, **4**, 114 (166) [*JETP Lett.*, **4**, 78], 1966.

Zöllich F., Wibberenz G., Kunov H. and Green G., Corotating events in the energy range 4–13 MeV as observed on board Helios 1 and 2 in 1975 and 1976, *Adv. Space Res.*, **1**, 89–92, 1981.

6

Radiation belts and ring current

Daniel N. Baker and Ioannis A. Daglis

In terms of mass, energy and associated dynamic phenomena the inner magnetosphere is the locus of two highly important magnetospheric particle populations: the radiation belts and the ring current. The radiation belts and the ring current are part of the chain that interconnects the Sun and interplanetary space with the terrestrial magnetosphere, ionosphere, and atmosphere – and often even the surface of the Earth. This chapter discusses the origin, formation and dynamics of these populations. We further describe some critical aspects of space weather that involve these populations, such as spacecraft surface and bulk charging, single event effects, high radiation doses, and effects of geomagnetically induced currents on power grids and other land-based infrastructures.

6.1 INTRODUCTION AND HISTORICAL CONTEXT

The trapped radiation belts were the first component of magnetospheric plasma to be discovered, at the dawn of the space era. The discovery was made by James Van Allen's group, using Geiger–Mueller tubes placed onboard the Explorer I spacecraft (Van Allen *et al*., 1958). Those measurements were interpreted as the result of intense corpuscular radiation (Van Allen, 1959).

The foundations of modern magnetospheric research were laid earlier by Chapman and Ferraro (1930, 1931), who proposed that a transient stream of outflowing solar ions and electrons were responsible for terrestrial magnetic storms. Chapman and Ferraro claimed that once the solar stream had reached the Earth, charged particles would leak into the magnetosphere and drift around the Earth, creating a current whose field would oppose the main geomagnetic field. This is remarkably close to what we know today. The only major element of Chapman's theory that has changed is the existence of a continuous – instead of transient – stream of ionized gas from the Sun. This stream was named the 'solar wind' by

Eugene Parker (Parker, 1958), and its existence was later confirmed by measurements performed by the Venus-bound Mariner 2 spacecraft (Neugebauer and Snyder, 1962).

The central idea of the Chapman and Ferraro theory was a huge *ring current* in space, circling the Earth and leading to magnetic perturbations on the Earth's surface. This idea was further elaborated by Singer (1956, 1957), and was eventually confirmed by *in situ* spacecraft measurements, starting with Explorer I.

The actual ability of the geomagnetic field to trap high-energy electrons was experimentally verified by the Argus experiment, which was proposed by Nicholas C. Christofilos in 1957 and carried out in 1959 (Christofilos, 1959). Christofilos – an unconventional Greek scientist who had been working as an engineer designing elevator systems in Athens before migrating to the US in 1953 – had actually communicated to the US Army in the early 1950s, that many charged particles, due to the dipole magnetic field, could be trapped around the Earth. He further proposed that an artificial radiation belt, due to beta decay, could be created by exploding one or more small nuclear fission bombs at high altitude ($\sim$200 km). This proposal evolved into Argus – the first active experiment in space, which was successfully performed in 1959.

The Earth's magnetosphere is now known to be an efficient accelerator and trapping device for energetic particles. The sources, losses, acceleration mechanisms and transport processes of energetic particles remain primary issues in magnetospheric physics. High-energy electrons and ions hold special interest because of their ubiquitous presence in the Earth's magnetosphere and their importance to other scientific and technological issues. Modern instrumentation has given an unprecedented combination of sensitivity, energy resolution, time resolution, and measurement duration that has made it possible to observe previously unknown energetic particle phenomena. Long-term measurements have revealed many key features of such particles, and have also presented a great variety of new challenges in understanding the dynamics of energetic particles in the Earth's magnetosphere.

Important space weather effects relate to large and long-lasting enhancements of radiation belt and ring current particle fluxes. Such particles can be damaging to near-Earth spacecraft (Vampola, 1987; Wrenn, 1995; Baker, 2000), as well as to humans in low-Earth orbit (Weyland and Golightly, 2001). Physical phenomena include dose and bulk charging effects in space vehicles in most near-Earth orbits (Violet and Frederickson, 1993; Koons and Gorney, 1992). Recent research has provided a reasonably clear picture of the solar and solar wind drivers of general radiation belt changes (Baker *et al.*, 2001; Li *et al.*, 2001a) and ring current enhancements (Tsurutani, 2001; Daglis *et al.*, 2003). Analyses of long-term data sets allow us to characterize the variation of the Earth's outer radiation belts and ring current, and to describe the most extreme conditions over approximately the last sunspot activity cycle.

6.2 RADIATION BELT SOURCES

Long-term studies of energetic electron fluxes in the Earth's magnetosphere have revealed many of the temporal occurrence characteristics and their relationships to solar wind drivers. Figure 6.1 is a schematic diagram of the radiation belt structure. Early work showed the obvious and powerful role played by solar wind speed in producing subsequent highly relativistic electron enhancements. More recent work has also pointed out the key role that the north–south component of the IMF plays. In order to observe relativistic electron enhancement, there must typically be an interval of southward IMF along with a period of high ($V_{SW} \geq 500\,\text{km/s}$) solar wind speed. This has led to the view that enhancement in geomagnetic activity (such as magnetospheric substorms) is a key first step in the acceleration of magnetospheric electrons to high energies. A second step is found to be a period of intense wave activity (that is closely related to high values of V_{SW}). Hence, substorms appear to provide a 'seed' population, while high-speed solar wind drives the acceleration to relativistic energies in this two-step geomagnetic storm scenario. This picture seems to apply to most storms examined, whether associated with high-speed streams or with CME-related events.

Recent mechanisms have been studied that might account for acceleration of electrons to relativistic energies during geomagnetic storms. An important correlation has been found between electron flux enhancements and ULF wave power in the magnetosphere (Baker *et al.*, 1998c). The lower panel of Figure 6.2 shows the wave power measured at several different ground stations in the CANOPUS system in Canada for the period 30 April through 15 May 1998. The data show increases from quiet day wave power by as much as a factor of 1,000 in the frequency range 0.8–20 mHz. It has been argued that these ULF waves can play a key role in electron

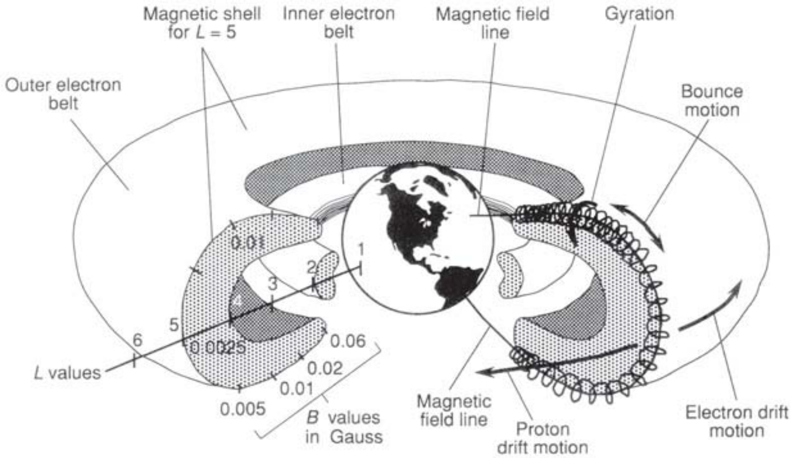


Figure 6.1. A three-dimensional representation of the inner and outer radiation belts around the Earth (Mitchell, 1994).

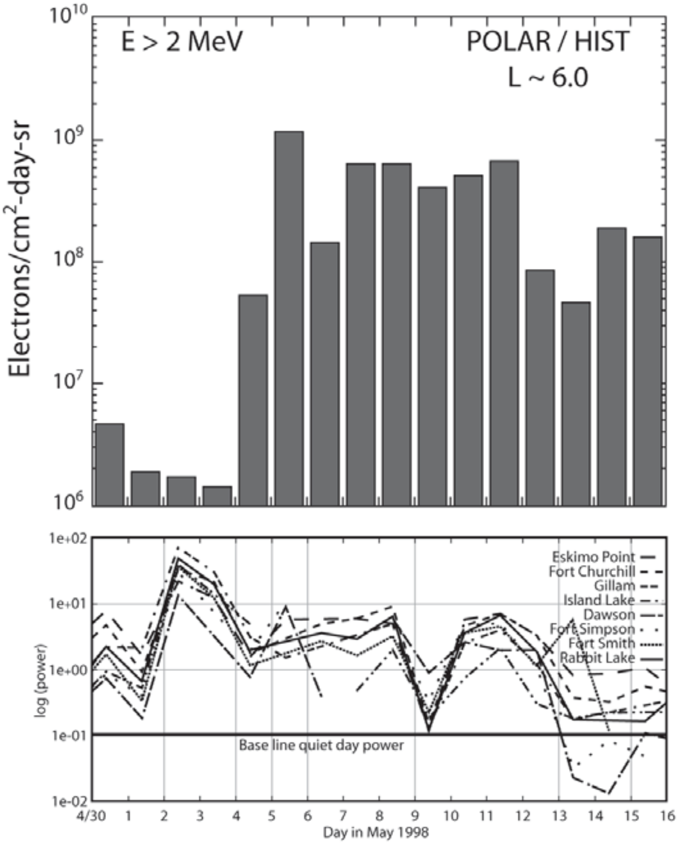


Figure 6.2. The ULF wave power (lower panel) measured each day during late April and early May 1998. Several CANOPUS ground stations are represented (as labelled). The wave frequency range is 0.8–20 mHz. (Courtesy of G. Rostoker). The upper panel shows electron flux increases associated with this wave power increase as measured at $L \sim 6.0$ by the POLAR spacecraft (from Baker *et al.*, 2005).

acceleration (Rostoker *et al.*, 1998; Hudson *et al.*, 2000). The upper panel shows concurrent measurements of electron flux obtained by the POLAR spacecraft at $L \sim 6.0$. Obviously, there was a large increase in electron flux in the radiation belts associated with the ULF power increase.

Radiation belt electrons are formed by accelerating lower energy ambient electrons. There are really two possible sources of lower energy electrons. One source is electrons at larger L that can be energized by being transported radially inward. This is usually called radial diffusion. Another source is lower-energy electrons at the same spatial location that can be energized by wave–particle interactions. Both possible sources usually have a substantially larger phase space density than the radiation belt electrons, and thus either of them could be a source of radiation belt electrons. Radial diffusion is usually thought to be the main acceleration mechanism (Schulz and Lanzerotti, 1974). Recently, a greater emphasis has been placed on *in situ* heating of electrons by VLF waves on the same *L*-shell (Temerin, 1994; Summers, 1998; Horne and Thorne, 1998; Meredith *et al.*, 2001; Meredith *et al.*, 2002; Albert, 2002). However, the relative effectiveness of these acceleration mechanisms has not yet been fully quantified.

By analysing plasma wave and particle data from the CRRES satellite during three case studies, Meredith *et al.* (2002) suggested that the gradual acceleration of electrons to relativistic energies during geomagnetic storms can be effective only when there are periods of prolonged substorm activity following the main phase of the geomagnetic storm. They argue that the prolonged substorm activity provides sustained VLF wave activity, which in turn accelerates some substorm injected electrons to higher energies. Thus, there is an emerging consensus that magnetospheric substorms are key to providing the seed population (Baker *et al.*, 1998b).

Figure 6.3 shows a schematic diagram which illustrates energy flow from the solar wind into, and through, the magnetosphere–ionosphere system (adapted from Baker *et al.*, 1997). The explosive dissipation of stored magnetotail energy that

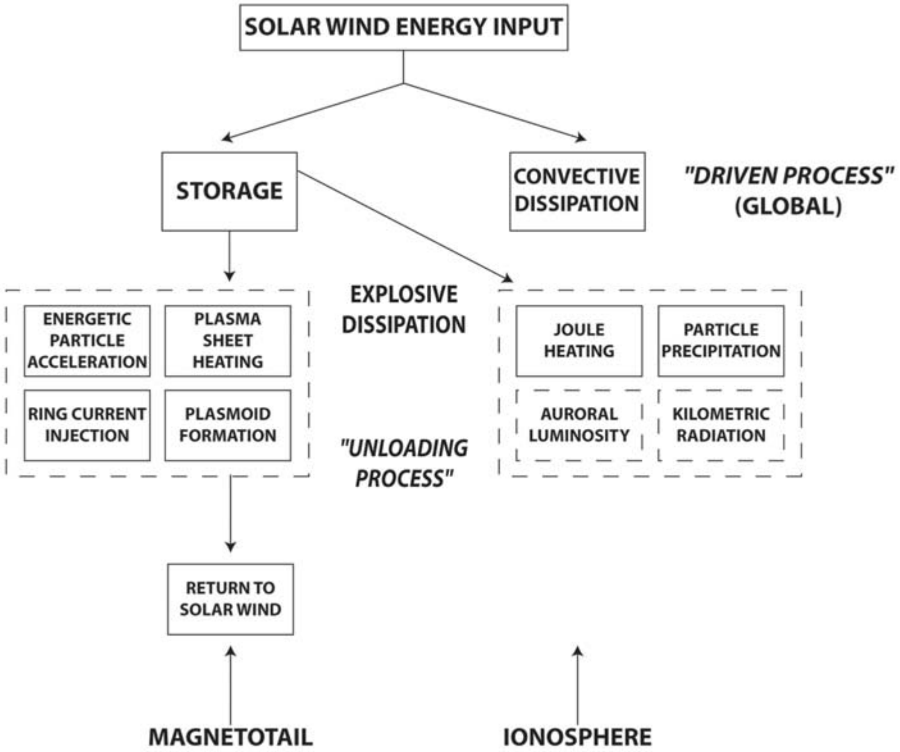


Figure 6.3. The flow of energy into and through the magnetosphere during periods of enhanced geomagnetic activity (adapted from Baker *et al.*, 1997).

occurs during substorms leads to many forms of energy output including plasmoid formation, ring current injection, plasma sheet heating, and particle acceleration. However, the typical substorm seldom directly accelerates electrons to energies much above 200–300 keV (Baker *et al.*, 1997). A further, second-step acceleration process involving radial diffusion and wave particle interaction is almost certainly necessary to reach the relativistic energies characteristic of the outer Van Allen belt.

The energy that can be gained by radial transport (whether in the form of radial diffusion or fast injections) through the violation of the third adiabatic invariant is limited by the ratio of the magnetic field magnitudes within the region of radial transport. Thus, as noted above, radial transport as an energization mechanism normally requires a substantial source population. For a given value of the first and second adiabatic invariants, the phase space density usually increases with increasing L for $L = 3$–6.6 and beyond (Selesnick and Blake, 1997). Thus there should be a region outside of geosynchronous orbit where the phase space density at constant first and second adiabatic invariants peaks. The central plasma sheet region is a probable source region. Indeed, study of several Wind spacecraft perigee passes in conjunction with POLAR spacecraft measurements suggests that the phase space density for given first adiabatic invariant continues to increase toward larger radial distances ($\sim$11–14 Re) and precipitously decreases once the Wind satellite went out of the magnetosphere (Li *et al.*, 1997a, b).

Based on the standard radial diffusion equation, a model (Schulz and Lanzerotti, 1974; Li *et al.*, 2001a) was developed to make quantitative prediction of the intensity of multi-million electron Volt (MeV) electrons at geosynchronous orbit using only measured solar wind parameters. The radial diffusion equation was solved by setting the phase space density larger at the outer boundary than at the inner boundary and by making the diffusion coefficient a function of the solar wind parameters. The most important parameters were the solar wind velocity and the southward component of the IMF. Figure 6.4 displays a comparison of two different years of daily averages of the MeV electron flux measured at geostationary orbit with a prediction based solely on measurements of the solar wind. Both the shorter time scale and the longer seasonal effects are reproduced. Furthermore, the model provides a physical explanation for several features of the correlation between the solar wind and the MeV electron flux at geostationary orbit such as the approximate 1–2 day delay between the peak in the solar wind velocity and the peak in the MeV electron flux at geostationary orbit. The delay is primarily due to the fact that it takes some time for the electrons to diffuse inward to geostationary orbit in response to changes in the solar wind input, and also some time for such changes to decay (Li *et al.*, 2001a).

Baker and Li (2003) have shown that magnetospheric substorms and geomagnetic storms are closely related to one another when it comes to energetic electron phenomena. It would be remarkable that a southward turning of the IMF that opens the magnetosphere to energy input would lead to two totally separate and disconnected phenomena. The original view of S. Chapman and many other researchers, that storms are just a superposition of substorms, was clearly too limited. A more supportable view is that substorms are an important, indeed, key step along the way

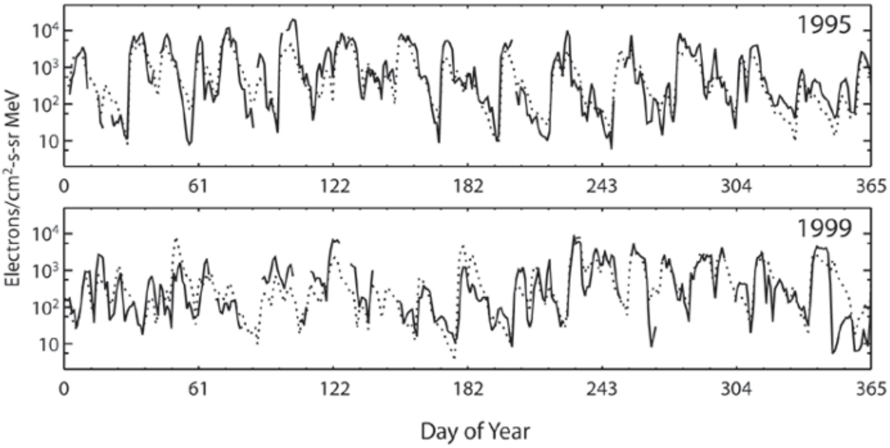


Figure 6.4. A comparison of daily averages of 1.8–3.6 MeV electron flux measured at geosynchronous orbit with the predicted results based solely on measurements of the solar wind, for 1995 and 1999 as examples. The solid line shows electron flux measured at geosynchronous orbit, data gaps are not plotted. The dotted line shows predicted results. The vertical axis shows electron flux (cm^2–s–sr–MeV). The horizontal axis shows the day of the year (courtesy of X. Li).

to geomagnetic storms. The magnetosphere crosses many thresholds in its progression of development, and it begins to allow many new forms of energy dissipation as it is driven harder and harder by the solar wind (Figure 6.5). Substorms are an elementary (and essential) component in this progression. As the magnetosphere progresses toward major storms, however, the external driver (the strong flow of the solar wind energy) overwhelms and drives the magnetosphere into a mode of powerful 'direct' response. This is somewhat analogous to a large-scale and highly organized hurricane which moves across sea and land in a dominant and coherent way. Strong radiation belt enhancements are one part of a similar type of large-scale process in the magnetospheric context.

6.3 RADIATION BELT STRUCTURE AND DYNAMICS

Records of outer zone high-energy electron fluxes (Figure 6.6) for the solar cycle in the 1990s suggests an 11-year cycle in the outer-belt electron flux (see, also Baker *et al.*, 1993a). The long-term geostationary-orbit record indicates that outer radiation belt electrons peak in the declining phase of the sunspot cycle (in association with high-speed solar wind streams and recurrent storms), not at the time of sunspot maximum. The AE-8 model (Vette, 1991) flux level estimates show that the static NASA models give substantially higher flux values at $L = 6.6$ than are actually observed and, furthermore, none of the solar cycle flux variations are really captured by AE-8.

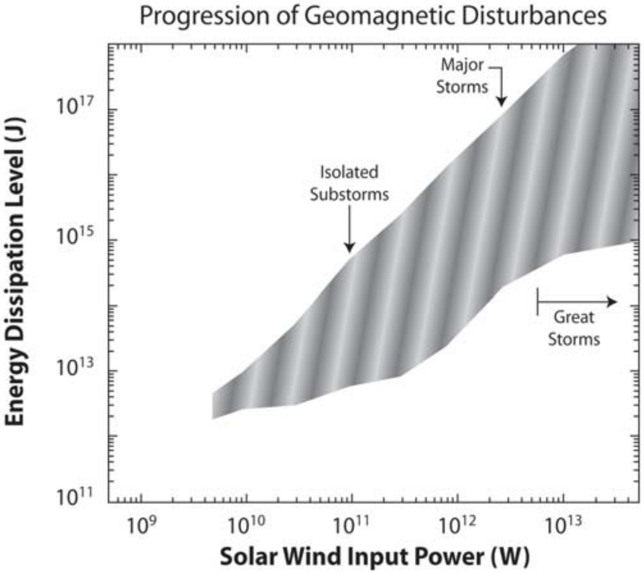


Figure 6.5. The progression of energy dissipation in the magnetosphere with the increasing solar wind energy input rate. The vertical shading shows the range of energy dissipation typically occurring during different input power levels (adapted from Baker *et al.*, 2001).

There is ample evidence of a ‘global coherence’ of relativistic electron behaviour such that electron fluxes vary throughout the entire outer radiation belt with a fair degree of synchronicity (Kanekal *et al.*, 1999; Baker *et al.*, 2001). Geostationary orbit data as shown in Figure 6.6 provide an important and useful monitoring of outer radiation zone particle fluxes. However, it is reasonably evident that the fluxes of relativistic electrons near $L = 4$ are much higher and more slowly varying than those at $L = 6.6$ (shown in Figure 6.7, colour section). Thus the electron fluxes are quantitatively, if not qualitatively, different at the heart of the outer zone ($L = 4$) than at its outer fringes ($L = 6.6$).

Figure 6.7 (see colour section) shows a colour-coded representation of the flux of 2–6 MeV electrons (top panel) measured by the Solar, Anomalous, and Magnetospheric Particle Explorer (SAMPEX) (see Baker *et al.*, 1993b) from launch in mid-1992 to 2004. The lower panel of Figure 6.7 shows equivalent data ($E > 2$ MeV) from the Comprehensive Energetic Particle Pitch Angle Detector (CEPPAD) experiment (Blake *et al.*, 1995) on the POLAR spacecraft from the time of its launch in February 1996 until 2004. The horizontal axis is time in years and the vertical axis in each panel is the McIlwain L-value. The colour-coded intensity of electrons for each panel is shown as integral flux (electrons $(\mathrm{cm}^2\text{–s–sr})^{-1}$) according to the colour bar to the right of the figure. A 27-day running-average smoothing function has been run over the entirety of both the SAMPEX and POLAR data to reduce the high-frequency fluctuations that can be present.

This figure shows many features pertaining to the structure and variability of the

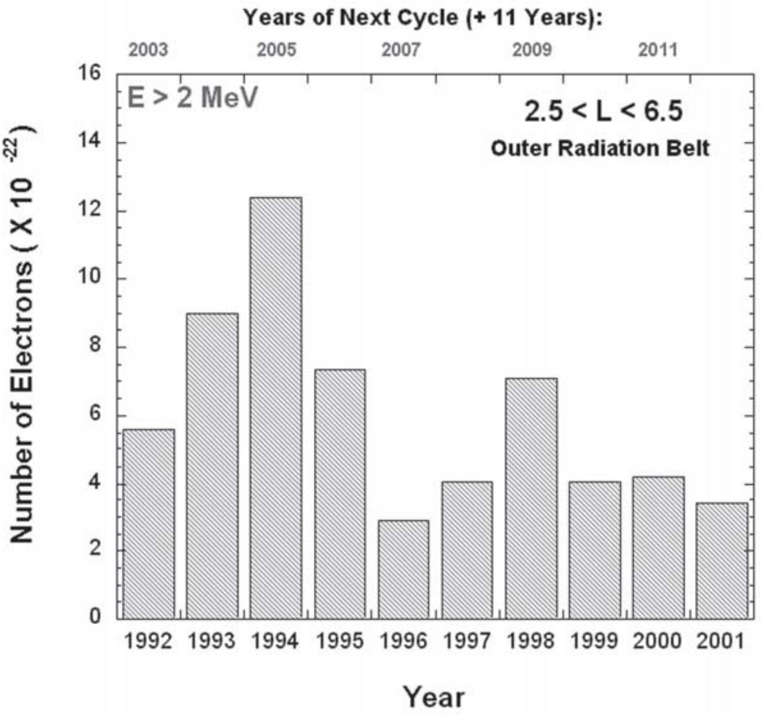


Figure 6.6. Annual fluxes of electrons with $E > 1.4$ MeV from 1992 through 2001 throughout the outer radiation zone. The upper horizontal scale shows the approximate corresponding years for the present solar cycle (adapted from Baker and Li, 2003).

outer zone electron population. The high intensity and great breadth of the outer radiation belt in 1993–95 is quite clear. This contrasts with the weaker and narrower outer belt in 1996–97. The resurgence of the electron fluxes (especially in 1998 near the inner portion of the outer belt) after the time of solar minimum is evident. As discussed by previous authors (Baker *et al.*, 1998a, 2001; Kanekal *et al.*, 1999), SAMPEX and POLAR data show many similar features in space and time. However, the POLAR data, being obtained typically nearer the magnetic equator, show generally higher absolute electron intensities.

While Figure 6.7 provides a broad global view of outer zone electron flux variations, a more quantitative view is afforded by taking 'cuts' of the data at various L-values. Figure 6.8 shows such cuts for $L = 2.0$, 3.0 and 4.0. The plots show integral flux on a daily basis, but again with the 27-day smoothing filter applied. Notice the high peak flux values ($\sim > 2 \times 10^4$ electrons/cm^2–s–sr) in late 1993 and early 1994 at $L = 4.0$. At later times, the $L = 4$ fluxes were much lower on average. Notice also the extended interval in 1996 when the $L = 4$ fluxes were near 10^1 electrons (cm^2–s–sr)$^{-1}$. Clearly, the radiation belt electron flux can, even in the heart of the outer zone, exhibit three to four orders of magnitude differences in monthly average intensities over the course of the solar cycle. Close inspection of

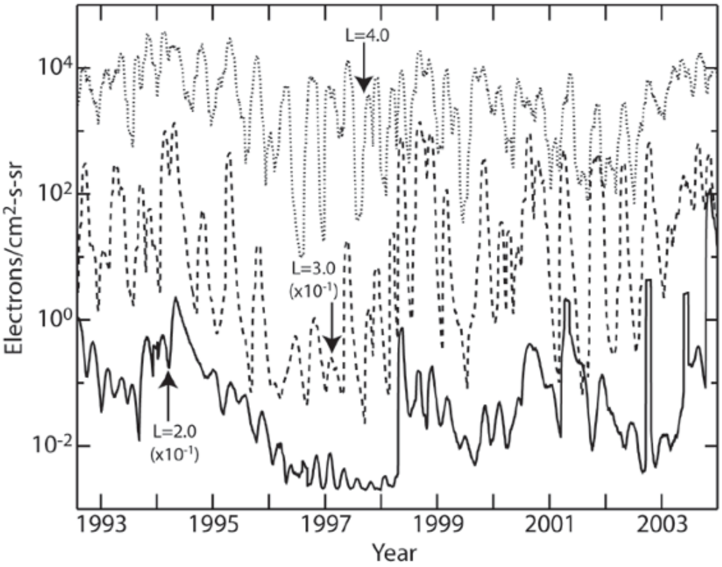


Figure 6.8. Plots of 'cuts' at selected L-values for fluxes of electrons measured by SAMPEX from 1992 to 2004. Cuts for $L = 4.0$, $L = 3.0$, and $L = 2.0$ are shown separately by the different curves (as labelled).

the data for $L = 4.0$ shows an important feature not in models: the seasonal (solstice/equinox) effect reported by Baker *et al.* (1999), and recently discussed further by Li *et al.* (2001b): on average, electron fluxes are much higher around equinox than they are at solstice times.

Figure 6.8 also shows the $L = 3$ electron flux profile near the 'slot' region between the inner and outer radiation belts. Notice in 1996–1997 (around solar minimum conditions) that average flux levels of $E > 2\,\text{MeV}$ electrons almost never went above 10 electrons/cm^2–s–sr. On the other hand, in 1993–1994 and again in 1998–2000, the electron fluxes were quite high, often reaching peak values near 10^4 electrons/cm^2–s–sr. During most periods – during 1992–1995 and 1998–2001 – the observed fluxes were considerably higher than either the AE-8 (Max) or AE-8 (Min) would predict (see Vette, 1991).

The $L = 2.0$ profile in Figure 6.8 shows a behaviour similar to $L = 3.0$ in which fluxes were near background levels during sunspot minimum, but were much more elevated in the years away from minimum. The AE-8 models (Vette, 1991) would suggest that there would be no significant flux at $L = 2.0$ for SAMPEX altitudes. It is clear that this is not the case. From Figure 6.8 we see that spacecraft within the inner magnetosphere ($L \sim< 4$) could experience vastly different fluences on a daily and monthly basis depending on which phase of the solar cycle they were actually operating in. Certainly, the static (AE) models are not a good description of observed flux values. More dynamic models (such as Brautigam *et al*., 1992) clearly are needed to characterize observed variability.

As noted, examination of the $L = 4.0$ record in Figure 6.8 shows that generally the highest electron fluxes were seen in late 1993 and early 1994. More detailed inspection of the unsmoothed SAMPEX data supports this interpretation. Thus the 1993–1994 interval was the most extreme period of $E > 2\,\text{MeV}$ electron radiation in the past solar cycle. We have examined data from SAMPEX electron sensors with thresholds from 0.4 MeV to 6 MeV over the lifetime of the mission in order to judge flux levels. The 1993–1994 interval was also found to be the time of the highest flux values over the entire relativistic energy range.

Figure 6.9 (see colour section) is a representation of the daily fluxes of 2–6 MeV electrons measured by SAMPEX during 1994. As in Figure 6.7 (see colour section), the flux values are colour-coded and plotted for various L-values versus day of year. Also shown are the white vertical arrows, 27-days apart, along the top of the figure. There was a clear and prominent 27-day periodicity in the electron flux enhancements. This was well associated with solar wind velocity enhancements. Thus, during the approach to sunspot minimum, high-energy electrons are at their highest levels throughout the outer radiation belt, and this population is well associated with recurrent geomagnetic storms.

The period of time around solar maximum is characterized not by high-speed solar wind streams, but rather by episodic geomagnetic storms driven by coronal mass ejections (CMEs). Powerful CMEs are often preceded by strong interplanetary shock waves that can greatly compress and distort the magnetosphere. Many CME-driven geomagnetic storms give rise to relativistic electron enhancements in the magnetosphere (Reeves, 1998). Specific CME-related events have been examined in detail in various papers (Baker *et al*., 1998a, 2004).

An interesting case occurred in October–November 2003. The Sun was very active in late 2003, with numerous flares and CMEs (Baker *et al*., 2004). The result was a set of major geomagnetic storms with a minimum Dst value of $\sim -400\,\text{nT}$ and an extended interval of high Kp and AE activity. As can be inferred from careful examination of Figure 6.7 (see colour section), the period of late 2003 was a time of very substantial electron flux enhancement throughout the entire outer radiation belt. Electron intensities were elevated from the ‘slot’ region ($L \sim 2.0$) all the way out to the vicinity of geostationary orbit ($L \sim 6.6$).

It is important to note that the October–November 2003 interval was quite prominent from a space weather standpoint: Several spacecraft anomalies and failures occurred (Lopez *et al*., 2004). Most of these spacecraft problems occurred near geostationary orbit or elsewhere in the outer magnetosphere. Figure 6.10 (see colour section) shows a detail of relativistic electron fluxes (2–6 MeV) measured by SAMPEX at $L = 6.6$ for the period 17 October through 15 December 2003. As can be seen, the electron flux, after being relatively low in early October, went up abruptly to high levels in late October. With some fluctuations in absolute intensity, the fluxes remained elevated until the end of the year. As discussed by Lopez *et al*. (2004), numerous failures may have been related to deep dielectric charging from the elevated electron flux.

6.4 RING CURRENT STRUCTURE, SOURCES AND FORMATION

The terrestrial ring current is carried by energetic charged particles flowing toroidally around the Earth, and creating a ring of westward electric current, centred at the equatorial plane and extending from geocentric distances of about $2\,R_E$ to roughly $9\,R_E$. This current has a permanent existence due to the natural properties of charged particles in the geospace environment, yet its intensity is variable. It becomes more intense during geospace magnetic storms. Changes in this current are responsible for global decreases in the Earth's surface magnetic field, which is the defining feature of geomagnetic storms.

Charged particles trapped by the geomagnetic field undergo an azimuthal drift in the inner magnetosphere. This drift is the net effect of the gradient and curvature of the magnetic field and it is oppositely directed for ions and electrons. The most energetic of these trapped particles comprise the radiation belts, which were discussed in Sections 6.2 and 6.3.

Ions in the medium-energy range of ~ 10 keV to a few hundreds of keV contribute, due to their abundance, substantially to the total current density and to the global geomagnetic disturbances on the Earth surface. These ions constitute the ring current (Daglis *et al.*, 1999). As described in several excellent reference books (such as Northrop, 1963; Roederer, 1970), the general motion of ring current particles (which also holds for radiation belt particles) in the magnetic mirror geometry of the Earth's quasi-dipolar magnetic field is subject to three quasi-periodic motions:

(1) A cyclotron motion or gyration around the so-called guiding centre.
(2) A bounce motion along field lines and between mirror points.
(3) A drift motion around the Earth on a closed three-dimensional drift shell around the magnetic field axis.

Each motion is associated with an adiabatic invariant; namely the first, second and third adiabatic invariant (μ, J, Φ) respectively (Daglis *et al.*, 1999). The theoretically foreseen adiabatic invariants were experimentally verified by the Argus experiment (Christofilos, 1959).

As first worked out by Parker (1957), the total current density perpendicular to the magnetic field $j_\perp$ for a plasma under equilibrium with the magnetic stress balancing the particle pressure, is:

$$j_\perp = \frac{\vec{B}}{B^2} \times \left[\nabla P_\perp + (P_\parallel - P_\perp) \frac{(\vec{B} \cdot \nabla)}{B^2} \vec{B} \right]$$

The current does not depend on gradients of the magnetic field, while in the case of an isotropic distribution ($P_\parallel = P_\perp$), or a straight magnetic field line geometry, the magnetic field configuration plays no role in the current build-up. The current system is then established only by particle pressure gradients.

The first long-lived dispute in ring current research referred to the question of its origin. From the dawn of the space era to the mid-1980s the Sun was the preferred

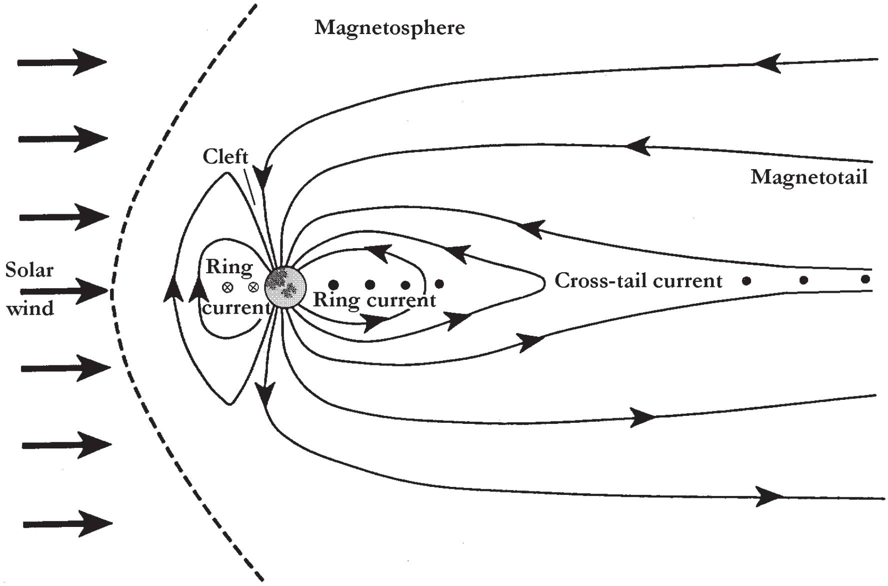


Figure 6.11. Schematic side-view of the terrestrial magnetosphere, with the ring current locus presented near the Earth (from Daglis *et al.*, 1999).

ring current source. According to the solar origin paradigm, solar wind ions penetrate inside the magnetosphere and populate it. The solar origin paradigm dominated for a couple of decades, despite the discovery of ionospheric ions in the magnetosphere in the early 1970s. For many years, the significance of the ionosphere as a source of ring current ions (or magnetospheric ions in general) was underestimated or considered negligible.

It was not before the AMPTE mission that the question of ring current sources was answered satisfactorily. In May 1985, a few months after the launch of AMPTE, the source and the composition of the ring current and its formation process were totally unknown; the data from AMPTE were expected to bring long awaited answers (Williams, 1985).

Several missions in the 1980s showed the significance of the ionospheric particle source during storms (Young *et al.*, 1982). However, the conclusive composition measurements covering the whole energy range important for the storm-time ring current were performed by the AMPTE mission. The charge–energy–mass (CHEM) spectrometer (Gloeckler *et al.*, 1985a) onboard AMPTE/CCE was the first experiment to investigate the near-Earth magnetosphere with multi-species ion measurements in the energy range of the bulk ring current ($\sim$20–300 keV).

AMPTE showed that H^+ ions usually dominate the inner magnetosphere and the ring current, and that O^+ ions originating in the ionosphere are an important part of the ring current population in the disturbed geospace (Gloeckler *et al.*, 1985b;

Hamilton *et al.*, 1988; Daglis *et al.*, 1993). In particular, the AMPTE mission showed that the major source of the quiet-time ring current is the solar wind (via the storage region of the plasma sheet), while the storm-time ring current is increasingly terrestrial in origin.

The next important milestone was achieved through measurements by the Magnetospheric Ion Composition Spectrometer (MICS) onboard the Combined Release and Radiation Effects Satellite (CRRES), which operated around solar maximum (Wilken *et al.*, 1992). Daglis (1997a) demonstrated that in all intense storms observed during the CRRES lifetime, ionospheric O^+ was the dominant ion species in the ring current. While Hamilton *et al.* (1988) had shown an O^+ dominance in the inner ring current only (Figure 6.12) for just one storm, Daglis (1997a) demonstrated that O^+ dominated not only in the inner ring current, but throughout the ring current and for all intense storms in 1991. During the particularly intense storm of March 24–26, 1991, the peak O^+ contribution to the total ring current energy was over 80% (Daglis *et al.*, 1999b). It is noteworthy that MICS was underestimating the O^+ contribution to the total energy density, because of its high energy threshold of 50 keV.

In general, the source question has been answered satisfactorily (Daglis *et al.*, 1999a, table 1). However, there is a remaining uncertainty regarding the percentage of H^+ originating in the terrestrial atmosphere rather than in the solar wind. Magnetospheric H^+ ions originate both in the ionosphere and in the solar wind – which complicates the identification of the dominant source. For the quiet-time ring current, Gloeckler and Hamilton (1987) had estimated that $\sim$35% of H^+ ions in the outer ring current ($L = 5$–7) and $\sim$75% of H^+ ions in the inner ring current ($L = 3$–5) are of ionospheric origin. For the storm-time ring current, Gloeckler and Hamilton (1987) estimated that $\sim$30% of H^+ ions in the outer ring current and $\sim$65% of H^+ ions in the inner ring current are of ionospheric origin. Solar wind He^{++} ions usually contribute less than 4% of the ring current, except in the case of great storms.

The general picture of ring current formation and intensification is understood quite well. Particles that originate from the solar wind or the ionosphere are transported from the magnetotail and the plasma sheet to the inner magnetosphere during time intervals of enhanced convection.

The basic transport and acceleration process for ions moving from the magnetotail and the plasma sheet to the inner magnetosphere is the $\mathbf{E} \times \mathbf{B}$ drift imposed by the large-scale electric field in the night-side magnetosphere. In the magnetotail the particles presumably gain energy while they move from regions of weaker to stronger magnetic field, conserving their first adiabatic invariant. While approaching the inner magnetosphere, the particles are transported across magnetic field lines primarily by gradient and curvature drift, as well as by $\mathbf{E} \times \mathbf{B}$ drift in a complicated combination of potential and induction electric fields.

The large-scale potential electric fields in the night-side magnetosphere are due to prolonged southward interplanetary magnetic fields, while the impulsive induced electric fields are due to magnetic field reconfigurations during substorms. There is an ongoing dispute on the relative importance of the large-scale convection electric

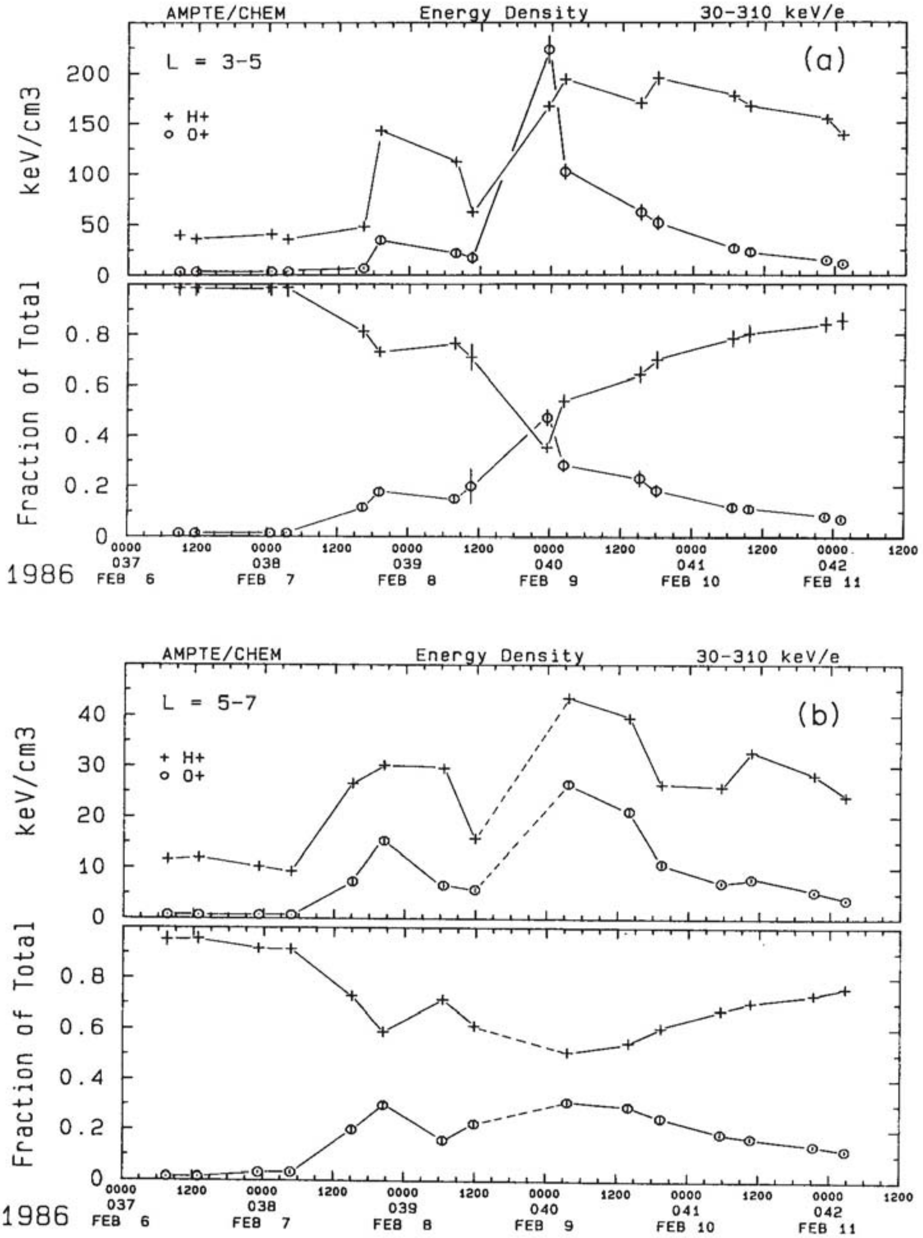


Figure 6.12. The ring current during the intense storm of February 1986: (a) the top panel shows the H^+ and O^+ energy density in the inner ring current ($L = 3$–5); the bottom panel shows the fraction each species comprises of the total ion energy density in that region. (b) The same data for the outer ring current ($L = 5$–7) (Hamilton *et al.*, 1988, Figure 4).

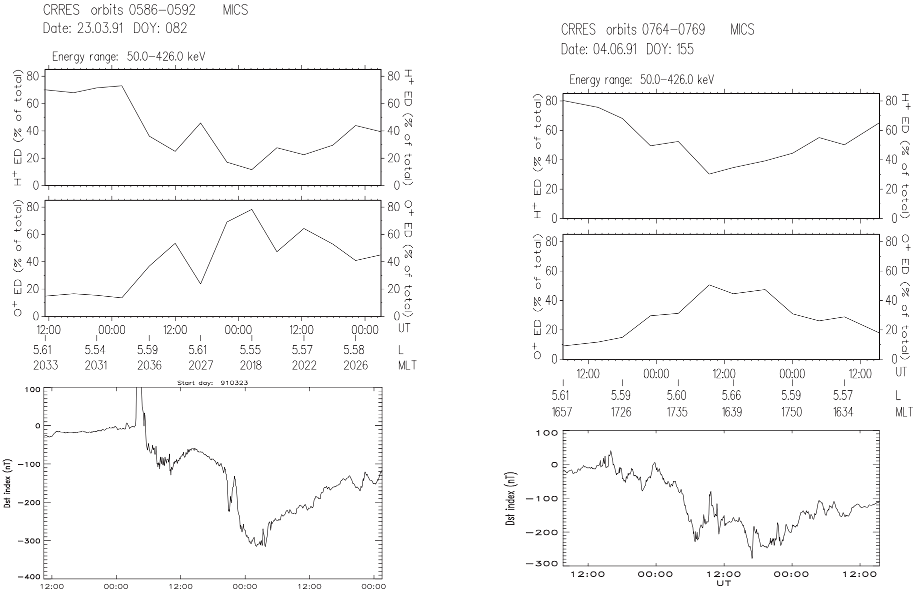


Figure 6.13. Ring current composition measurements by MICS onboard CRRES during two intense storms in March 1991 (left) and June 1991 (right). The upper panels show the H^+ (top) and O^+ (middle) contribution total ion energy density in the energy range 50–426 keV. The bottom panel shows the 5-minute D_{st} index prepared by the Solar–Terrestrial Environment Laboratory of Nagoya University. It is remarkable that D_{st} minima and O^+ maxima occur concurrently (adapted from Daglis *et al.*, 1999a).

field and the substorm-associated impulsive electric fields in the energization and transport of ions into the ring current, which will be discussed in the next section.

6.5 RING CURRENT DYNAMICS

Open questions and critical issues of ring current dynamics pertain on build-up/enhancement, as well as decay mechanisms of the ring current. Presently, one of the most prevalent disputes on ring current dynamics often appears as the ‘substorm effects’ dispute. The name stems from the storm–substorm paradigm, according to which storms are the result of a superposition of successive ‘substorms’. This was the view of Sydney Chapman, who thought of substorms as being the key elements of a magnetic storm and thus named them ‘substorms’ to evoke this idea (Chapman, 1962; Akasofu, 1968). In the picture originally drawn by Sydney Chapman and Syun-Ichi Akasofu, substorms had the role of magnetic pumps, each of which inflates the inner magnetosphere with hot plasma. During substorm expansion, magnetospheric particles are accelerated and injected into the inner magnetosphere, where they become trapped and ultimately form the ring current. According to the classic Chapman–Akasofu picture, magnetic storms occur when substorms deliver hot plasma to the inner magnetosphere faster than it can be dissipated.

This storm–substorm relation paradigm has been questioned in recent years. Several studies have addressed the issue, without however achieving conclusive evidence (Kamide, 1992; Kamide and Allen, 1997; Kamide *et al.*, 1998; Daglis *et al.*, 2000; Ebihara and Ejiri, 2000). The foundations of the storm–substorm dispute lie in correlation studies between auroral electrojet indices AE and the Dst index. While early studies showed a causal relationship between substorms and storms, investigations that followed have questioned this relationship. Such investigations have suggested that substorm occurrence is incidental to the main phase of storms, and that ion transport into the ring current is accomplished solely by enhanced large-scale magnetospheric convection.

To mention a couple of these studies, Iyemori and Rao (1996) identified a total of 28 geospace storms 'containing' substorm expansion onsets and showed that after substorm onset there was no storm development noticeable in the average SYM-H value (a high-resolution Dst index). The authors concluded that the ring current development is not the result of the frequent occurrence of substorms, but the result of enhanced convection caused by large southward IMF. Along the same line, McPherron (1997) noted that visual inspections of AL and Dst time series during storms generally show that substorms occur at times when there is no ring current development or when Dst is recovering.

Intrigued by such studies, Sun and Akasofu (2000) suggested that it is more appropriate to examine the relationship between the corrected ring current intensity Dst^* and the upward field-aligned current density, instead of the standard Dst and AE indices. Sun and Akasofu proposed the new approach in order to accommodate the dominance of ionospheric ions in the ring current during intense storms (Daglis, 1997a, b; Daglis *et al.*, 1999b). Using the Method of Natural Orthogonal Components (Sun *et al.*, 1998), Sun and Akasofu showed that the directly driven component (DD) of the upward field-aligned currents is poorly correlated to the corrected Dst index (correlation coefficient of 0.33), while the unloading component (UL) correlates much stronger (correlation coefficient of 0.81). This indicates that the upward field-aligned currents during substorms play an important role in the formation of the ring current. Sun and Akasofu further concluded that the poor correlation between DD and Dst indicates that the formation of the ring current is not the result of enhanced convection. The strong correlation between UL and Dst, on the other side, is consistent with the observational evidence that the magnetosphere–ionosphere coupling plays an important role in the ring current growth. We will come back to this issue a few paragraphs later, when discussing magnetospheric ion composition.

The central question is whether substorms can substantially contribute to increased particle acceleration leading to the storm-time ring current build-up. A basic argument of the anti-substorm polemic is that very often the occurrence of substorms does not have any effect on ring current intensity and is therefore not followed by storm development. This argument is based on the unspoken *a priori* assumption that storm-time substorms do not differ from non-storm substorms: to 'substorm-opponents', the inability of non-storm substorms to produce storms condemns all substorms to 'storm-impotence'. However, no proof exists that all species in the substorm-zoo are the same with regard to storm-efficiency.

Daglis and Kamide (2003) and Daglis *et al.* (2004b) have argued that the role of substorm induced impulsive electric fields is decisive in obtaining the massive ion acceleration observed during storms, and that there is a synergy between large-scale convection and substorm dipolarizations in building up the storm-time ring current.

Some studies have already addressed this aspect, showing that indeed substorm particle acceleration is very efficient, and that substorms can contribute to storm development. For example, Fok *et al.* (1999) investigated the effect of substorms in kinetic simulations and concluded that global convection and substorm dipolarizations cooperate to inject plasma energy more deeply into the magnetosphere than either one would do individually.

The strong connection between prominent compositional changes in the inner magnetosphere and both intense storms and substorm occurrence, can provide important clues on the storm–substorm relation and the ring current buildup. Observational studies consistently show that the energy density of O^+ increases drastically during storms, much more drastically than the energy density of H^+ (Daglis, 1997a; Nosé *et al.*, 2001), implying that the acceleration mechanisms for ions in the near-Earth plasma sheet are mass dependent. The demonstration of wide ionospheric dominance in intense storms (Daglis, 1997a, b; Daglis *et al.*, 1999b) made massive ionospheric ion supply to the ring current a key issue. What is the cause of explosive O^+ enhancements during intense storms. Previously, Daglis *et al.* (1994) and Daglis and Axford (1996) had shown the occurrence of fast and bulky ionospheric ion feeding of the inner magnetosphere during substorms. The O^+ dominance during intense storms may well be due to substorm occurrence.

Further evidence for a storm–substorm synergy comes from a model by Rothwell *et al.* (1988), predicting that higher concentrations of O^+ in the nightside magnetosphere would permit substorm onset at successively lower L-values. Remarkably, an older storm study by Konradi *et al.* (1976) had shown that the substorm injection boundary was displaced earthward with each successive substorm during the storm. Combining these studies, and the fact that O^+ abundance increases with storm size (Daglis, 1997a), we arrive at the following storm–substorm scenario. During series of intense substorms, which always occur during large storms, the ionospheric feeding of the magnetosphere is most effective, both in quantity and in spatial extent. Such intense O^+ outflows can facilitate successive inward penetration of substorm ion injections, according to the Rothwell *et al.* (1988) model and consistent with the observations reported by Konradi *et al.* (1976) and Daglis (1997a). The result of successive inward penetration of substorm injections would be the trapping of increasingly more energetic ions, resulting in the build-up of an ever stronger ring current. The scenario of progressively earthward and duskward substorm onsets and accompanying ion injections during storm-time substorms can be investigated through global imaging techniques.

The very large energies attained by ring current ions can perhaps only be

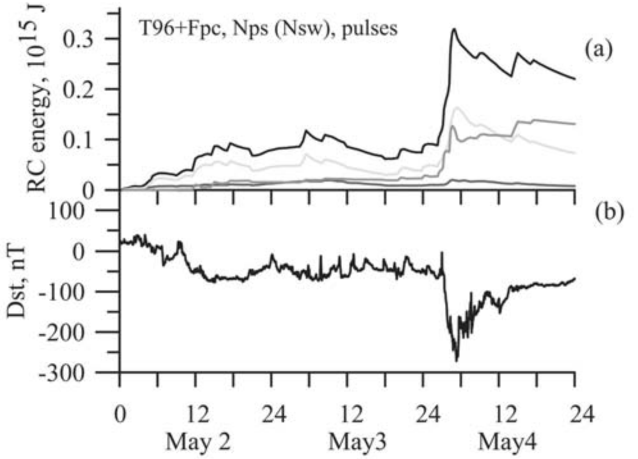


Figure 6.14. (Upper panel) calculated ring current proton energy in Joules for four energy ranges, incorporating several electromagnetic pulses simulating substorm-associated impulsive electric fields. (Lower panel) Dst index for the modelled time interval 2–4 May 1998 (Ganushkina *et al.*, 2005).

accounted for by the action of substorm-associated impulsive electric fields. Recently, Ganushkina *et al.* (2005) used a particle tracing code to investigate the role of substorm-associated impulsive electric fields in the transport and energization of ring current particles to energies above 80 keV, in the energy range that plays a dominant role during the storm recovery phase. They showed that the observed flux of high-energy protons cannot be obtained by simply using variable intensity of the large-scale convection electric field or by changing the initial distribution and/or temperature. Only the impulsive localized substorm-associated electric fields were able to yield the observed fluxes of high-energy protons.

Transient impulsive electric fields with amplitudes of up to 20 mV/m were noticed by Wygant *et al.* (1998) in the great storm of March 1991, although they only studied the large-scale convection electric field in detail. The detailed relation and possible combined action of convection and induction electric fields has still to be explored. Aggson and Heppner (1977) had shown that the impulsive induction electric fields observed in the inner magnetosphere can cause violation of the second and third adiabatic invariant for H^+, and violation of all three adiabatic invariant for heavier ions. Delcourt (2002) and Metallinou *et al.* (2005) have performed detailed simulations and showed that O^+ ions with initial energies of 100 eV may reach energies above 100 keV, while H^+ ions with the same initial energies are limited to final energies below 20 keV. The reason for this preferential acceleration of O^+ ions is related to the breakdown of the first adiabatic invariant, and its dependence on particle mass. This is further supported by the fact that results of studies assuming adiabaticity (such as Mauk, 1986; Lewis *et al.*, 1990) are incompatible with the actual energy spectra observed during substorms (such as Kistler *et al.*, 1990).

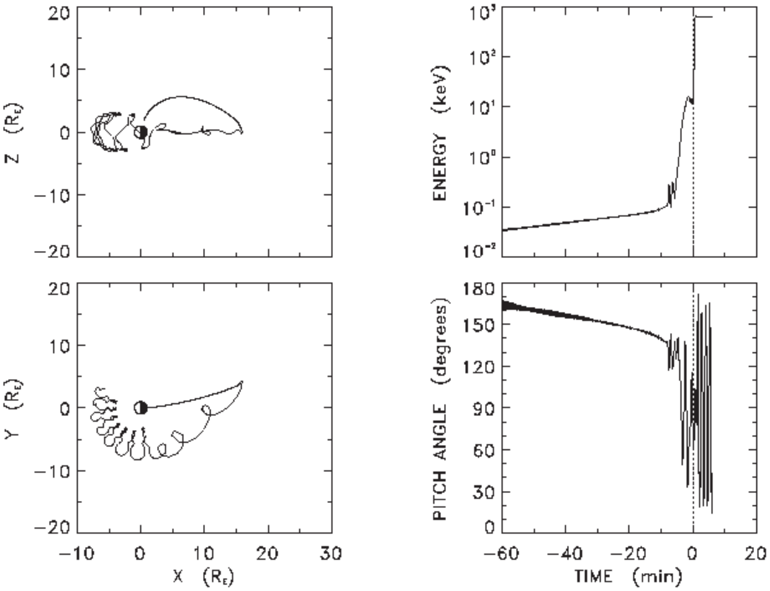


Figure 6.15. Sample trajectory of an O^+ ion launched from the night-side auroral zone (0000 MLT) with an initial energy of 30 eV and encountering a substorm dipolarization at time $t = 0$. The efficient acceleration of the ion is evident in the top right panel (Daglis *et al.*, 2004b).

Finally, studies of DMSP (Defense Meteorological Satellite Program) measurements (Shiokawa and Yumoto, 1993) showed that high-latitude field-aligned potential drops, which are efficient agents of ionospheric ion acceleration, exhibit strong enhancements during the substorm recovery phase. Such enhancements will increase the rate of ionospheric ion feeding of the inner plasma sheet during substorm recovery, especially in a series of successive substorms. The obvious result is a substorm-driven net growth of the ring current far from substorm expansion. The criticism against substorm build-up of the ring current has focused on substorm expansion, leaving the case open as to whether other substorm phases influence the ring current growth.

The debate on the 'storm–substorm relationship' will presumably continue for some time, as no school of thought has managed to present conclusive evidence for their case. McPherron (1997) – who is often misquoted with regard to the role of substorms in ring current growth – actually based his argumentation on Dst timing and prediction. He showed that Dst begins to decrease before the first substorm expansion, and that the solar wind dynamic pressure and electric field account for over 85% of the variance of Dst. Accordingly he concluded that the ring current growth is not the result of substorm expansion. However, three important points have to be taken into account: 1, Dst is not a purely ring current index; 2, substorm expansion has mixed effects on Dst variation; 3, substorms are not just expansion phase.

Substorm expansion has a double effect: injection of energetic particles to the inner magnetosphere, with a negative effect on Dst; and disruption of the cross-tail current in the near-Earth plasma sheet, with a positive effect on Dst (Ohtani *et al.*, 2001). The net effect can well be positive in Dst, despite the fact that the injected particles contribute to the ring current build-up.

A crucial issue of ring current dynamics relates to the extent of interplanetary driving of ring current enhancement. Combined simulation and observation studies indicate that plasma sheet density is of critical importance to the eventual effect of interplanetary drivers: variations in the plasma sheet density significantly modify the geoeffectiveness of southward IMF regions in the solar wind. Energy input from the solar wind varies as southward IMF (or westward interplanetary electric field), but is also modulated by the plasma sheet density. It appears that ring current intensity follows the profile of the modulated energy input rather than that of the original energy input (Daglis *et al.*, 2003).

The final part of this section concerns the ring current decay and consequent storm recovery. In the past, charge exchange between ring current ions and cold neutral hydrogen of the geocorona had been generally assumed to be the main killer of the storm-time ring current (Daglis and Kozyra, 2002). The often observed feature of the two-step recovery of intense storms has been interpreted as the result of a plasma composition change during the storm recovery (Hamilton *et al.*, 1988; Daglis, 2001a). The reason for this interpretation is that the charge exchange lifetimes of the main ring current ion species, namely H^+ and O^+, are radically different. In the energy range of several tens of keV (50–100 keV), where the bulk of the storm-time ring current energy is contained, the O^+ lifetime can be an order of magnitude shorter than that of H^+. This difference is even large at higher energies.

For example, at $L = 5$ and a mirror latitude of 14°, the charge-exchange lifetime of a 100 keV O^+ ion is $\sim$46 hrs; for the same energy the lifetime of H^+ ions is $\sim$470 hrs. These time-scales become considerably shorter in the inner ring current, because the geocoronal density increases. At $L = 3.5$ the respective 100 keV O^+ and H^+ lifetimes are 11 and 110 hrs. These numbers are consistent with observations showing a dominance of high-energy H^+ (having long lifetimes) during storm recovery.

In contrast, in the lower energy range O^+ has a much longer lifetime than H^+: at $L = 5$ the 10 keV O^+ and H^+ lifetimes are $\sim$56 hrs and $\sim$17 hrs respectively. At $L = 3.5$ the lifetimes become 28 and 5.5 hrs respectively. The implications of these differences are very important. A large (not to mention a dominant) O^+ component can, during intense storms, induce: 1, a rapid initial decay after the storm maximum, due to the rapid loss of high-energy O^+; 2, a decrease of the decay rate during the recovery phase, due to the relatively long lifetimes of high-energy H^+ and low-energy O^+.

It is obvious that variations in the relative abundance of the two main ion species H^+ and O^+ play a regulatory role in the decay rate of the storm-time ring current, and should therefore be taken into account in any comprehensive modelling study. However, some case studies and simulations indicated limitations of charge-exchange losses as the main cause of ring current decay during storm recovery. In particular, the two-phase recovery of intense storms has not been reproducible by charge-exchange losses alone (Fok *et al.*, 1995; Liemohn *et al.*, 1999; Kozyra *et al.*, 2002). Alternative mechanisms for ring current losses are wave–particle interactions, precipitation, and convective drift escape through the day-side magnetopause. The relative contributions of these processes seem to be different from case to case.

Liemohn *et al.* (2001a) suggested that the two-phase recovery of intense storms is due to the transition between fast time-scale 'flow-out' losses (associated with open drift paths) and much slower charge-exchange losses (associated with closed drift paths) rather than the transition between fast O^+ and slower H^+ charge-exchange losses (associated with a trapped ring current). Liemohn *et al.* (2001a) attributed the dramatic loss of O^+ compared to H^+, typically observed in the early recovery phase of major storms, to composition changes in the inner plasma sheet rather than to composition-dependent decay mechanisms in the ring current. This increasingly oxygen-poor plasma convects into the inner magnetosphere along open drift paths, which are being converted to closed drift paths as the convection electric field weakens (magnetic activity subsides).

However, a study of two intense storms that occurred close to solar maximum and solar minimum respectively showed that charge-exchange indeed dominated as decay process during the initial fast recovery of the storm with a large O^+ component (Daglis *et al.*, 2003). This was presumably due to the relatively short charge-exchange lifetime of O^+, as discussed above. The interplay between flow-out and charge exchange losses during the particular storm was the combined effect of the ring current composition and the fluctuations in the interplanetary magnetic field. The IMF fluctuations influence the convection electric field – and, consequently, the convection pattern of the incoming ions and their loss through the day-side magnetopause – while the O^+ dominated ring current favours rapid decay through charge exchange. This is evident in the early, fast recovery phases of the storm, and even more in the late recovery, when charge-exchange losses become much larger than flow-out losses (by a factor of 5–10). In the other storm of the same study, which did not have such a large O^+ component, the initial recovery phase was dominated by convective flow-out losses.

In summary, the various different studies of ring current decay cannot yet be considered conclusive. Nevertheless, it seems that charge-exchange loss is increasingly important in the main and early recovery storm phase for ring currents with an increasingly large O^+ component.

In recent years a paradigm shift has made ring current enhancements quite interesting from the space weather perspective – specifically in the framework of geomagnetically-induced currents (GIC) on power grids and other land-based infrastructures (see Chapter 10). In the past, large impulsive geomagnetic field disturbances associated with auroral electrojet intensifications at mid- and high latitudes had been understood as a concern for power grids in close proximity to these disturbance regions. However, research and observational evidence has determined that the GIC risks are much more broad and more complex than this traditional view.

Surprisingly, it has been shown that turbulent ground-level geomagnetic field disturbances driven by intensifications of the ring current can create large GIC flows at low latitudes, which have actually been observed in central Japan (Daglis *et al.*, 2004a). These disturbance events have also been observed to produce GICs of unusually long duration. Consequently, in future, forecasting of GIC-related problems must also take into account ring current intensifications.

6.6 SYNOPSIS

In this chapter we discussed the origin, formation and dynamics of two major magnetospheric particle populations: the radiation belts and the ring current. The radiation belts and the ring current are part of the space weather chain interconnecting the Sun, interplanetary space and geospace.

The radiation belts are an important and interesting part of the geospace environment both from a basic science view and from a practical standpoint. Although the radiation belts were the first aspect of the space environment discovered at the dawn of the space age, they have continued to intrigue and challenge space researchers. Dramatic – and virtually unprecedented – changes can occur in the radiation belts, as demonstrated during the Halloween Storms of 2003 (Baker *et al.*, 2004). Events of such magnitude – though very rare – serve to illustrate the intimate relationships between the radiation belt particle populations, the cold plasmas of the plasmasphere, and the medium-energy particles that form the ring current. The inner magnetosphere is clearly a tightly coupled and highly organized region within geospace.

Smoothed and/or broadly averaged data were used in this review in order to characterize the outer radiation belt electron environment at different time scales. We have shown that at any L-value, including right at the heart of the outer zone ($L = 4$), there can be tremendous ranges of average particle fluxes. The lowest intensities of electrons are found during summer and winter solstice times around sunspot minimum. The maximum fluxes are found, generally, at equinox times during the declining phase of the sunspot activity cycle. These analyses present quite a different picture of the relativistic electron fluxes in the outer radiation belt as compared to the standard static models. More modern radiation belt models, such as those from CRRES (Brautigam *et al.*, 1992), present somewhat more realistic descriptions of observed radiation belt behaviour than do the essentially static models, but few models yet capture the rich diversity of the real radiation belts.

We have also discussed the systematic response of MeV electrons at geostationary orbit to the variation of solar wind. We have shown some of the predictability of the daily-averaged electron fluxes based on solar wind parameters only. Modern research is using data from spacecraft such as SAMPEX and POLAR to describe variations of the electron populations across all relevant L-shells.

It is crucially important to bear in mind the practical importance of the radiation belts in light of their potential deleterious effects on human technology. High-energy electrons and the deep dielectric charging they can cause remains one of the very challenging aspects of space weather that requires ‘situational awareness’ by spacecraft operators, and very probably demands effective prediction and mitigation strategies (Baker, 2002).

The ring current enhancement is an essential, if not the major, component of geospace magnetic storms. The evolution of instrumentation and the launch and successful operation of a couple of space missions in the 1980s and early 1990s brought significant progress in ring current research. In particular, they clarified the origin of ring current particles by demonstrating the significance of both the solar wind and the terrestrial ionosphere. The two particle sources contribute in a variable manner to the ring current population. The estimated ratio of solar wind to ionospheric contribution ranges from 2 : 1 for quiet times, to 1 : 1 for small to medium storms, to 3 : 7 for intense storms (Daglis *et al.*, 1999a).

Research on ring current dynamics has also advanced significantly in the last decade. This has been achieved through simulations and their interplay with observations. Comprehensive modelling studies of particle acceleration (Delcourt, 2002; Daglis *et al.*, 2004; Metallinou *et al.*, 2005), magnetospheric configuration and particle transport (Liemohn *et al.*, 2001b), ring current decay (Liemohn *et al.*, 2001a; Jordanova *et al.*, 2003), wave-particle interactions (Thorne and Horne, 1994) and global energy balance (Kozyra *et al.*, 2002) have been instrumental in achieving vital progress in ring current research.

Nevertheless, there are still open questions and issues under debate. For example, although the O^+-dominated ring current during intense storms is presumably a substorm effect, it is not clear why not all substorms appear equally efficient in changing ring current composition (Daglis and Kamide, 2003). Is it a matter of preconditioning, of long-term memory of the system? Does it depend on the type of the operating mechanism of ion acceleration? Does it depend on the duration of the mechanism operation? Ionospheric ions are rather cold, and a variety of successive acceleration mechanisms therefore have to act on them to raise their energy from about an eV to tens or hundreds of keV (Daglis and Axford, 1996, and references therein).

Another open issue is the cause of the fast ring current decay following storm maximum of intense storms. It is not yet clear whether charge-exchange losses or convective drift loss account for the massive decay of the ring current observed in the initial recovery phase of intense storms. Such issues have to be addressed by combined modelling and observational studies, which have proved to be very efficient in solving puzzles of geospace dynamic processes.

6.7 REFERENCES

Aggson, T.L., and J.P. Heppner, Observations of large transient magnetospheric electric fields, *J. Geophys. Res.*, **82**, 5155–5164, 1977.

Akasofu, S.-I., *Polar and magnetospheric substorms*, D. Reidel, Dordrecht, 1968.

Albert, J.M., Nonlinear interaction of outer zone electrons with VLF waves, *Geophys. Res. Lett., 10.1029/2001GL013941*, 30 April 2002.

Allen, J.H., and D.C. Wilkinson, Solar–Terrestrial Activity Affecting Systems in Space and on Earth, in *Solar–Terrestrial Predictions – IV*, **1**, 75–107, Ottawa, Canada, May 18–22, 1992.

Baker, D.N., E.W. Hones Jr., D.T. Young, and J. Birn, The possible role of ionospheric oxygen in the initiation and development of plasma sheet instabilities, *Geophys. Res. Lett.*, **9**, 1337–1340, 1982.

Baker, D.N., R.A. Goldberg, F.A. Herrero, J.B. Blake, and L.B. Callis, Satellite and rocket studies of relativistic electrons and their influence on the middle atmosphere, *J. Atmos. Terr. Phys.*, **55**, 1619, 1993a.

Baker, D.N., G.M. Mason, O. Figueroa, G. Colon, J. Watzin, and R. Aleman, An overview of the SAMPEX mission, *IEEE Trans. Geosci. Elec.*, **31**, 531, 1993b.

Baker, D.N., *et al.*, Recurrent geomagnetic storms and relativistic electron enhancements in the outer magnetosphere: ISTP coordinated measurements, *J. Geophys. Res.*, **102**, 14141–14148, 1997.

Baker, D.N., T.I. Pulkkinen, M. Hesse, and R.L. McPherron, A quantitative assessment of energy storage and release in the Earth's magnetotail, *J. Geophys. Res.*, **201**, 7159, 1997.

Baker, D.N., T.I. Pulkkinen, X. Li, S.G. Kanekal, J.B. Blake, R.S. Selesnick, M.G. Henderson, G.D. Reeves, H.E. Spence, and G. Rostoker, Coronal mass ejections, magnetic clouds, and relativistic magnetospheric electron events: ISTP, *J. Geophys. Res.*, **103**, 17279, 1998a.

Baker, D.N., X. Li, J.B. Blake, and S. Kanekal, Strong electron acceleration in the Earth's magnetosphere, *Adv. Space Res.*, **21**, 609, 1998b.

Baker, D.N., T.I. Pulkkinen, X. Li, S.G. Kanekal, K.W. Ogilvie, R.P. Lepping, J.B. Blake, L.B. Callis, G. Rostoker, H.J. Singer, and G.D. Reeves, A strong CME-related magnetic cloud interaction with the earth's magnetosphere: ISTP observations of rapid relativistic electron acceleration on May 15, 1997, *Geophys. Res. Lett.*, **25**, 2975–2978, 1998c.

Baker, D.N., S.G. Kanekal, T.I. Pulkkinen, and J.B. Blake, Equinoctial and solstitial averages of magnetospheric relativistic electrons: A strong semiannual modulation, *Geophys. Res. Lett.*, **26**, No. 20, 3193, 1999.

Baker, D.N., The occurrence of operational anomalies in spacecraft and their relationship to space weather, *IEEE Trans. Plasma Sci.*, **28**, 2007, 2000.

Baker, D.N., S.G. Kanekal, J.B. Blake, and T.I. Pulkkinen, The global efficiency of relativistic electron production in the Earth's magnetosphere, *J. Geophys. Res.*, **106**, 19169, 2001.

Baker, D.N., How to cope with space weather, *Science*, **297**, 30 Aug 2002.

Baker, D.N., S.G. Kanekal, X. Li, S.P. Monk, J. Goldstein, and J.L. Burch, An extreme distortion of the Van Allen belt arising from the 'Hallowe'en' solar storm in 2003, *Nature*, doi:10.1038/nature03116, 2004.

Baker, D.N., Specifying and forecasting space weather threats to human technology, in *Effects of Space Weather on Technology Infrastructure*, edited by I.A. Daglis, pp. 1– 25, Kluwer Academic Publishers, Dordrecht, 2004.

Baker, D.N., and X. Li, Relativistic electron flux enhancements during strong geomagnetic activity, *AGU Monograph Disturbances in Geospace: The Storm–Substorm Relationship. Monograph*, **142**, 217–230, doi: 10.1029/142GM18, 2003.

Blake, J.B., *et al.*, CEPPAD: Comprehensive energetic particle and pitch angle distribution experiment on POLAR, *Space Sci. Rev.*, **71**, 531, 1995.

Brautigam, D.H., M.S. Gussenhoven, and E.G. Mullen, Quasi-static model of outer zone electrons, *IEEE Trans. Nucl. Sci.*, **39**, 1797, 1992.

Chapman, S., and V.C.A. Ferraro, A new theory of magnetic storms, *Nature*, **126**, 129–130, 1930.

Chapman, S., and V.C.A. Ferraro, A new theory of magnetic storms, I. The initial phase, *Terrest. Magn. Atmosph. Elec.*, **36**, 77–97, 1931.

Chapman, S., Earth storms: Retrospect and prospect, *J. Phys. Soc. Japan*, **17** (Suppl. A-I), 6–16, 1962.

Chen, M.W., L. Lyons, and M. Schultz, Simulations of phase space distributions of storm time proton ring current, *J. Geophys. Res.*, **99**, 5745–5759, 1994.

Christofilos, N.C., The Argus Experiment, *J. Geophys. Res.*, **64**, 869–875, 1959.

Daglis, I.A., The role of magnetosphere-ionosphere coupling in magnetic storm dynamics, in *Magnetic Storms, Geophys. Monogr. Ser.*, **98**, edited by B.T. Tsurutani, W.D. Gonzalez, Y. Kamide, and J.K. Arballo, pp. 107–116, American Geophysical Union, Washington, DC, 1997a.

Daglis, I.A., Terrestrial agents in the realm of space storms: Missions study oxygen ions, *Eos Trans. AGU*, **78**(24), 245–251, 1997b.

Daglis, I.A., Space storms, ring current and space-atmosphere coupling, in *Space Storms and Space Weather Hazards*, edited by I. A. Daglis, pp. 1–42, Kluwer Academic Publishers, Dordrecht, 2001a.

Daglis, I.A., The storm-time ring current, *Space Sci. Rev.*, **98**, 343–363, 2001b.

Daglis, I.A., Magnetic Storm – still an adequate name?, *Eos Trans. AGU*, **84**(22), 207–208, 2003.

Daglis, I.A., and W.I. Axford, Fast ionospheric response to enhanced activity in geospace: Ion feeding of the inner magnetotail, *J. Geophys. Res.*, **101**, 5047–5065, 1996.

Daglis, I.A., and Y. Kamide, The role of substorms in storm-time particle acceleration, in *Disturbances in Geospace: The Storm–Substorm Relationship*, edited by A.S. Sharma, Y. Kamide, and G.S. Lakhina, pp. 119–129, Geophysical Monograph Series 142, DOI 10.1029/142GM11, American Geophysical Union, Washington, DC, 2003.

Daglis, I.A., and J.U. Kozyra (2002), Outstanding issues of ring current dynamics, *J. Atmos. Sol.–Terr. Phys.*, **64**, 253–264.

Daglis, I.A., E.T. Sarris, and B. Wilken, AMPTE/CCE observations of the ion population at geosynchronous altitudes, *Ann. Geophys.*, **11**, 685–696, 1993.

Daglis, I.A., S. Livi, E.T. Sarris, and B. Wilken, Energy density of ionospheric and solar wind origin ions in the near-Earth magnetotail during substorms, *J. Geophys. Res.*, **99**, 5691–5703, 1994.

Daglis, I.A., R.M. Thorne, W. Baumjohann, and S. Orsini, The terrestrial ring current: Origin, formation, and decay, *Rev. Geophys.*, **37**, 407–438, 1999a.

Daglis, I.A., G. Kasotakis, E.T. Sarris, Y. Kamide, S. Livi, and B. Wilken, Variations of the ion composition during a large magnetic storm and their consequences, *Phys. Chem. Earth*, **24**, 229–232, 1999b.

Daglis, I.A., Y. Kamide, C. Mouikis, G.D. Reeves, E.T. Sarris, K. Shiokawa, and B. Wilken, 'Fine structure' of the storm–substorm relationship: Ion injections during Dst decrease, *Adv. Space Res.*, **25**(12), 2369–2372, 2000.

Daglis, I.A., J.U. Kozyra, Y. Kamide, D. Vassiliadis, A.S. Sharma, M.W. Liemohn, W.D. Gonzalez, B.T. Tsurutani, and G. Lu, Intense space storms: Critical issues and open disputes, *J. Geophys. Res.*, **108**(A5), 1208, doi:10.1029/2002JA009722, 2003.

Daglis, I.A., D.N. Baker, J.G. Kappenman, M. Panasyuk, and E.J. Daly, Effects of space weather on technology infrastructure, *Space Weather*, **2**, S02004, doi:10.1029/2003SW000044, 2004a.

Daglis, I.A., D. Delcourt, F.-A. Metallinou, and Y. Kamide, Particle acceleration in the frame of the storm–substorm relation, *IEEE Trans. Plasma Sci.*, **32**(4), 1449–1454, 2004b.

Delcourt, D.C., Particle acceleration by inductive electric fields in the inner magnetosphere, *J. Atmos. Sol. Terr. Phys.*, **64**, 551–559, 2002.

Ebihara, Y., and M. Ejiri, Simulation study on fundamental properties of the storm-time ring current, *J. Geophys. Res.*, **105**, 15843–15859, 2000.

Fok, M.C., *et al.*, 3-Dimensional Ring Current Decay Model, *J. Geophys. Res.–Space Phys.*, **100**, 9619–9632, 1995.

Fok, M.-C., T.E. Moore, and D.C. Delcourt, Modelling of inner plasma sheet and ring current during substorms, *J. Geophys. Res.*, **104**, 14,557–14,569, 1999.

Gloeckler, G., and D.C. Hamilton, AMPTE ion composition results, *Phys. Scr.*, **T18**, 73–84, 1987.

Gloeckler, G., *et al.*, The charge-energy-mass (CHEM) spectrometer for 0.3 to 300 keV/*e* ions on the AMPTE/CCE, *IEEE Trans. Geosci. Remote Sens.*, *GE-23*, 234–240, 1985a.

Gloeckler, G., B. Wilken, W. Stüdemann, F.M. Ipavich, D. Hovestadt, D.C. Hamilton, and G. Kremser, First composition measurements of the bulk of the storm time ring current (1 to 300 keV/*e*) with AMPTE/CCE, *Geophys. Res. Lett.*, **12**, 325–328, 1985b.

Gosling, J.T., The solar flare myth, *J. Geophys. Res.*, **98**, 18937–18949, 1993.

Hamilton, D.C., G. Gloeckler, F.M. Ipavich, W. Stüdemann, B. Wilken, and G. Kremser, Ring current development during the great geomagnetic storm of February 1986, *J. Geophys. Res.*, **93**, 14343–14355, 1988.

Horne, R.B., and R.M. Thorne, Potential waves for relativistic electron scattering and stochastic acceleration during magnetic storms, *Geophys. Res. Lett.*, **25**, 3011, 1998.

Hudson, M.K., S.R. Elkington, J.G. Lyon, and C.C. Goodrich, Increase in relativistic electron flux in the inner magnetosphere: ULF wave move structure, *Adv. Space Res.*, **25**, 2327, 2000.

Iyemori, T., and D.R.K. Rao, Decay of the Dst field of geomagnetic disturbance after substorm onset and its implication to storm–substorm relation, *Ann. Geophys.*, **14**, pp. 608–618, 1996.

Jordanova, V.K., L.M. Kistler, M.F. Thomsen, and C.G. Mouikis, Effects of plasma sheet variability on the fast initial ring current decay, *Geophys. Res. Lett.*, **30**, 1311, doi:10.1029/2002GL016576, 2003.

Kamide, Y., Is substorm occurrence a necessary condition for a magnetic storm?, *J. Geomagn. Geoelectr.*, **44**, 109–117, 1992.

Kamide, Y., and J.-H. Allen, Some outstanding problems of the storm/substorm relationship, in *Solar–Terrestrial Predictions*, edited by G. Heckman, K. Marubashi, M.A. Shea, D.F. Smart, and R. Thompson, pp. 207–216, Communications Research Laboratory, Tokyo, 1997.

Kamide, Y., *et al.*, Current understanding of magnetic storms: Storm–substorm relationships, *J. Geophys. Res.–Space Phys.*, **103**, 17705–17728, 1998.

Kanekal, S.G., D.N. Baker, J.B. Blake, B. Klecker, R.A. Mewaldt, and G.M. Mason, Magnetospheric response to magnetic cloud (CME) events: Energetic particle observations from SAMPEX and POLAR, *J. Geophys. Res.*, **104**, 14885, 1999.

Kistler, L.M., E. Möbius, B. Klecker, G. Gloeckler, F.M. Ipavich, and D.C. Hamilton, Spatial variations in the suprathermal ion distribution during substorms in the plasma sheet, *J. Geophys. Res.*, **95**, 18871–18885, 1990.

Konradi, A., C.L. Semar, and T. A Fritz, Injection boundary dynamics during a geomagnetic storm, *J. Geophys. Res.*, **81**, 3851–3865, 1976.

Koons, H.C., and D.J. Gorney, The relationship between electrostatic discharges on spacecraft P78-2 and the electron environment, The Aerospace Corp., El Segundo, CA, *TR-0091 (6940 06)-2*, Mar. 1992.

Kozyra, J.U., M.W. Liemohn, C.R. Clauer, A.J. Ridley, M.F. Thomsen, J.E. Borovsky, J.L. Roeder, V.K. Jordanova, and W.D. Gonzalez, Multistep Dst development and ring current composition changes during the 4–6 June 1991 magnetic storm, *J. Geophys. Res.*, **107**(A8), DOI 10.1029/2001JA00023, 2002.

Krimigis, S.M., G. Gloeckler, R.W. McEntire, T.A. Potemra, F.L. Scarf, and E.G. Shelley, Magnetic storm of 4 September 1985: A synthesis of ring current spectra and energy densities measured with AMPTE-CCE, *Geophys. Res. Lett.*, **12**, 329–332, 1985.

Lewis, Z.V., S.W.H. Cowley, and D.J. Southwood, Impulsive energization of ions in the near-Earth magnetotail during substorms, *Planet. Space Sci.*, **38**, 491–505, 1990.

Li, X., D.N. Baker, M. Temerin, D. Larson, R.P. Lin, G.D. Reeves, M.D. Looper, S.G. Kanekal, and R.A. Mewaldt, Are energetic electrons in the solar wind the source of the outer radiation belt? *Geophys. Res. Lett.*, **24**, 923, 1997a.

Li, X., D.N. Baker, M. Temerin, D. Larson, R.P. Lin, G.D. Reeves, J.B. Blake, M. Looper, R. Selesnick, and R.A. Mewaldt, Source of Relativistic Electrons in the Magnetosphere: Present Knowledge and Remaining Questions, presented at the IAGA, Uppsala, Sweden, 1997b.

Li, X., M. Temerin, D. N. Baker, G. D. Reeves, and D. Larson, Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements, *Geophys. Res. Lett.*, **28**, 1887, 2001a.

Li, X., D.N. Baker, S.G. Kanekal, M. Looper, and M. Temerin, Long term measurements of radiation belts by SAMPEX and their variations, *Geophys. Res. Lett.*, **28**, 3827, 2001b.

Li, X., Variations of 0.7–6.0 MeV Electrons at geosynchronous orbit as a function of solar wind, *Space Weather*, **2**, No. 3, S0300610.1029/2003SW000017, 2004.

Liemohn, M.W., J.U. Kozyra, V.K. Jordanova, G.V. Khazanov, M.F. Thomsen, and T.E. Cayton, Analysis of early phase ring current recovery mechanisms during geomagnetic storms, *Geophys. Res. Lett.*, **26**, 2845–2849, 1999.

Liemohn, M.W., J.U. Kozyra, C.R. Clauer, and A.J. Ridley, Computational analysis of the near-Earth magnetospheric current system during two-phase decay storms, *J. Geophys. Res.*, **106**, 29531–29542, 2001a.

Liemohn, M.W., J.U. Kozyra, M.F. Thomsen, J.L. Roeder, G. Lu, J.E. Borovsky, and T.E. Cayton, Dominant role of the asymmetric ring current in producing the stormtime Dst*, *J. Geophys. Res.*, **106**, 10883–10904, 2001b.

Lopez, R.E., D.N. Baker, and J.H. Allen, Sun unleashes Halloween Storm, *Eos*, **85**, No. 11, 105, 108, 2004.

Mauk, B.H., Quantitative modeling of the convection surge mechanism of ion acceleration, *J. Geophys. Res.*, **91**, 13423–13431, 1986.

Meredith, N.P., R.B. Horne, and R.R. Anderson, Substorm dependence of chorus amplitudes: Implications for the acceleration of electrons relativistic energies, *J. Geophys. Res.*, **106**, 13165, 2001.

Meredith, N.P., R.B. Horne, R.H. Iles, R.M. Thorne, D. Heynderickx, and R.R. Anderson, Outer zone relativistic electron acceleration associated with substorm enhanced whistler mode chorus, *J. Geophys. Res.*, **107,** 1144, doi: 10.1029/2001JA900146, 2002.

Metallinou, F.-A., I.A. Daglis, and D. Delcourt, The influence of substorm-induced electric fields on the build-up of particle radiation during geospace magnetic storms, *Annales Geophysicae*, submitted 2005.

Mitchell, D.G., The space environment, in *Fundamentals of Space Systems*, edited by V.L. Pisacane and R.C. Moore, pp. 45–98, Oxford University Press, Oxford, 1994.

Neugebauer, M., and C. Snyder, The mission of Mariner II: Preliminary observations: Solar plasma experiments, *Science*, **138**, 1095–1096, 1962.

Northrop, T.G., *The Adiabatic Motion of Charged Particles*, Interscience, New York, 1963.

Nosé, M., S. Ohtani, K. Takahashi, A.T.Y. Lui, R.W. McEntire, D.J. Williams, S.P. Christon, and K. Yumoto, Ion composition of the near-Earth plasma sheet in storm and quiet intervals: Geotail/EPIC measurements, *J. Geophys. Res.*, **106**, 8391–8404, 2001.

Ohtani, S., M. Nosé, G. Rostoker, H. Singer, A.T.Y. Lui, and M. Nakamura, Storm–substorm relationship: Contribution of the tail current to Dst, *J. Geophys. Res.*, **106**, 21199–21209, 2001.

Parker, E.N., Newtonian development of the dynamical properties of the ionised gases at low density, *Phys. Rev.*, **107**, 924–933, 1957.

Parker, E.N., Interaction of the solar wind with the geomagnetic field, *Phys. Fluids*, **1**, 171, 1958.

Reeves, G.D., Relativistic electrons and magnetic storms: 1992–1995, *Geophys. Res. Lett.*, **25**, 1817, 1998.

Roederer, J.G., *Dynamics of Geomagnetically Trapped Radiation*, Springer Verlag, Berlin, 1970.

Rostoker, G., S. Skone, and D.N. Baker, On the origin of relativistic electrons in the magnetosphere associated with some geomagnetic storms, *Geophys. Res. Lett.*, **25**, 3701, 1998.

Rothwell, P.L., L.P. Block, M.B. Silevitch, and C.-G. Fälthammar, A new model for substorm onsets: The pre-breakup and triggering regimes, *Geophys. Res. Lett.*, **15**, 1279–1282, 1988.

Schulz, M. and L.J. Lanzerotti, *Particle Diffusion in the Radiation Belts*, Springer Verlag, New York, 1974.

Selesnick, R.S., and J.B. Blake, Dynamics of the outer radiation belt, *Geophys. Res. Lett.*, **24**, 1347, 1997.

Simpson, S., Massive solar storms inflict little damage on Earth, *Space Weather*, **1**, 1012, doi:10.1029/2003SW000042, 2003.

Singer, S.F., Trapped orbits in the Earth's dipole field, *Bull. Am. Phys. Soc. Series II*, **1**, 229 (A), 1956.

Singer, S.F., A new model of magnetic storms and aurorae, *Trans. Am. Geophys. Union*, **38**, 175–190, 1957.

Summers, D., R.M. Thorne, and F. Xiao, Relativistic theory of wave-particle resonant diffusion with application to electron acceleration in the magnetosphere, *J. Geophys. Res.*, **103**, 20487, 1998.

Sun, W., and S.-I. Akasofu, 'On the formation of the storm-time ring current belt', *J. Geophys. Res.*, **105**, 5411–5418, 2000.

Sun, W., W.-Y. Xu, and S.-I. Akasofu, Mathematical separation of directly driven and unloading components in the ionospheric equivalent currents during substorms, *J. Geophys. Res.*, **103**, 11695, 1998.

Temerin, M., I. Roth, M.K. Hudson, J.R. Wygant, New paradigm for the transport and energization of radiation belt particles, *Eos Trans AGU*, 538, Nov. 1, 1994.

Thorne, R. M., and R.B. Horne, Energy transfer between energetic ring current H^+ and O^+ by electromagnetic ion cyclotron waves, *J. Geophys. Res.*, **99**, 17275–17282, 1994.

Tsurutani, B.T., The interplanetary causes of magnetic storms, substorms and geomagnetic quiet, in *Space Storms and Space Weather Hazards*, edited by I.A. Daglis, pp. 103–130, Kluwer Academic Publishers, Dordrecht, 2001.

Vampola, A.L., Thick dielectric charging on high-altitude spacecraft, *J. Electrostatics*, **20**, 21, 1987.

Van Allen, J.A., G.H. Ludwig, E.C. Ray, and C.E. McIlwain, Observations of high intensity radiation by satellites 1958 Alpha and Gamma, *Jet Propul.*, **28**, 588–592, 1958.

Van Allen, J. A., The geomagnetically trapped corpuscular radiation, *J. Geophys. Res.*, **64**, 1683–1689, 1959.

Vette, J., The AE-8 trapped electron model environment, *National Space Science Data Center, Report 91-24*, Greenbelt, MD, 1991.

Violet, M.D., and A.R. Fredrickson, Spacecraft anomalies on the CRRES satellite correlated with the environment and insulator samples, *IEEE Trans. Nucl. Sci.*, **40**, 1512, 1993.

Weyland, M., and M. Golightly, Results of radiation monitoring on the International Space Station and associated biological risks, *Proceedings of International Space Environment Conference*, Queenstown, New Zealand, July 23–27, 2001.

Wilken B., W. Weiß, D. Hall, M. Grande, F. Søraas, and J.F. Fennell, Magnetospheric Ion Composition Spectrometer onboard the CRRES spacecraft, *J. Spacecraft Rockets*, **29**, 585–591, 1992.

Williams, D.J., Exploration and understanding in space physics, *Geophys. Res. Lett.*, **12**, 303, 1985.

Wrenn, G.L., Conclusive evidence for internal dielectric charging anomalies in geosynchronous communications spacecraft, *J. Spacecraft Rockets*, **32**, 3, 514, 1995.

Young, D.T., H. Balsiger, and J. Geiss, Correlations of magnetospheric ion composition with geomagnetic and solar activity, *J. Geophys. Res.*, **87**, 9077–9096, 1982.

7

Ionospheric response

Kristian Schlegel

7.1 INTRODUCTION

This chapter describes the most important effects of space weather on the ionosphere and atmosphere. It is presupposed that the reader is familiar with the basic physics and terminology of this part of near-Earth space. For reference, a figure is included showing the temperature, ion density and neutral density as a function of altitude (Figure 7.1). Recent reviews of modern ionospheric physics have been published by Kelley (1989), Kohl *et al.* (1990), Prölss (2001), and Hagfors and Schlegel (2001).

The main space weather effects can be summarized in a flow chart, as displayed in Figure 7.2 (see colour section). 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 $\vec{V}$ in the presence of the Earth's magnetic field causes a system of electric fields according to the relation:

$$\vec{E} = -\vec{V} \times \vec{B} \tag{7.1}$$

The enhanced conductivity σ and the electric field together cause strong electric currents, mainly in the auroral zone within the dynamo region (90–150 km altitude), according to Ohm's law:

$$\vec{j} = \sigma \vec{E} \tag{7.2}$$

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 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

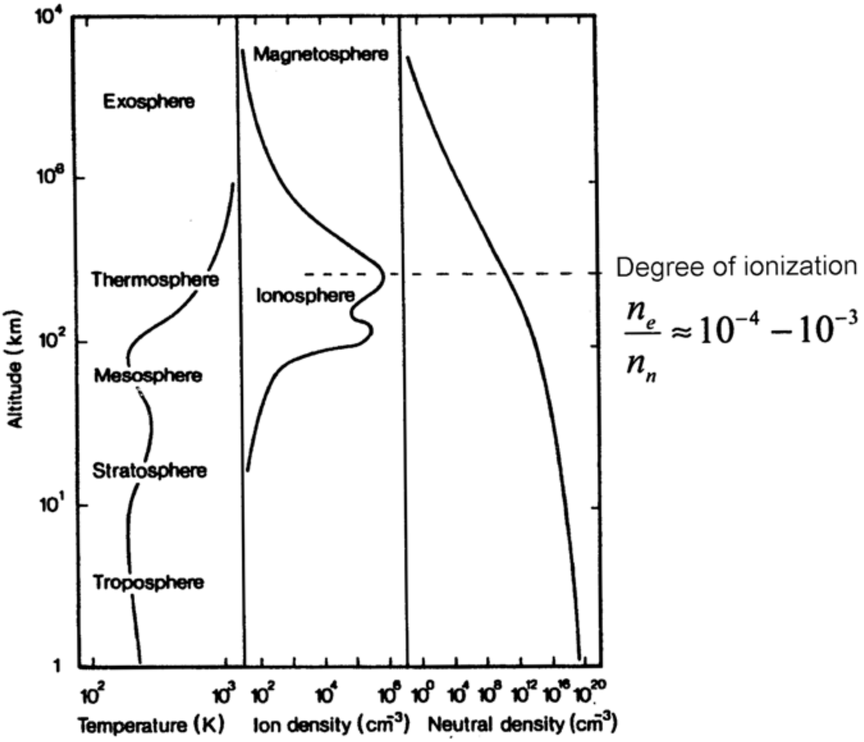


Figure 7.1. Basic quantities of the atmosphere and ionosphere as a function of height.

neutral atmosphere by frictional processes, and influences the global circulation of the neutral gas (Stubbe, 1982).

In the following sections we will describe the above processes in more detail.

7.2 PARTICLE PRECIPITATION

The particles precipitating from the magnetosphere into the ionosphere during geomagnetic storms are mainly electrons with energies of a few keV. They are spiralling around the geomagnetic field lines, and interact with neutrons and ions by collisions. Neutrons are ionized by these collisions and the ion production as a function of particle energy and altitude can be computed. Typical ionization rates as a function of altitude are plotted in Figure 7.3. They are calculated under the assumption of mono-energetic electrons. 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 obtained from, for example, satellite observations. 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 (Kirkwood and Osepian, 2001).

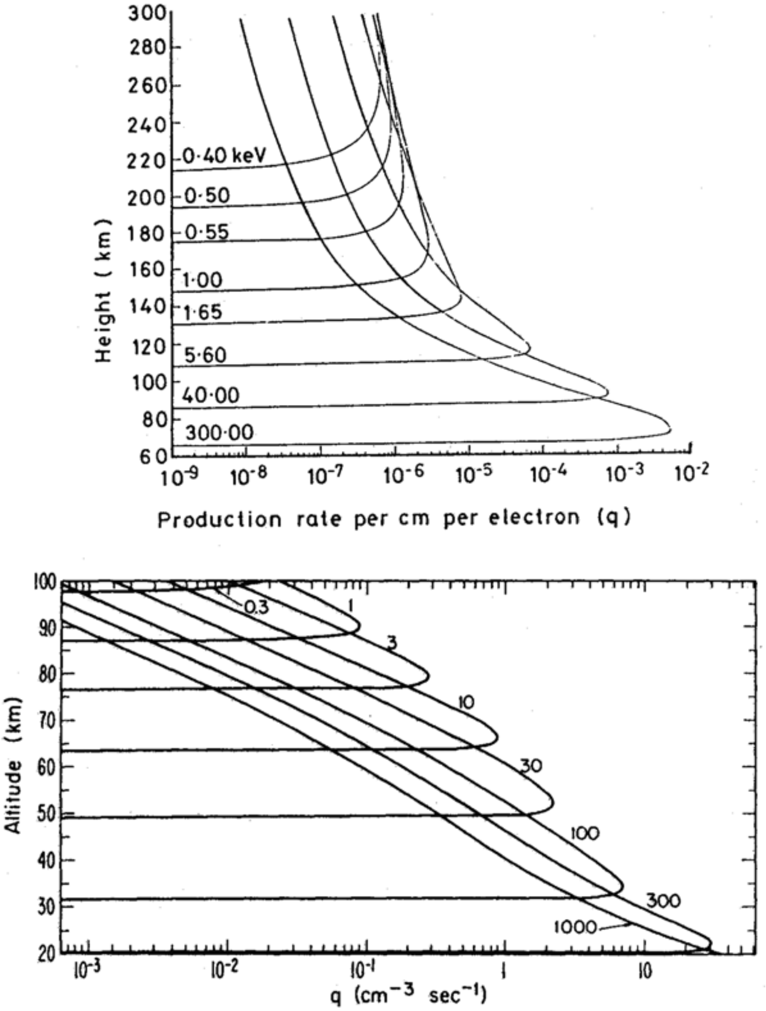


Figure 7.3. Ionization rates caused by precipitation particles: (upper panel), electrons; (lower panel), protons.

It is obvious from Figure 7.3 that a peak electron energy of a few keV causes a maximum of ionization in the lower E region. At high latitudes this ionization can be more than an order of magnitude higher than the 'usual' ionization by solar EUV radiation. The resulting ion density N_i (= electron density N_e, because of charge neutrality) from the ionization by particles Q_{Part} and solar radiation Q_{EUV}, can be calculated from the continuity equation:

$$\frac{\partial N_i}{\partial t} = Q_{\text{EUV}} + Q_{\text{Part}} - L - div(N_i \vec{V}_i) \tag{7.3}$$

Figure 7.4. Example of an electron density enhancement in the auroral *E* region during particle precipitation (solid line) compared to quiet conditions (dashed line).

where L represents the ionization loss processes and the last term expresses transport effects. Figure 7.4 shows an example of electron densities measured in the high latitude E region with the incoherent scatter technique (Alcaydé, 1997). Figures 7.4 to 7.9 show quantities derived from data of the incoherent scatter system EISCAT during a strongly disturbed period on 29–30 October 2003, with Kp-values (see Section 7.4) rising to 9_{o}.

During solar proton events (a consequence of solar flares), high-energy protons (up to more than 100 MeV) can precipitate into the ionosphere. Figure 7.3 also shows production rate profiles for mono-energetic protons. They are able to cause ionization much deeper in the atmosphere than electrons (see also Figure 7.19 in Section 7.7).

7.3 CONDUCTIVITIES AND CURRENTS

Currents in the ionosphere are caused by moving charged particles; thus the current density is given by:

$$\vec{j} = eN_e(\vec{V}_i - \vec{V}_e) \tag{7.4}$$

where e is the elementary charge, and V_i and V_e are ion and electron drifts, respectively. Using the steady-state momentum equations, the current can be expressed as:

$$\vec{j} = \tilde{\sigma}(\vec{E} + \vec{U} \times \vec{B}) \tag{7.5}$$

with the neutral gas velocity U and the conductivity tensor:

$$\tilde{\sigma} = \begin{Bmatrix} \sigma_P & \sigma_H & 0 \\ -\sigma_H & \sigma_P & 0 \\ 0 & 0 & \sigma_\parallel \end{Bmatrix} \tag{7.6}$$

The three components of this tensor are:

$$\text{Pedersen conductivity} \quad \sigma_P = \frac{eN_e}{B}\left(\frac{\omega_e \nu_{en}}{\omega_e^2 + \nu_{en}^2} + \frac{\omega_i \nu_{in}}{\omega_i^2 + \nu_{in}^2}\right)$$

$$\text{Hall conductivity} \quad \sigma_H = \frac{eN_e}{B}\left(\frac{\omega_e^2}{\omega_e^2 + \nu_{en}^2} - \frac{\omega_i^2}{\omega_i^2 + \nu_{in}^2}\right)$$

$$\text{and parallel conductivity} \quad \sigma_\parallel = e^2 N_e\left(\frac{1}{m_e(\nu_{en} + \nu_{ei})} + \frac{1}{m_i \nu_{in}}\right) \tag{7.7}$$

Two important facts can be derived from these equations: 1, 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; 2, 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 (Figure 7.1). The latter is usually taken from a neutral atmospheric model (*http://nssdc.gsfc.nasa.gov/space/model/models/msis_n.html*).

Typical conductivity profiles are displayed in Figure 7.5. Note that σ_H has its peak around 110 km altitude and decreases rapidly with height, whereas σ_P has its peak somewhat higher 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 in the ionosphere because the parallel electric field is negligibly small.

For practical purposes the so-called conductance is often used. It is the conductivity integrated over the height range where it is most important:

$$\Sigma_{P,H} = \int_{90\,\text{km}}^{250\,\text{km}} \sigma_{P,H}(z)\, dz \tag{7.8}$$

It has the advantage that it can be conveniently plotted versus time, thus showing the ionospheric variability during space weather events. Figure 7.6 shows an example which has been computed from measured electron densities (incoherent scatter technique) together with model values of collision frequencies (Schlegel, 1988). During the daytime, 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 early morning and in late afternoon/evening, conductances are high and burst-like, which is characteristic for a geomagnetic storm.

Electric fields in the ionosphere are low – typically a few mV/m – during quiet conditions. Generally this field can be regarded as independent of height within the range of interest (90 km to several 100 km). During strong geomagnetic storms the field can well exceed 100 mV/m. Figure 7.7 shows, as an example the N–S and

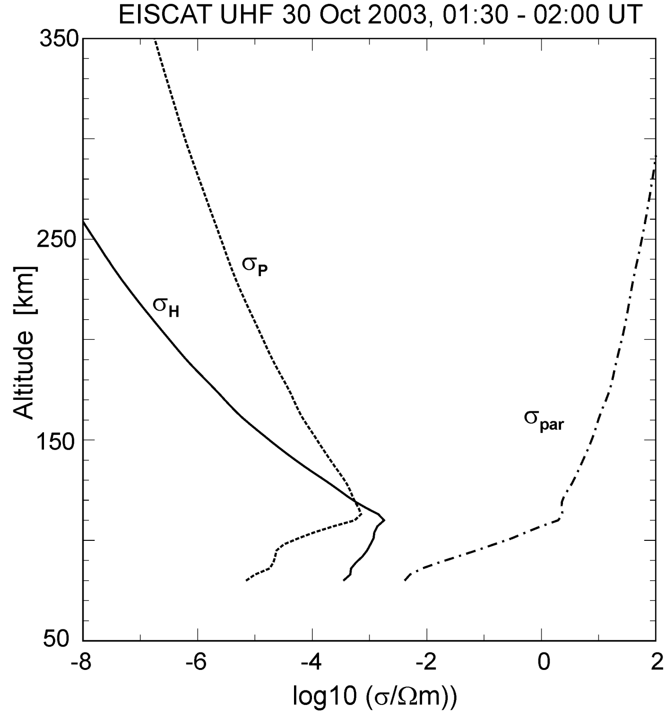


Figure 7.5. Typical ionospheric conductivity profiles during disturbed conditions, derived from EISCAT data.

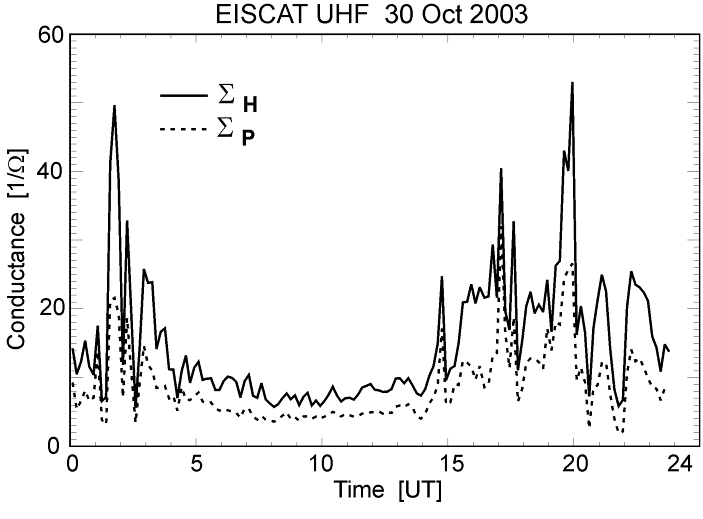


Figure 7.6. Hall and Pedersen conductances during a magnetic storm, computed from EISCAT data.

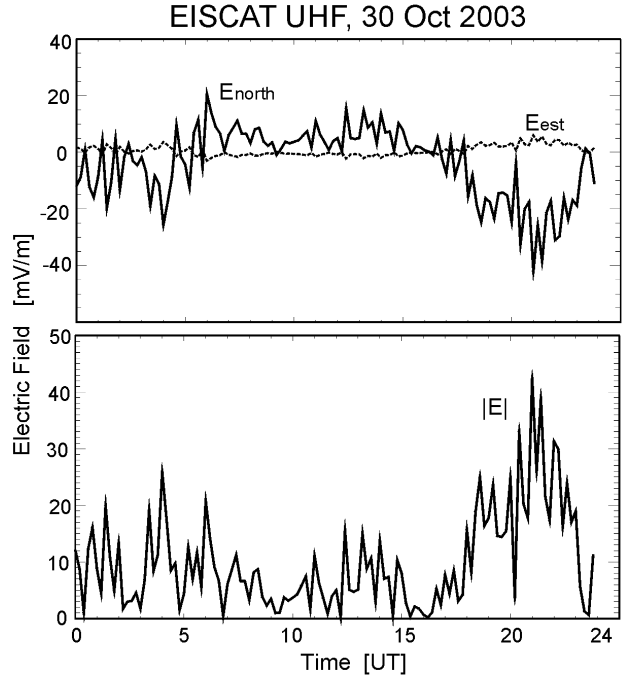


Figure 7.7. Typical electric field variations during a magnetic storm, derived from tristatic EISCAT measurements.

the E–W components of the electric field during the same time interval as in Figure 7.6. Northward electric fields prevail during the day between about 0500 and 1600 UT. The field then turns southward and remains in this direction until early the following morning. This is a very typical electric field behaviour during a magnetic storm. The burst-like enhancements in the morning sector correspond to substorms.

A typical current density profile during disturbed conditions in the morning sector is plotted in Figure 7.8. 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 eastward-directed in the afternoon sector and westward in the morning sector. The north–south extent of the electrojet is typically about 100 km. Therefore, the total current of the electrojet at the time given in Figure 7.8 was of the order of:

$$J \approx 50\,\mu\mathrm{A/m^2} \cdot 30000\,\mathrm{m} \cdot 100000\,\mathrm{m} = 0.15 \cdot 10^6\,\mathrm{A} \tag{7.9}$$

During very strong magnetic storms it can easily exceed 1 million A.

The power (per unit volume) dissipated in the E region (Joule heating) by the electrojet is given by:

$$P_J = \vec{j} \cdot \vec{E} = \sigma_P E^2 \tag{7.10}$$

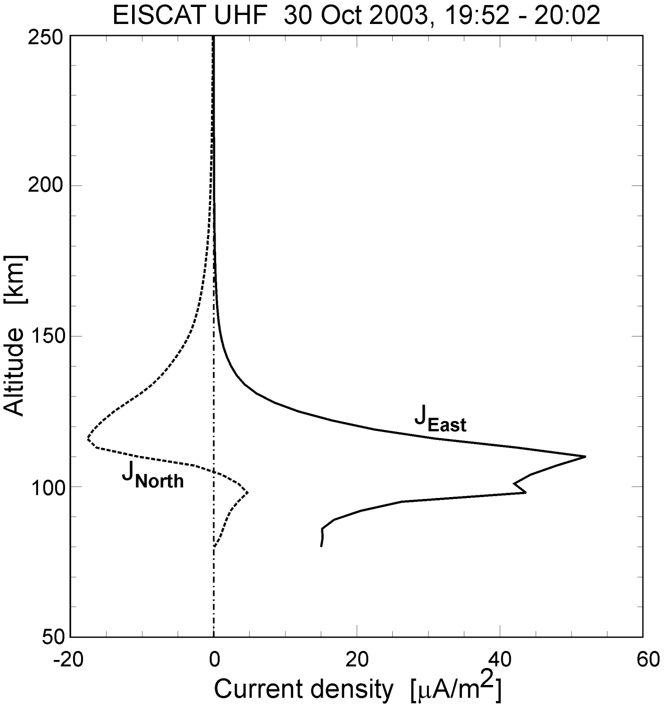


Figure 7.8. Current densities as a function of height, as derived from EISCAT data during particle precipitation and high electric fields.

The product of the expression for the current 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

$$Q_J = \Sigma_P E^2 \tag{7.11}$$

can be again conveniently plotted versus time, to provide an overview of the dissipated energy during a magnetic storm. Figure 7.9 shows an example. In addition to Joule heating there is also an energy deposition due to the precipitating particles which can be estimated with the relation

$$Q_P = 1/2\eta\alpha_{\text{eff}}N_e^2 \tag{7.12}$$

with $\eta = 35\,\text{eV}$ 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 – 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 whole auroral zone

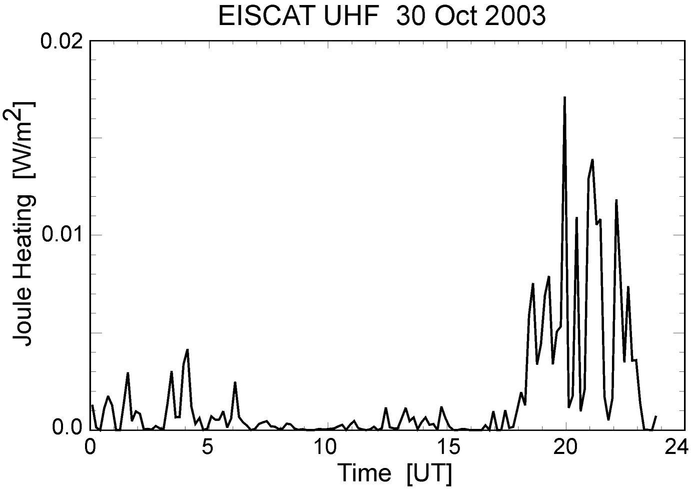


Figure 7.9. Joule heating in the auroral ionosphere during a magnetic storm, derived from EISCAT data.

was estimated to 13,000 TJ, which is about 40% of the average solar wind energy input into the magnetosphere (Schlegel and Collis, 1999). Another detailed estimate of the global energy dissipation during a magnetic storm has been presented by Pulkkinen *et al.* (2002).

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. These satellites normally experience a negligible friction at these altitudes because of the low air-density. During a magnetic storm, however, when the atmosphere expands, the neutral density can rise by more than a factor of ten at the satellite's orbit. This causes considerable air braking, which can disturb the satellite's orientation (with loss of communication, because the antenna may no longer point towards Earth), or in severe cases a deceleration of the spacecraft, which can finally lead to burn-up. In order to avoid this, affected satellites have to be brought back on their nominal orbit with their onboard jet engines, as illustrated in Figure 7.10.

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 changes only marginally 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 ionization is increased, as shown in Figure 7.11. Apart from the consequences for satellite trajectories, this also leads to important differences in shortwave radio propagation during the solar cycle, as

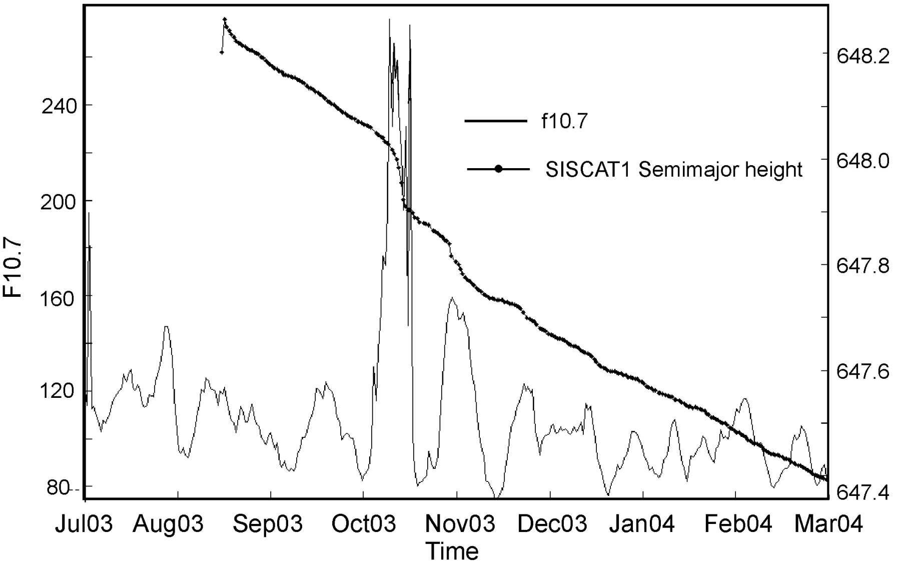


Figure 7.10. Orbit corrections to the Satellite SISCAT after air braking during high solar activity in autumn 2003. (Jean-Marc Noel, private communications.)

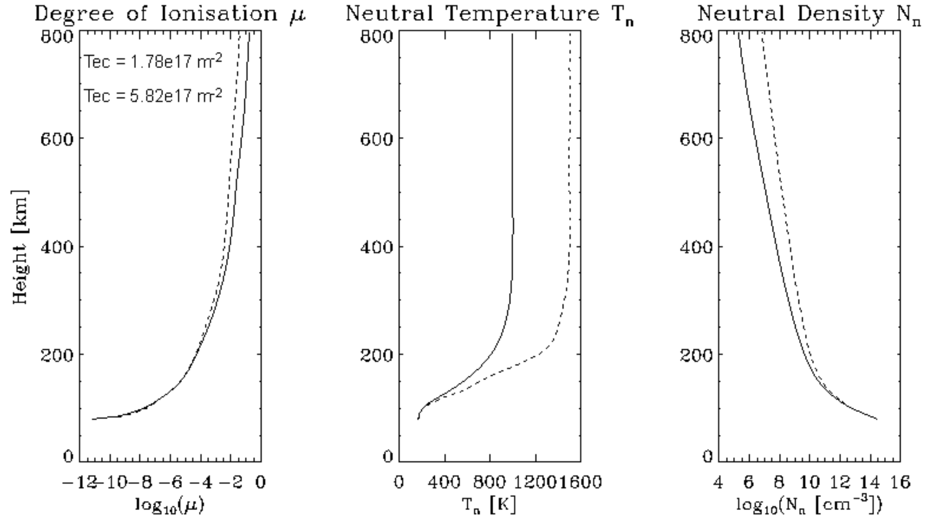


Figure 7.11. Difference between high (dashed) and low (solid) solar activity in atmospheric ionization, temperature and density. Whereas the total electron content increases more than threefold during high solar activity (from $1.78 * 10^{17}\,\mathrm{m}^{-2}$ to $5.82 * 10^{17}\,\mathrm{m}^{-2}$) because the ionosphere extends to greater heights, the degree of ionization $\mu = n_e/n_n$ is lower because the neutral density increases stronger than the electron density.

known for more than 70 years. Even radio amateurs enjoy the greater distances to be covered during solar maximum years.

7.4 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 $\Delta\vec{B}$. According to the 'right-hand rule', the magnetic perturbation appears mainly in the N component of ΔB 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 extent of the electrojet can be well established.

Ground-based magnetometers are similarly important 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 characterized by the index Kp (Chapman and Bartels, 1962). 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 thirteen 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 –, o, + (e.g 1_-, 1_o, 1_+, 2_-, ...), yielding 28 steps in total. Bartels also developed a convenient representation of Kp in terms of musical notes (Figure 7.12).

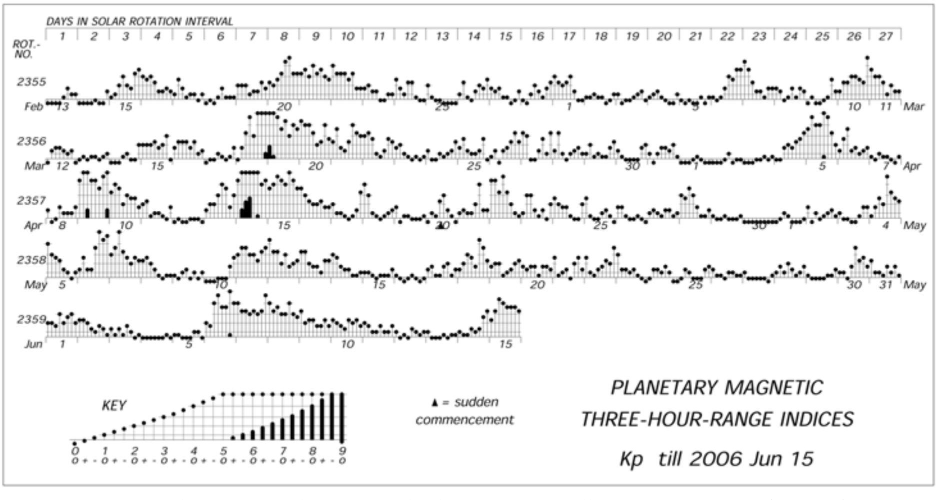


Figure 7.12. Bartels' Kp notation as musical notes. Every line represents 28 days – the average solar rotation. Recurrency trends can therefore be easily detected.

As already mentioned, Kp is expressed in a logarithmic scale and is consequently not well suited for averaging. Bartels therefore introduced the linear equivalent ap, where $Kp = 9_o$ corresponds to ap = 400 nT. The Ap index is a mean over eight 3-hour intervals of ap (over a full day). It consequently characterizes 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, but for earlier years not enough magnetometer stations were available. For many years the University of Göttingen issued the Kp and Ap indices, but since the beginning of 1997 this task has been taken over by the Adolf-Schmidt Observatorium für Geomagnetismus 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 AA index, was developed. It is similar to Ap, but is derived from the magnetograms of only two stations – one in the northern hemisphere (England), and one in 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 scientists have pushed the AA records even further back (Nevanlinna and Kataja, 1993).

The indices mentioned so far characterize geomagnetic variations, particularly at high and mid-latitudes, which are mainly related with the auroral electrojet. The magnetic variations due to the ring current (see Chapter 6) are described by the Dst index. Since 1957 it has been derived from the horizontal magnetic field component measured at four stations near the equator. The magnetic field of the ring current is directed opposite to the main geomagnetic field; consequently, strong disturbances are characterized by large negative Dst excursions.

Finally, magnetic disturbances at very high latitudes are characterized by the AE index which is derived from magnetic records of twelve stations at auroral latitudes. Details of the derivation of all indices can be found in Mayaud (1980), and their values are accessible on the Internet (*http://www.cetp.ipsl.fr/~isgi/lesdonne.htm*; *http://spidr.ngdc.noaa.gov/spidr/index.html*). A map with the stations for the various indices is shown in Figure 7.13. The convenience of geomagnetic indices is demonstrated in Table 7.1, listing the ten strongest storms, in terms of AA, since 1886.

7.5 AURORAE

Aurorae are 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 over a broad spectrum (infrared, visible and 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

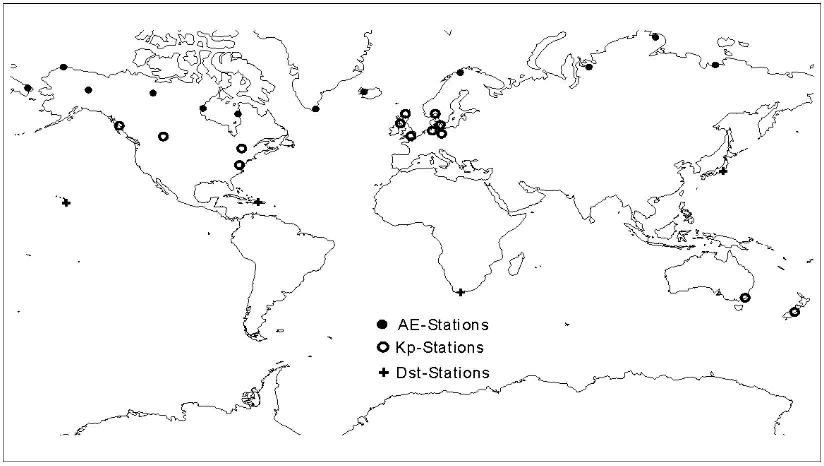


Figure 7.13. Map of the location of magnetometer stations from which the various magnetic indices are derived.

Table 7.1. The ten strongest geomagnetic storms since 1886, in descending order of strength. The second column gives $AA^* = \Sigma_{(3h)} aa/8$ 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 (since 1957) 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 (georgr. latitude)
1989 Mar 13–14	441	9_o	−589	A: Florida Keys ($\varphi \approx 24^\circ$ N)
1941 Sep 18–19	429	9_-	–	Si: Florida ($\varphi \approx 29^\circ$ N)
1940 Mar 24–25	377	9_o	–	Si: Korfu ($\varphi = 39^\circ$ N)
1960 Nov 12–13	372	9_o	−339	A: Atlantic ($\varphi = 28^\circ$ N)
1959 Jul 15–16	357	9_o	−429	Sch: $\varphi \approx 48^\circ$ N
1921 May 14–15	356	–	–	Si: Samoa ($\varphi = 14^\circ$ S)
1909 Sep 25–26	333	–	–	Si: Mallorca ($\varphi = 39^\circ$ N)
2003 Oct 29–30	332	9_o	−363	A: Florida
1946 Mar 28–29	329	9_o	–	Si: Queensland ($\varphi \approx 27^\circ$ S)
1928 Jul 7–8	325	–	–	Si: Atlantic ($\varphi = 24^\circ$ N)

induced or affected by these particles. Figure 7.14 shows a simplified spectrum of auroral emissions in the visible range.

The predominant green colour of aurorae is caused by the following reactions:

$$N^+ + O_2 \rightarrow NO^+ + O(^1S)$$

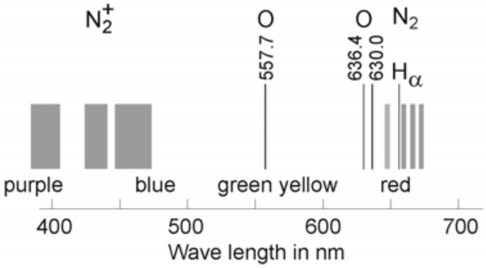


Figure 7.14. Simplified spectrum of auroral emissions in the visible range.

The excited oxygen atom then transits in a lower excitation state by emitting a photon:

$$O(^1S) \rightarrow O(^1D) + \text{Photon (555.7 nm)}$$

Red light is emitted when this metastable state goes to the ground state:

$$O(^1D) \rightarrow O(^3P) + \text{Photon (630 nm)}$$

The $O(^1D)$ state can also be directly excited by dissociative recombination:

$$O_2^+ + e \rightarrow O + O(^1D)$$

Nitrogen molecules can be ionized and excited by primary electrons:

$$N_2 + e_{\text{primary}} \rightarrow N_2^+(B^2\Sigma_u^+) + e_{\text{primary}} + e_{\text{secondary}}$$

This excited state of the ionized 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 (Rees, 1989, for details).

The green line and the red line (the latter is actually a doublet) are so-called ‘forbidden lines’. The corresponding excitation states have a relatively long lifetime of 1s (green) and 110s (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 with atomic oxygen was therefore a longstanding problem for spectroscopists, and was not finally solved until 1932.

The brightness of aurorae is characterized by the International Brightness Coefficient (IBC) (Table 7.2), according to four classes (1 Rayleigh = 10^6 photons/cm^2/s/sterad).

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’ (Figure 7.15, see colour section). 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, aurorae occur most frequently. During

Table 7.2. The International Brightness Coefficient specifying auroral intensity.

IBC	Intensity of the 557.7 nm line	Comparable brightness
I	1 kR (kilo-Rayleigh)	Milky Way
II	10 kR	Moonlight on thin cirrus
III	100 kR	Moonlight on cumulus
IV	1000 kR	Full Moon

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 aurorae therefore appear mainly at altitudes above 200 km, as a red glow. These red colours were associated with blood by our ancestors, and aurorae were therefore regarded as a bad omen of war and diseases (Schlegel, 2001).

The forms of aurorae depend on the topology of the currents flowing from the magnetotail into the polar regions. In principle, two basic manifestations exist: diffuse aurora (unstructured and extended) and discrete aurora (arcs, veils, bands). Some common forms are shown in Figure 7.16. 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), and is therefore considerably lower than the mean energy of the particles in the

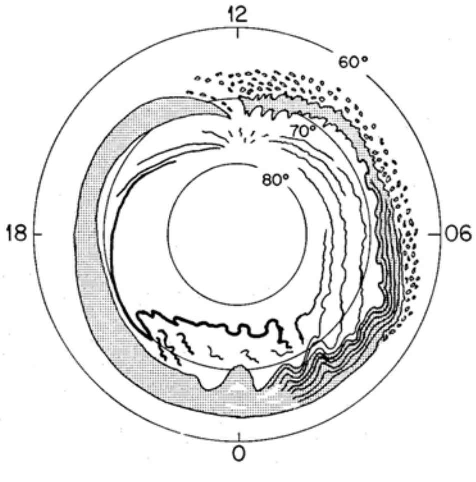


Figure 7.16. Schematic representation of the main auroral forms as a function of local time for latitudes >60°. The shaded area characterizes diffuse aurora, and the thick line a quiet arc which transforms into folded bands after about 2100 LT. In the morning hours, patchy aurora can often be found at the southern rim of the auroral oval. The short thin lines around local noon at about 75° latitude are daylight aurorae. (Akasofu, 1976.)

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, and several different mechanisms have been discussed (for example, Schlegel, 1991).

Particles causing the aurorae mentioned so far are electrons. The 'proton aurora' is much more rare, and is caused by energetic protons which are decelerated in the atmosphere by collisions and finally transformed to excited neutral hydrogen by charge transfer:

$$M + H \rightarrow M^{+} + H^{*}$$

where M is any neutral constituent. The excited hydrogen atoms emit $L\alpha$ (121.57 nm, UV) or $H\alpha$ (656.3 nm, red). The latter cannot be distinguished by eye from the red oxygen light (Figure 7.14). Proton aurorae are generally diffuse, and are often associated with PCA events (Section 7.5).

Figure 7.17 (see colour section), illustrates several types of aurora. In addition there are many Internet pages with splendid photographs (for example, *http://www.meteoros. de*; *http://www.exploratorium.edu/learning_studio/auroras/*; *http://www-pi.physics. uiowa.edu/vis/*).

7.6 CONSEQUENCES OF ELECTRON DENSITY ENHANCEMENTS AND FLUCTUATIONS

The enhancement of electron density by precipitating particles (as described in Section 7.2) has important consequences for 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 (see more details in Chapter 9).

The propagation of electromagnetic waves in the ionosphere is described by the magneto-ionic theory (Rawer, 1993). One important equation is the index of refraction of the waves, which in its simplest form is represented as:

$$n^2 = 1 - \frac{\omega_P^2}{\omega^2} \quad \text{with the plasma frequency} \quad \omega_P = \sqrt{\frac{N_e e^2}{\varepsilon_o m_e}} \tag{7.13}$$

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.

Equation 7.13 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 (see more details in Chapter 13).

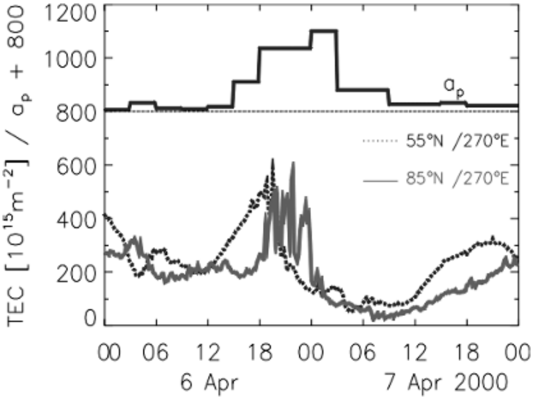


Figure 7.18. Variation of TEC during a magnetic storm. The upper panel shows ap. (Jakowski *et al.*, 2002.)

A very important ionospheric quantity in this context is the total electron content,

$$\mathrm{TEC} = \int_P N_e ds \tag{7.14}$$

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 7.18 shows an example. In the case of GPS measured distances, the TEC change translates into errors of

$$D(\mathrm{mm}) = 2.16 \text{ TEC (in TECU)} \tag{7.15}$$

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 (modified two-stream instability and gradient drift instability) which lead to a structuring of the normally uniform plasma (Schlegel, 1996). The formed plasma irregularities have a broad range of scale lengths λ_{irr}, from kilometres to tenth of metres, 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

$$\lambda_{\text{radio wave}} = 2\lambda_{\mathrm{irr}} \tag{7.16}$$

A corresponding effect which is often observed during space weather events is the over-range of VHF signals; for example, taxi drivers in Hamburg can listen to their colleagues in Helsinki over their usual communication channels. Radio amateurs also utilize this ‘auroral scatter’, as they termed it, for long-range communication. The effect is not limited to high latitudes. It occurs also in the equatorial ionosphere, together with other plasma instabilities.

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).

7.7 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 of 18 August 1979, for instance, the X-ray flux in the wavelength range 0.029–0.048 nm increased by a factor of 2,000, 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 ionization 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). A similar ionization increase is due to the high-energy protons which can penetrate well down into the stratosphere (see Figure 7.3). Figure 7.19 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 minutes after the flare onset, the energetic protons need a travel time of the order of 1 hour. The X-ray flux increase is peak-like with a duration of only about 10 minutes, whereas the enhanced proton flux usually persists for several days, and the large electron densities in the mesosphere and stratosphere persist 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). Short (MHz) and medium (kHz) wave communication is therefore strongly affected in such cases.

In the vicinity of the Earth, the energetic protons gyrate around the geomagnetic

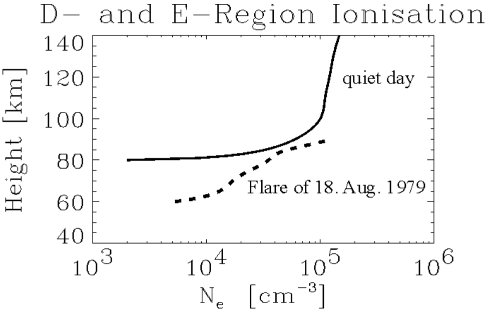


Figure 7.19. Electron density during quiet conditions and during the solar-flare of 18 August 1979 (after Zinn *et al.*, 1988).

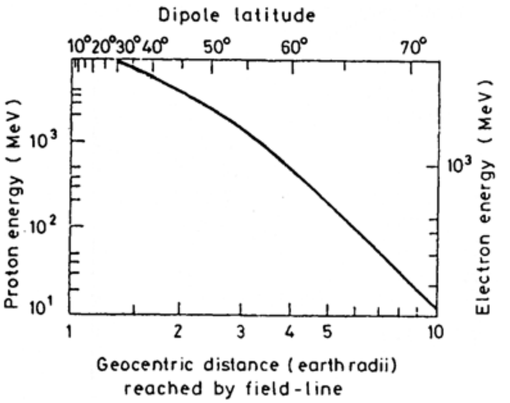


Figure 7.20. Penetration of energetic protons into the atmosphere.

field lines according to Störmer's theory (Walt, 1994). An important quantity for their propagation is the 'magnetic rigidity',

$$R = \frac{pc}{Ze} \tag{7.17}$$

where p is the particle momentum, c is the velocity of light, and Z is their charge number. (This equation also applies to particle with $Z > 1$, such as alpha particles.) All particles with the same rigidity have the same orbital parameters. It can be shown that all particles with a critical rigidity

$$R_c = 14.9 \cos^4 \lambda_c \tag{7.18}$$

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 Figure 7.20: protons with energies $E_p < 100\,\text{MeV}$ will penetrate the Earth's atmosphere only at high latitudes; the higher their energy, the lower latitudes they can reach. Since the peak of solar protons is normally below 100 MeV, the D-region ionization caused by them is usually strongest over the polar caps and consequently the above-mentioned radio wave damping. Such short-wave absorption events were reported in the 1930s, well before their true nature was recognized, and were called 'polar cap absorption events' (PCAs) – a term which is still in use in space weather investigations.

Besides this mainly 'technological' consequence of flares there is another very important climatological result. Through a complicated chain of chemical reactions, the enhanced proton flux causes a strong increase in atmospheric nitric oxide, which in turn destroys ozone. A considerable reduction of the total ozone content in the mesosphere and stratosphere has therefore been observed (Figure 7.21). Since ozone is a very important climate agent, frequent flares may well contribute to climate effects (Kallenrode, 2003).

All these consequences of the ionization of energetic solar protons also apply, in principle, to non-solar energetic particles – galactic cosmic rays (GCR). As explained in previous chapters, the flux of GCRs is anti-correlated with solar activity, and the

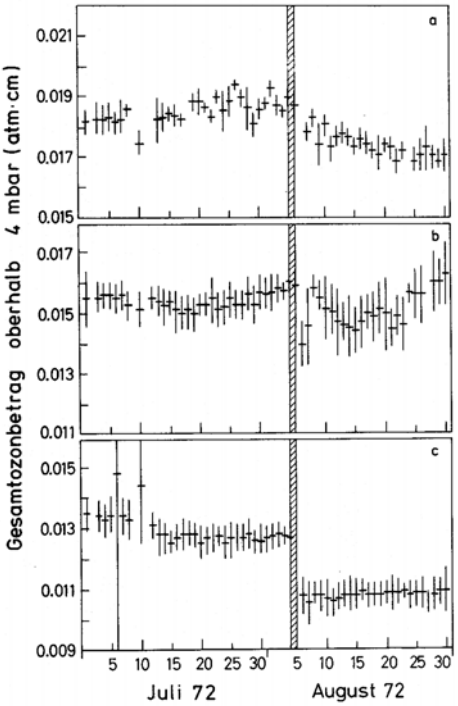


Figure 7.21. Total ozone content above 35 km for equatorial (a), mid-latitudes (b), and high latitudes (c) after the flare of 4 August 1972. (Heath *et al.*, 1977.)

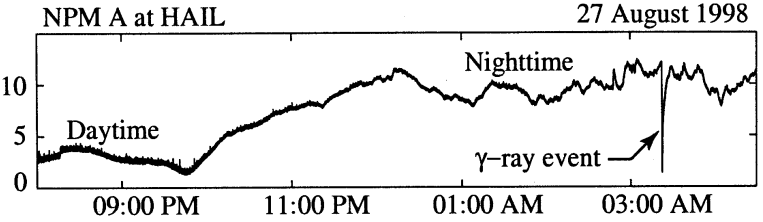


Figure 7.22. D-region electron density increase (shown here as spike in VLF wave propagation) due to an X-ray burst of the neutron star SGR1900 + 14.

ionization of the mesosphere and stratosphere is therefore generally higher during solar minimum years. This probably has a climatological impact.

Finally, it should be noted that not only the Sun, but also stars can be a cause of space weather effects in the ionosphere. During cosmic catastrophes, such as nova or supernova explosions, huge intensities of X-rays and gamma-rays are released. Such an event was registered on 28 August 1998, as a consequence of an X-ray burst of a neutron star. The D region experienced a brief spike of ionization (Figure 7.22), despite the source being 23,000 light-years from Earth. If such an event were to occur 'close' (within 50 light-years), the terrestrial ozone layer might be destroyed for several years, as model calculations show (Ruderman, 1974). It is obvious that this would have drastic consequences for the biosphere!

7.8 REFERENCES

Akasofu, S.-I., Recent progress in studies of DMSP auroral photographs, *Space Sci. Rev.*, **19**, 169, 1970.

Alcaydé, D., Incoherent scatter: theory, practice and science, Technical Rpt. 97/53, EISCAT Scienific Assoc., Kiruna, Sweden, 1997.

Basu, S. and K.M. Groves, Specification and forecasting of outages on satellite communication and navigation systems, in *Space Weather*, P. Song, H.J. Singer and G.L. Siscoe (eds), American Geophysical Union, Washington, D.C., 2001.

Chapman, S. and J. Bartels, *Geomagnetism*, Clarendon Press, Oxford, 1962.

Collis, P.N. and M.T. Rietveld, Mesospheric observations with the EISCAT UHF radar during polar cap absorption events, *Ann. Geophys.*, **8**, 809–824, 1990.

Jakowski, N., S. Heise, A. Wehrenpfennig, S. Schlüter, and R. Reimer, GPS/GLONASS-based TEC measurements as a contributor for space weather forecast, *J. Atmos. Sol.–Terr. Phys.*, **64**, Issue 5–6, 729–735, 2002.

Hagfors, T. and K. Schlegel, Earth's Ionosphere, in *The Century of Space Science*, Kluwer, Dordrecht, 2001.

Heath, D.F., A.J. Krueger, and P.J. Crutzen, Solar proton event: influence on stratospheric ozone, *Science*, **197**, 886, 1977.

Kallenrode, M.-B., Current views on impulsive and gradual solar energetic particle events, *J. Phys. G.*, **29**, 965–981, 2003.

Kelley, M.C., *The Earth's Ionosphere*, Academic Press, San Diego, 1989.

Kirkwood, S. and A. Osepian, Quantitative description of electron precipitation during auroral absorbtion events in the morning/noon local-time sector. *J. Atmos. Sol.–Terr. Phys.*, **63**, 1907–1922, 2001.

Kohl, H., R. Rüster and K. Schlegel (eds), *Modern Ionospheric Science*, European Geophysical Society, Katlenburg-Lindau, Germany, 1996.

Mayaud, P.N., Derivation, meaning and use of geomagnetic indices, American Geophysical Union, Washington D.C., 1980.

Nevanlinna, H. and E. Kataja, An extension of the geomagnetic activity index series aa for two solar cycles (1844–1868), *Geophys. Res. Lett.*, **20**, 2703–2706, 1993.

Prölss, G.W. *Physik des erdnahen Weltraums*, Springer, Berlin, 2001.

Pulkkinen, T.I. *et al.*, Energy dissipation during a geomagnetic storm, May 1998, *Adv. Space Res.*, **30**, 10, 2231–2240, 2002.

Rawer, K., *Wave Propagation in the Ionosphere*, Kluwer, Dordrecht, 1993.

Rees, M.H., *Physics and Chemistry of the Upper Atmosphere*, Cambridge University Press, 1989.

Ruderman, M.A., Possible consequences of nearby supernova explosions for the atmospheric ozone and terrestrial life, *Science*, **184**, 1079–1081, 1974.

Schlegel, K., *Vom Regenbogen zum Polarlicht: Leuchterscheinungen in der Atmospäre*, Spektrum Akad. Verlag, Heidelberg, 2001.

Schlegel, K., Coherent backscatter from ionospheric E-region plasma irregularities, *J. Atmos. Terr. Phys.*, **58**, 933–941, 1996.

Schlegel, K., Polarlichter, in *Plasmaphysik im Sonnensystem*, K.-H. Glassmeier and M. Scholer (eds), BI-Wissenschaftsverlag, Mannheim, 1991.

Schlegel, K., Auroral zone E-region conductivities during solar minimum derived from EISCAT data, *Ann. Geophys.*, **6**, 129–138, 1988.

Schlegel, K. and P.N. Collis, The storm of 10 Jan 1997: electrodynamics of the high latitude E region from EISCAT data, *J. Atmos. Solar.–Terr. Phys.*, **61**, 217–222, 1999.

Stubbe, P., Interaction of neutral and plasma motions in the ionosphere, in *Handbuch der Physik* (S. Flügge, ed.), Geophysik III, Part VI, p. 247, Berlin, 1982.

Walt, M., *Introduction to Geomagnetically Trapped Radiation*, Cambridge University Press, 1994.

8

Solar effects in the middle and lower stratosphere and probable associations with the troposphere

Karin Labitzke and Harry van Loon

8.1 INTRODUCTION

Nearly all the Earth's energy derives from the Sun, and it is therefore natural to look for links between variations in the Sun's irradiance and changes in the atmosphere and oceans. One of the first attempts to measure the total solar radiation (the solar constant) for this purpose was made by C. G. Abbot (1913), and despite many difficulties he succeeded in obtaining a mean value of the solar constant for the period 1902–1912. Figure 8.1 (from Abbot *et al.*, 1913) shows that there is also a change in the radiation from maximum to minimum in the decadal oscillation (the 11-year sunspot cycle (SSC)), in the sense that less radiation was emitted during solar minimum than during the maximum of the oscillation, despite the greater number of spots on the Sun during maximum.

In 1978 the first satellite observations of total solar radiation began (Figure 8.2, see colour section) (Fröhlich, 2000). Qualitatively, the satellite observations confirm Abbot's result that the values are higher during solar maxima, but the variation from maxima to minima within the 11-year solar cycle is appreciably smaller (0.1% difference between the extremes) in the satellite data (Fröhlich and Lean, 1998; Fröhlich, 2004).

The satellite observations of the total solar irradiance included the variability of the ultraviolet radiation. The variability of this quantity alone is considerably larger than that of the total solar radiation: 6–8% in those wavelengths in the ultraviolet (200–300 nm) that are important in the production of ozone and middle atmosphere heating (Chandra and McPeters, 1994; Haigh, 1994; Hood, 2003, 2004; Lean *et al.*, 1997).

Until recently it was generally doubted that the solar variability measured by satellites and in the decadal oscillation has a significant or even measurable influence on weather and climate variations (see, for example, Pittock's review, 1983, and

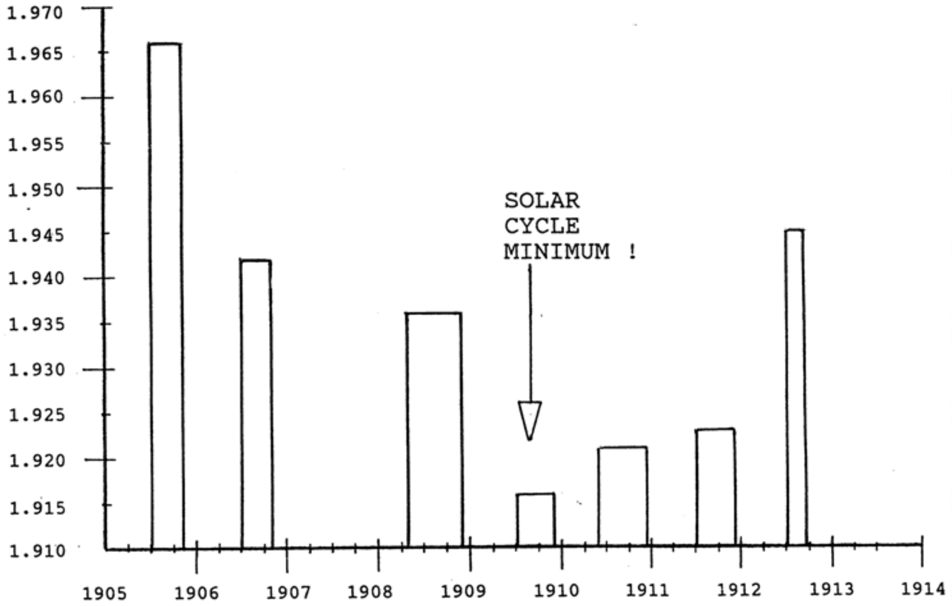


Figure 8.1. Values for the solar constant of radiation (calories) for the period 1905–1912 (Abbot *et al.*, 1913).

Hoyt and Schatten, 1997). But several studies, both empirical and modelling, have in recent years pointed to probable and certain influences. For instance, Labitzke (1982) suggested that the Sun influences the intensity of the north polar vortex in the stratosphere in winter, and that the Quasi-Biennial Oscillation (QBO, see below) modulates the solar signal (Labitzke, 1987; Labitzke and van Loon, 2000; Salby and Callaghan, 2000; Ruzmaikin and Feynman, 2002).

On time scales longer than the decadal scale Reid (2000) and Friis-Christensen and Lassen (1991, 1993), among others, have shown that long-term solar variability may be responsible for an appreciable component of the trends in global surface air temperature (see also discussion by Damon and Laut, 2004).

At present there is no agreement about the mechanism or mechanisms through which the solar effect is transmitted to the atmosphere. For example, Svensmark (1998) and Svensmark and Friis-Christensen (1997) have suggested that galactic cosmic rays (GCR) influence cloud formation and thus the Earth's radiation budget. However, Udelhofen and Cess (2001) showed by means of a 90-year record of cloud cover over the United States that the cloud variations are in phase with the solar variation, and not out of phase as suggested by Svensmark and Friis-Christensen. Kristjánson *et al.* (2004) found that when globally averaged low-cloud cover (18 years of satellite data) is considered, consistently higher correlations are found between low cloud variations and solar irradiance variations than between variations in cosmic-ray flux and low-cloud cover.

It is our approach to consider the increased UV radiation in the upper stratosphere during solar maxima as the main driver through which ozone is increased in the upper stratosphere which leads to warming and changes in the dynamics by changing the wind and the propagation of the planetary waves (see Section 8.5).

In Section 8.2 we explain which data and methods were used, and in Section 8.3 we describe the overall variability in the stratosphere and troposphere against which the influence of solar variability must be measured. Then follows, in Section 8.4, a summary of our diagnostic studies of the solar effect as observed during the past 50 years, supplemented by results of similar studies by others.

Finally, in Section 8.5, we discuss proposed mechanisms and modelling experiments.

8.2 DATA AND METHODS

The NCEP/NCAR reanalyses (Kalnay *et al.*, 1996) are used for the period 1968–2004 (except for Figures 7, 8, 14, and 15, where the data start in 1958). The reanalyses are less reliable for earlier periods, mainly because of the lack of radiosonde stations over the southern hemisphere, the lack of high-reaching balloons in the early years, and the scarce satellite information before 1979. However, we note that the inclusion of the early data nevertheless yields similar results.

The monthly mean values of the 10.7-cm solar flux are used as a proxy for the 11-year SSC. The flux values are expressed in solar flux units: $1\,\text{s.f.u.} = 10^{-22}\,\text{W}\,\text{m}^{-2}\,\text{Hz}^{-1}$. This is an objectively measured radio wave, highly and positively correlated with the SSC and particularly with the UV part of the solar spectrum (Hood, 2003).

For the range of the SSC, the mean difference of the 10.7-cm solar flux between solar minima (about 70 units) and solar maxima (about 200 units) is used – 130 units. Any linear correlation can be represented also by a regression line with $y = a + bx$, where x in this case is the 10.7-cm solar flux, and b is the slope. This slope is used here, multiplied by 130, in order to obtain the differences between solar minima and maxima, as presented in Sections 8.4.1 and 8.4.2 (Labitzke, 2003).

It is difficult to determine the statistical significance of the correlations, because we have less than four solar cycles and the degrees of freedom are therefore limited. However, using the same data, Ruzmaikin and Feynman (2002), as well as Salby and Callaghan (2004), found a high statistical significance of results similar to ours.

The QBO is an oscillation in the atmosphere which is best observed in the stratospheric winds above the equator, where the zonal winds change between east and west with time (Figure 8.3). The period of the QBO varies in space and time, with an average value near 28 months at all levels (see reviews by Naujokat, 1986, and Baldwin *et al.*, 2001).

Because the QBO modulates the solar signal, and in turn is modulated by the Sun, it is necessary to stratify the data into years for which the equatorial QBO in the lower stratosphere (at about 45 hPa (Holton and Tan, 1980)) was in its westerly or easterly phase (QBO data set in Labitzke and collaborators, 2002).

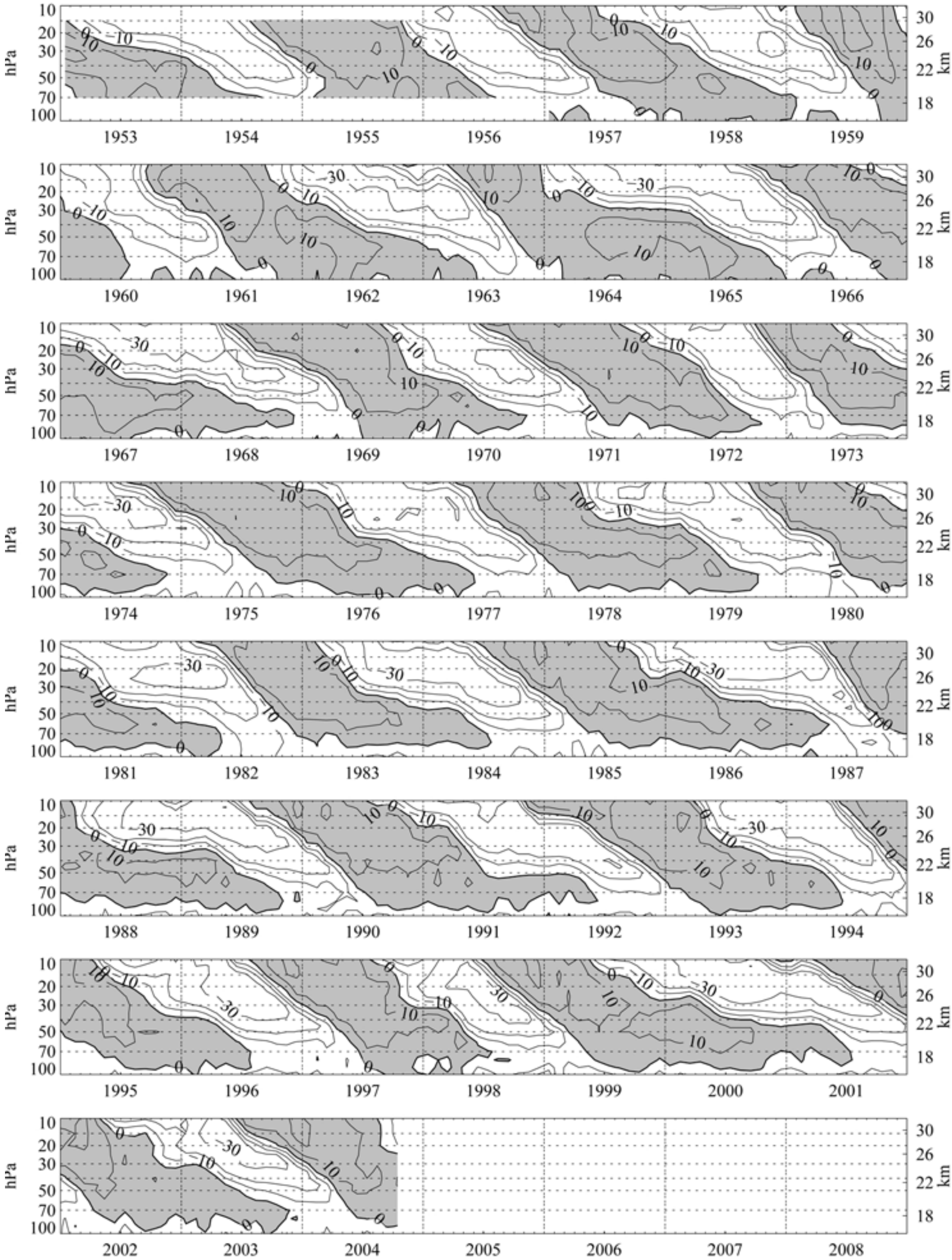


Figure 8.3. Time-height section between 100 and 10 hPa (16 and 32 km) of monthly mean zonal winds (m/s) at equatorial stations: Canton Island, 3°S/172°W (Jan 1953–Aug 1967); Gan/Maledive Islands, 1°S/73°E (Sep 1967–Dec 1975); and Singapore, 1°N/104°E (since Jan 1976). Isopleths are at 10 m/s intervals; westerlies are shaded. (Updated from Naujokat, 1986.)

8.3 VARIABILITY IN THE STRATOSPHERE

The stratosphere in the northern hemisphere reaches its highest variability in winter (Figure 8.4). It is remarkable that the standard deviations in the Arctic winter are three to four times larger than those in the Antarctic winter; however, when the Antarctic westerly vortex breaks down in spring (September–November) this process varies so much from one spring to another that the standard deviation at the south pole in October (Figure 8.4) approaches that at the north pole in January and February (Labitzke and van Loon, 1999).

In the respective summers, the variability is low in both hemispheres, below 1 K. A minor maximum is observed on the equator due to the QBO.

In Figure 8.5, 30-hPa temperatures at the north pole in January since 1956 show how the large standard deviation comes about: out of the 49 years, 14 years are well above the rest. In most of these years, Major Mid-Winter Warmings occurred (Labitzke and van Loon, 1999), and these warmings are associated with a breakdown of the cold westerly vortex.

The state of the Arctic westerly vortex in northern winter is influenced by several factors (van Loon and Labitzke, 1993) (Table 8.1):

- The QBO, Figure 8.3, consists of downward propagating west and east winds, with an average period of about 28 months, and is centred on the equator. An

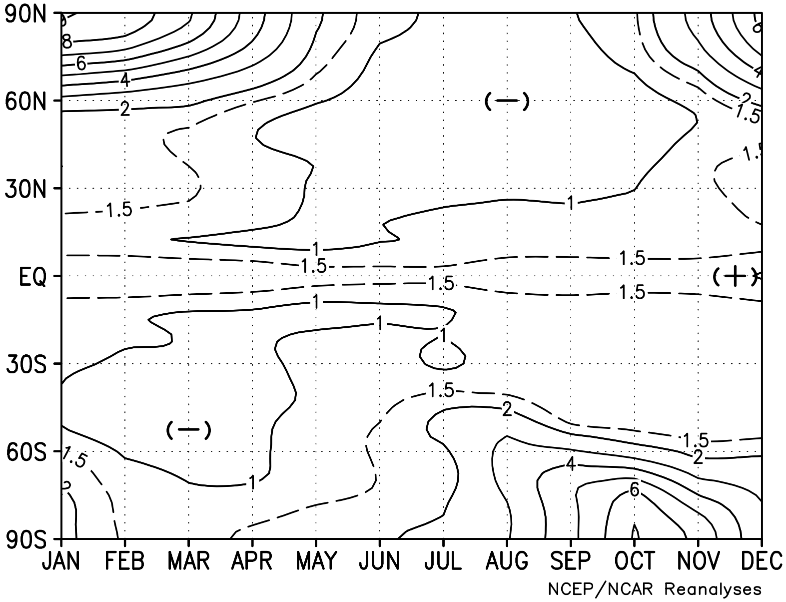


Figure 8.4. The global distribution of standard deviations (K) of the zonal mean 30-hPa monthly mean temperatures for the period 1968–2002 (NCEP/NCAR reanalyses). (Labitzke and van Loon, 1999, update of Figure 2.11.)

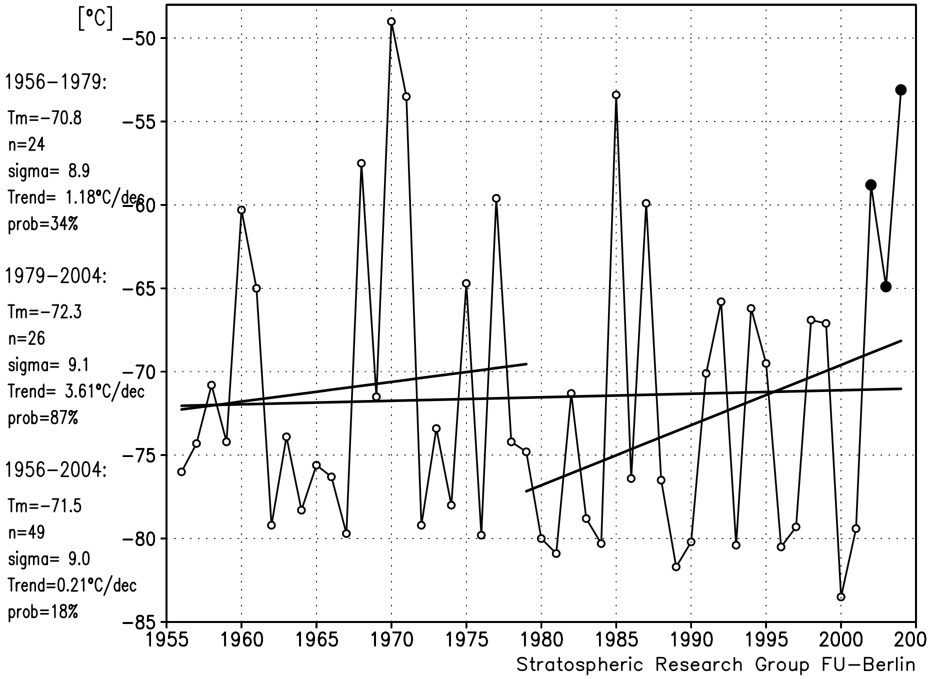


Figure 8.5. Time series of monthly mean 30-hPa temperatures (°C) at the north pole in January 1956–2004. Linear trends have been computed for three different periods. (Data: Meteorological Institute, Free University Berlin until 2001, then ECMWF.) (Labitzke and collaborators, 2002, updated.)

Table 8.1. Different forcings influencing the stratospheric circulation during the northern winters: the Quasi-Biennial Oscillation (QBO), the Southern Oscillation (SO), and the 11-year solar cycle, as well as volcanoes in the tropics.

Arctic	Polar	Vortex
QBO	west phase	cold and strong
	east phase	warm and weak
SO	cold event	cold and strong
	warm event	warm and weak
Solar cycle	solar min.	like QBO
	solar max.	opposite to QBO
Volcanoes		cold and strong

historical review and the present explanation of the QBO can be found in Labitzke and van Loon (1999). The QBO modulates the Arctic and also the Antarctic polar vortex (Labitzke, 2004b) (Table 8.1), but this modulation changes sign depending on the solar cycle (see Sections 4.1 and 4.2).

- Another quantity whose effect is felt in the stratosphere is the Southern Oscillation (SO). The SO is defined as a see-saw in atmospheric mass (sea-level pressure) between the Pacific Ocean and the Australian–Indian region, (see, for example, Labitzke and van Loon, 1999). Its influence is global, and, as shown in the following text, it reaches into the stratosphere. The anomalies in the lower stratosphere associated with extremes of the SO are described in van Loon and Labitzke (1987), where they are discussed in terms of other influences such as the QBO and volcanic eruptions. Figure 8.6 (see colour section) shows the temperature and geopotential height anomalies in four warm and four cold extremes of the Southern Oscillation. The years chosen are years in solar minima: that is, years when the Sun supposedly does not disturb the atmosphere (see Section 8.4) (Labitzke and van Loon, 1999, p. 86). In the warm extremes of the SO the stratospheric temperatures and heights at higher latitudes are well above normal (about 1 standard deviation), and conversely in the cold extremes.
- Yet another influence on the stratosphere is solar variability, which until recently received little attention. The influence of the 11-year SSC will be discussed in the next section.

8.4 SOLAR INFLUENCES ON THE STRATOSPHERE AND TROPOSPHERE

8.4.1 The stratosphere during the northern winter

Based on results published in 1982, Labitzke (1987) and Labitzke and van Loon (1988) found that a signal of the 11-year SSC emerged when the Arctic stratospheric temperatures and geopotential heights were grouped into two categories determined by the direction of the equatorial wind in the stratosphere (QBO). The reality and significance of using this approach have been confirmed by Naito and Hirota (1997), Salby and Callaghan (2000, 2004), and Ruzmaikin and Feynman (2002).

An example of this approach is shown in Figure 8.7 (see colour section), for the 30-hPa heights. On the left-hand side the correlations between the solar cycle and the 30-hPa heights are shown, with the winters in the east phase of the QBO in the upper part of the figure, and the winters in the west phase of the QBO in the lower part. The pattern of correlations is clearly very different in the two groups, with negative correlations over the Arctic in the east phase and large positive correlations there in the west phase. The respective height differences between solar maxima and minima are given on the right-hand side. In the east phase of the QBO the heights tend to be below normal over the Arctic in solar maxima and above normal to the south, whereas in the west phase the Arctic heights tend to be well above normal in solar maxima.

The modulation of the solar signal by the QBO is at its maximum in late winter (February). Figure 8.8 (see colour section) shows a vertical meridional section of correlations between the solar cycle and zonally averaged temperatures, as well as the corresponding temperature differences between solar maxima and minima. When all years are used in February, the correlations and the corresponding temperature differences (top left and right, respectively) are small to the point of insignificance; but in the east phase of the QBO the correlations of the zonally averaged temperatures with the solar data are positive from 60° N to the south pole in the summer hemisphere, and negative north of 60° N in the winter hemisphere. On the right-hand side in the middle panel are the average temperature differences between solar maxima and minima in the east phase of the QBO which correspond to the correlations on the left side. The shading is the same as that in the correlations where it denotes correlations above 0.4.

In the west phase of the QBO (Figure 8.8, bottom, see colour section), the correlations with the solar flux are highly positive over the Arctic and near zero or weakly negative elsewhere. The large positive correlations are associated with the frequent Major Mid-Winter Warmings which occur when the QBO is in the west phase at solar maxima (van Loon and Labitzke, 2000). The Arctic temperatures and heights in the stratosphere are then caused by strong subsidence. Outside the Arctic, the lower latitudes are expected to warm at solar maximum, but because of the subsidence and warming in the Arctic, the warming to the south is counterbalanced by rising motion and cooling, well into the southern (summer) hemisphere.

The expected opposite tendencies between the Arctic and lower latitudes are termed ‘teleconnection’ (Table 8.2). The meaning of this term is that when the atmosphere in one region changes in a given direction – for instance if the pressure rises – there will be a compensating drop in pressure in another region. This concept was recognized by Dove (1839) and later by Angström (1935).

Table 8.2. Schematic representation of the expected meridional changes to follow from the influence of the solar cycle in the winter stratosphere on the Northern Hemisphere (top); and the observed response to the solar influence as modulated by the equatorial Quasi-Biennial Oscillation (bottom). In the east years teleconnection and Sun work in the same direction; in the west years they oppose each other. (Labitzke and van Loon, 2000.)

Expected teleconnections from solar influence

	Solar maximum	Solar minimum
Arctic	Low	High
Extra–Arctic	High	Low

Observed response to solar influences

	East QBO		West QBO	
	Solar maximum	Solar minimum	Solar maximum	Solar minimum
Arctic heights and temperatures	Low	High	High	Low
Extra–Arctic	High	Low	Negligible	Negligible

Teleconnections work through vertical and horizontal motion, or both. Teleconnections in the stratosphere are shown in Shea *et al.* (1992).

The height and temperature changes shown in Figures 8.7 and 8.8 (see color section), also indicate that the solar cycle influences the Mean Meridional Circulation (MMC) – also called Brewer–Dobson Circulation (BDC). Forced by planetary waves, the MMC regulates wintertime polar temperatures through downwelling and adiabatic warming (Kodera and Kuroda, 2002; Kuroda and Kodera, 2002; Hood and Soukharev, 2003; Labitzke, 2003; Hood, 2004; Salby and Callaghan, 2004).

During the west phase of the QBO, the MMC is intensified during solar maxima (and vice versa during solar minima) with large positive anomalies over the Arctic (intensified downwelling and warming), and concurrent weak anomalies (anomalous upwelling/adiabatic cooling) over the tropics and subtropics (lower maps in Figure 8.7 (see colour section), and lowest panels in Figure 8.8 (see colour section).

During the east phase the MMC is weakened in solar maxima, with reduced downwelling (anomalous upwelling/cooling) and negative anomalies over the Arctic in solar maxima, and concurrent anomalous downwelling with positive anomalies over the tropics and subtropics.

The weakened BDC with an intensified polar vortex is also called positive Northern or Southern Annular Mode (NAM or SAM) (Baldwin and Dunkerton, 2001).

8.4.2 The stratosphere during the northern summer

The interseasonal shift between hemispheres of the solar–stratosphere relationship is evident in Figure 8.9. The curves on the left-hand side show the correlations between the 30-hPa zonally averaged temperature in May–August – the four months centred on the northern summer solstice (dashed line). The biggest correlations, above 0.4, lie between 10° N to 50° N, and a secondary peak is found at 15° S. This picture reverses in the four months November–February, which are centred on the southern solstice (solid line), when the largest correlations are found between 15° S and 65° S, and a secondary peak is found at 15° N. The temperature differences between solar maxima and minima are shown on the right-hand side of the figure. They are almost everywhere positive, with the largest differences (more than 1 K) over the respective summer hemisphere.

Spatially, this interseasonal movement is illustrated in Figure 8.10 for the northern summer: on top of the figure and for the period 1968–2004 the subtropical to mid-latitude peak dominates the circumference in the northern hemisphere, and the secondary peak in the southern hemisphere spans that hemisphere. However, when the data are divided into the east and west phases of the QBO, the picture is different. Originally we made this division according to the phase of the QBO only in the winter data; but it turns out that it is also a valid approach for the rest of the year (Labitzke, 2003, 2004a, b, and 2005).

In the east phase of the QBO (Figure 8.10 middle), both the major and the minor peaks are accentuated, whereas in the west phase of the QBO the solar relationship is weaker. In other words, the correlations for all years in the top panel are dominated

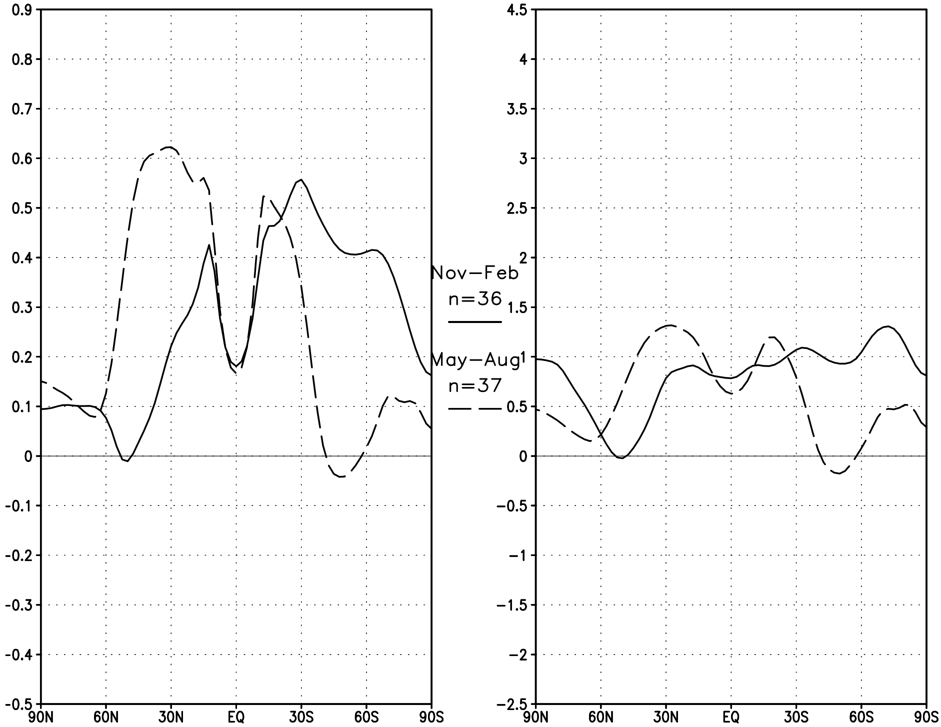


Figure 8.9. (Left) Correlations between the 10.7-cm solar flux and the detrended zonally averaged 30-hPa temperatures in May–June–July–August (dashed line) and November–December–January–February (solid line) (van Loon and Labitzke, 1999, figure 9, updated.) (Right) The respective temperature differences (K) between solar maxima and solar minima. (NCEP/NCAR reanalyses, 1968–2004.)

by the east years. This is further emphasized in the scatter diagrams in Figure 8.11. The two points in this figure were chosen for their high correlations in the east years: $r = 0.87$ at 25° N, and $r = 0.92$ at 20° S. In the west years $r = 0.36$ at the northern position, and only 0.13 in the south. In the east phase the temperature differences between solar maxima and minima are about 2.5 K, which is more than 2 standard deviations, as shown in Figure 8.4.

The correlations on the left in Figure 8.12 show, for July, the vertical distribution of the solar relationship from the upper troposphere to the middle stratosphere. Again, the data are grouped in all years and in the east and west phase of the QBO; the corresponding temperature differences are on the right in the diagram. The results for the east phase (middle panels) are most striking: two centres with correlations above 0.8 are found over the subtropics between 20 and 30 hPa; further down, the double maximum in the east phase changes into one maximum, centred on the equatorial tropopause (Labitzke, 2003, figure 4). The temperature differences between solar maxima and minima are large – more than two standard deviations

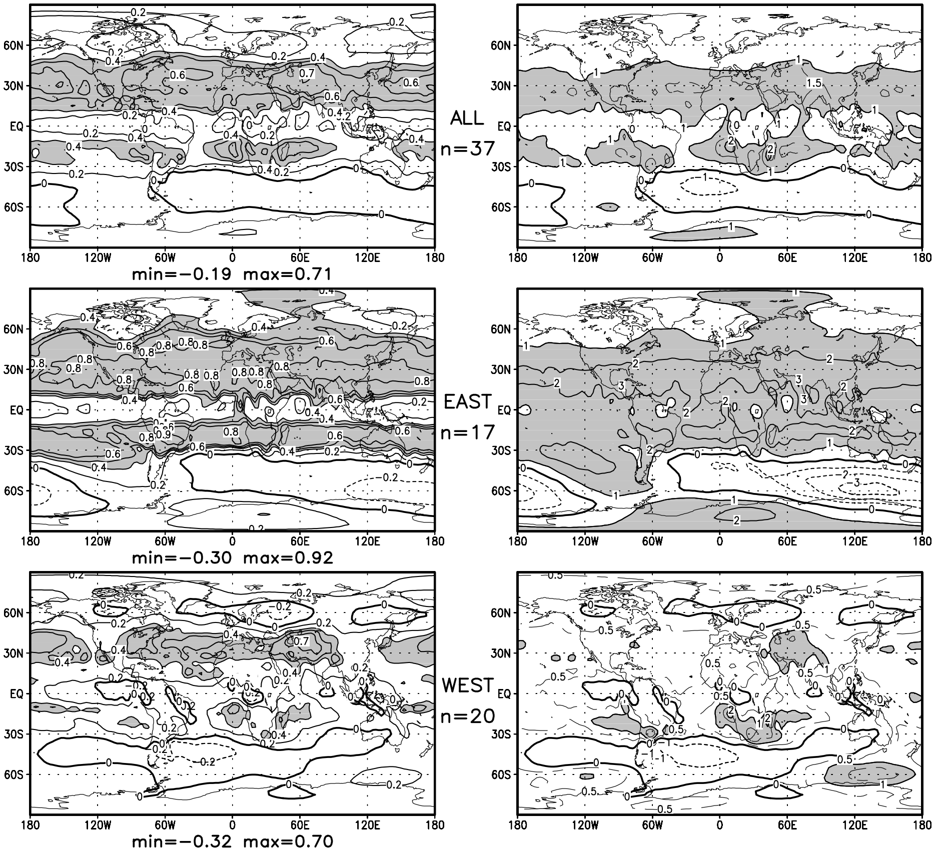


Figure 8.10. (Left) Correlations between the 10.7-cm solar flux and the detrended 30-hPa temperatures in July, shaded for emphasis where the correlations are above 0.4. (Right) The respective temperature differences (K) between solar maxima and minima, shaded where the differences are above 1 K. (Upper panels) all years; (middle panels) only years in the east phase of the QBO; (lower panels) only years in the west phase of the QBO. (NCEP/NCAR re-analyses, 1968–2004); (Labitzke, 2003, figure 1, updated.)

in some regions. This warming (positive anomalies) can only be explained by downwelling over the subtropics and tropics – roughly between 60°N and 30°S – which in other words means a weakening of the BDC in solar maxima/east phase of the QBO, as discussed above for the northern winter (Kodera and Kuroda, 2002; Shepherd, 2002).

The solar signal is much weaker in the west phase. It hints at an intensification of the Hadley Circulation (HC) over the northern hemisphere, with stronger rising over the equator (warming due to latent heat release) and some anomalous heating (downwelling) over the northern summer hemisphere.

Figure 8.13 shows, for the 30-hPa heights, the same correlations and differences as does Figure 8.10 for the temperatures: the east phase dominates the solar signal.

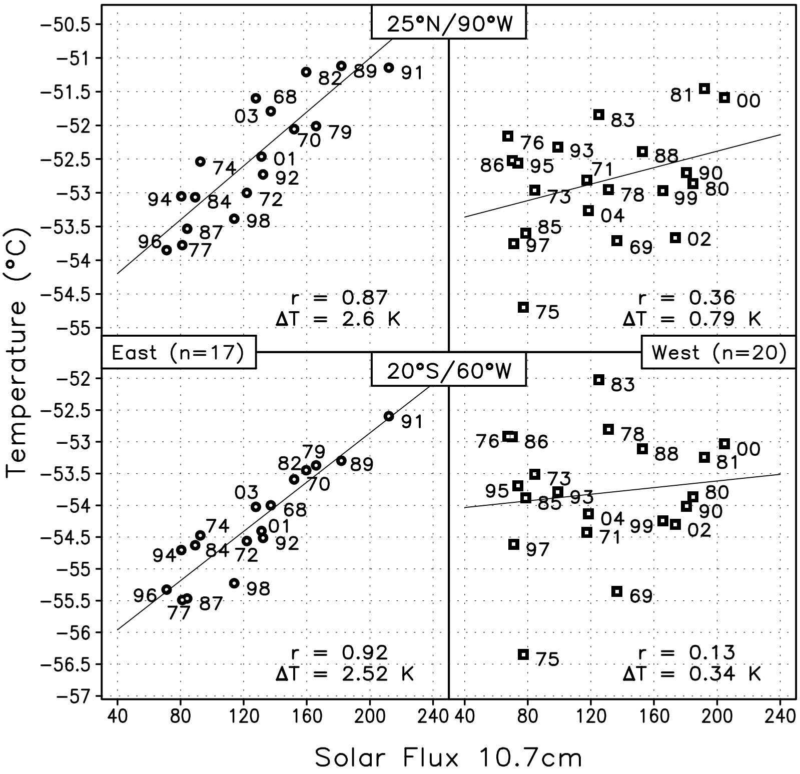


Figure 8.11. Scatter diagrams (detrended 30-hPa temperatures (°C) in July against the 10.7-cm solar flux at two grid points). (Upper panels) 25° N/90° W; (lower panels) 20° S/60° W. (Left) years in the east phase of the QBO ($n = 17$); (right) years in the west phase of the QBO ($n = 20$). The numbers indicate the respective years. Period: 1968–2004. (Labitzke, 2003, figure 3, updated.)

In addition, the anomalous zonal (west–east) wind in the equatorial belt is affected by the solar variability on the decadal scale; on the right-hand side, in the middle panel, an anomalous high value is centred on the equator. It means that an anomalous anticyclonic circulation is centred on the equator in the solar maximum east years, connected with anomalous winds from the west. Therefore, during solar maxima in east years the low-latitude east wind is weakened, and conversely in the solar minimum years.

In the west phase of the QBO (bottom right in Figure 8.13), the geopotential heights are lowest on the equator in the solar maxima and the anomalous winds are from the east, around the anomalous low on the equator; and conversely in the solar minima and west years. The QBO thus not only modulates the solar signal on the decadal scale, but is itself modulated by the solar variability (Salby and Callaghan, 2000; Soukharev and Hood, 2001; Labitzke, 2003).

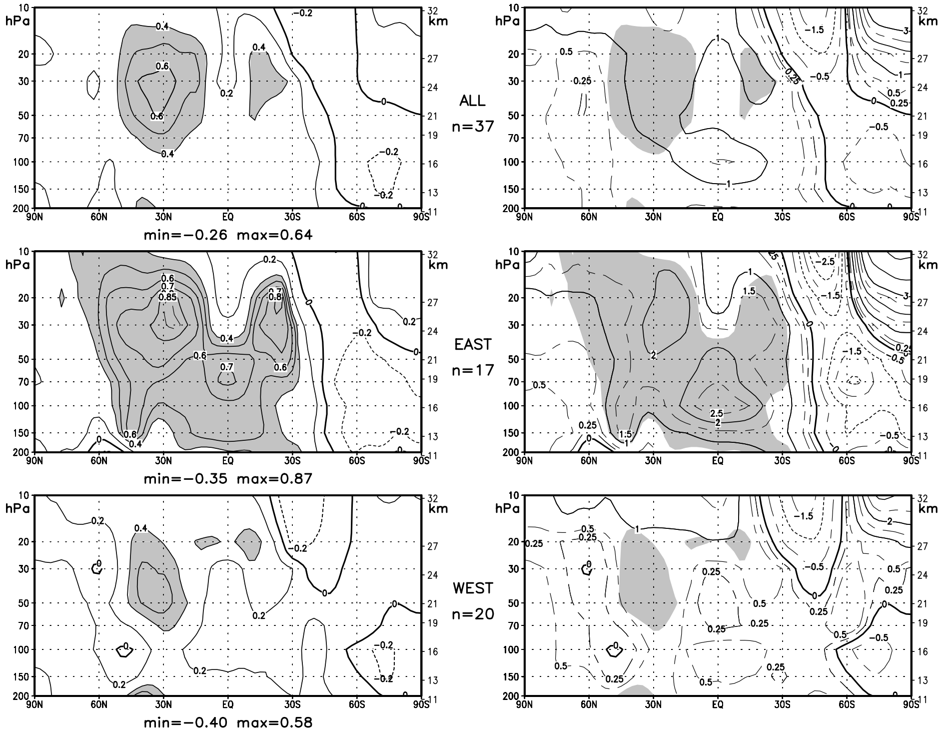


Figure 8.12. Vertical meridional sections between 200 and 10 hPa of (left) the correlations between the detrended zonally averaged temperatures for July and the 10.7-cm solar flux, shaded for emphasis where the correlations are larger than 0.4. (Right) The respective temperature differences (K) between solar maxima and minima, shaded where the corresponding correlations on the left-hand side are above 0.4. (Upper panels) all years; (middle panels) only years in the east phase of the QBO; (lower panels) only years in the west phase of the QBO. (NCEP/NCAR reanalyses, 1968–2004; Labitzke, 2003, figure 4, updated.)

8.4.3 The troposphere

There are several indications that solar forcing can also affect the troposphere. For instance, Labitzke and van Loon (1992 and 1995) and van Loon and Labitzke (1994) noted that radiosonde stations in the tropics and subtropics of the northern hemisphere showed a marked difference in the vertical distribution of temperature between maxima and minima in the solar decadal oscillation, the temperatures being higher in the maxima in the troposphere and stratosphere, and lower or little changed in the tropopause region. These results from single radiosonde stations are supplemented here with the temperature differences in space between 20° S and 40° N, averaged over the longitudes from which the radiosonde data were obtained (Figure 8.14 (from van Loon *et al.*, 2004)). The grid-point data from the NCEP/NCAR analyses in Figure 8.14 agree well with the radiosondes used by

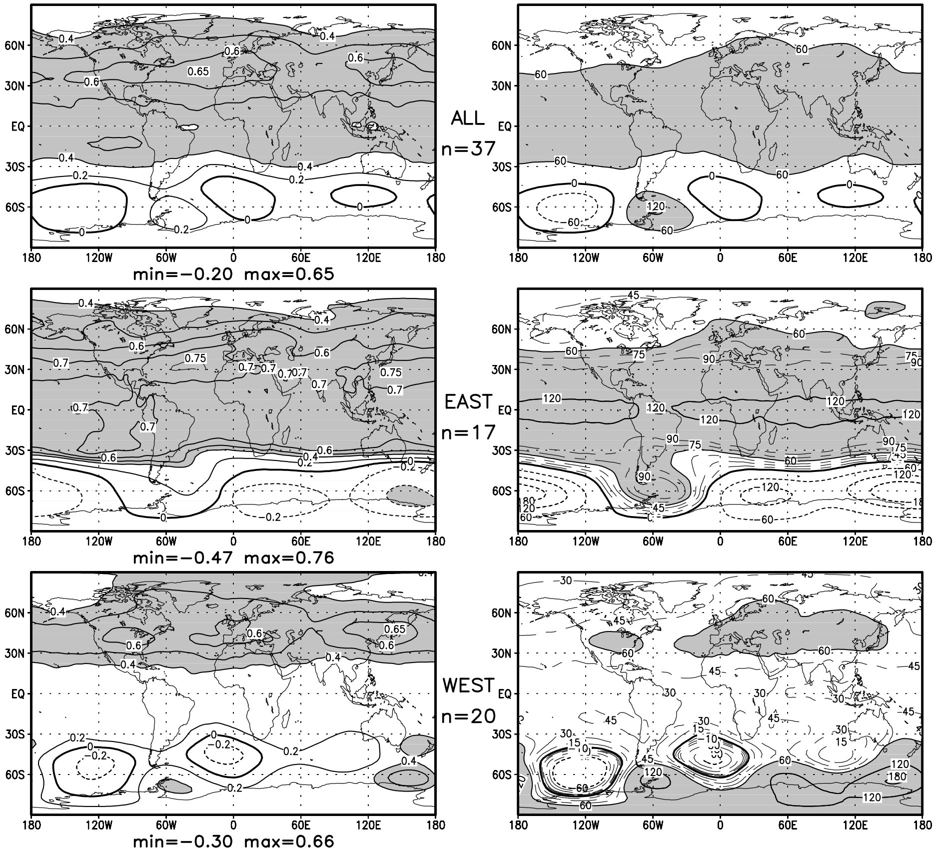


Figure 8.13. (Left) Correlations between the 10.7-cm solar flux and the detrended 30-hPa heights in July, shaded for emphasis where the correlations are above 0.4. (Right) The respective height differences (geopot. m) between solar maxima and minima, shaded where the differences are above 60 gpm. (Upper panels) all years; (middle panels) only years in the east phase of the QBO; (lower panels) only years in the west phase of the QBO. (NCEP/NCAR reanalyses, 1968–2004; Labitzke, 2003, figure 5, updated.)

van Loon and Labitzke, and they extend their analysis by two more solar periods. Gleisner and Thejl (2003) and Coughlin and Tung (2004) found similar positive anomalies in the troposphere of the tropics and subtropics, associated with an increase in the solar irradiance.

Furthermore, van Loon and Labitzke (1998, figures 10 and 11) demonstrated that EOF 1 in the 30-hPa temperatures and heights follows the interannual course of the solar 10.7-cm flux, and accounts for over 70% of the interannual variance in the summer of both hemispheres. This eigenvector at 30 hPa is well correlated with the temperatures in the troposphere.

In addition, the three-year running, area-weighted means of the zonally averaged temperature of the entire northern hemisphere – in the layer between

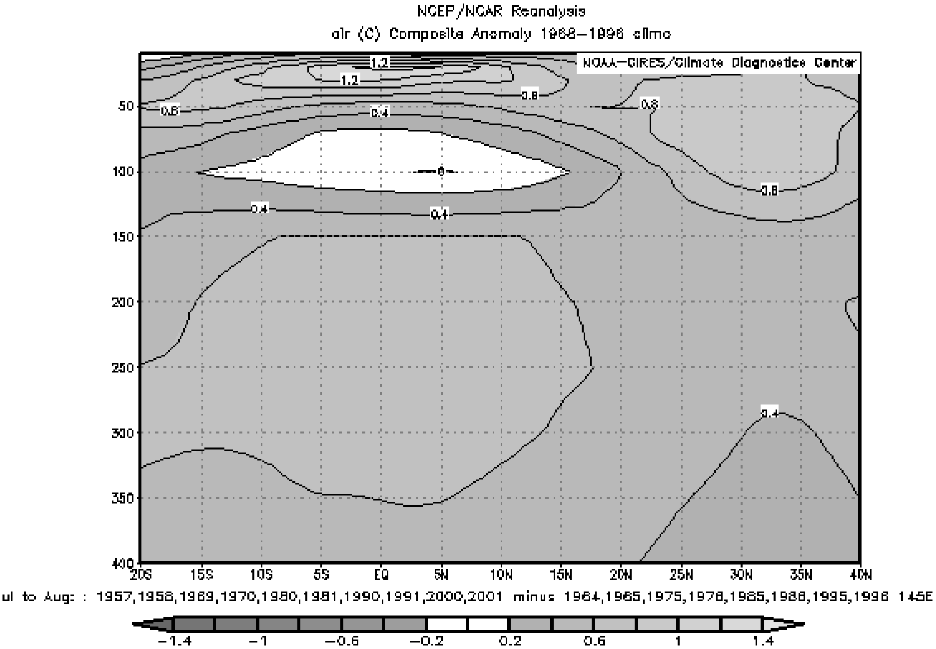


Figure 8.14. The zonally averaged temperature difference (K) (from 145° E eastward to 68° W) between five solar maxima (1957–1958, 1969–1970, 1980–1981, 1990–1991, 2000–2001) and four solar minima (1964–1965, 1975–1976, 1985–1986, 1995–1996). Between 400 hPa and 10 hPa. (NCEP/NCAR reanalyses, 1957–2001; van Loon and Shea, 2000, updated.)

700 hPa and 200 hPa in July–August – was shown by van Loon and Shea (1999 and 2000) to follow the decadal solar oscillation, with higher temperatures in the solar maxima than in the minima. The temperature of the nearly 9-km thick layer correlated with the solar oscillation at $r = 0.65$ for July–August (Figure 8.15 (van Loon and Shea, 2000)) and $r = 0.57$ for the 10-month average March to December (van Loon and Shea, 1999, not shown).

Van Loon and Shea (2000) found the strongest solar signal at 30 hPa to 20 hPa in the zonally averaged temperatures, Figure 8.14; and the signal decreased with decreasing height, approaching zero near the ground. This does not mean that a solar signal does not exist at the surface locally. Figures 8.16–8.18 (see colour section) (van Loon *et al.*, 2004) emphasize the danger of relying on zonal averages: the difference in the tropical rainfall in the eastern Indian–western Pacific Oceans – a climatically sensitive region – between maxima and minima in the solar decadal oscillation is markedly positive (higher rainfall in solar maxima) in the east and negative in the west (Figure 8.16). This is reflected in the anomalies in the vertical motion (Figure 8.17) where there is stronger upward motion (negative values) in solar maxima above the higher rainfall than over the Indian Ocean, where the rainfall is lower in the solar maxima and the vertical motion is anomalously downward (positive values). Further confirmation is obtained by the differences in

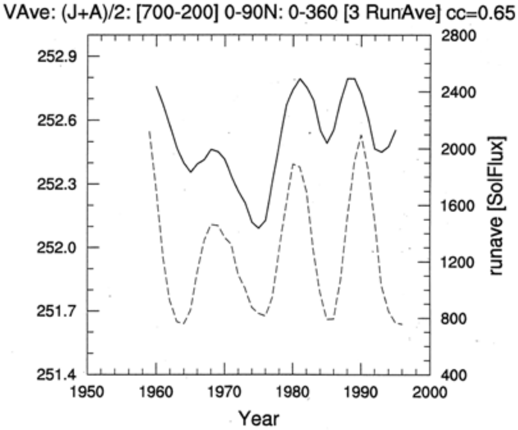


Figure 8.15. (Solid line) Three-year running means of the temperature (K) in the layer between 700 hPa and 200 hPa, averaged over the northern hemisphere in July–August and area weighted. (Dashed line) The 10.7-cm solar flux, used as index of the solar cycle. (NCEP/NCAR reanalyses, 1957–2001; van Loon and Shea, 2000.)

the outgoing longwave radiation (Figure 8.18). The negative values of the OLR in the east are due to the fact that the cloud tops are higher over the larger rainfall in the east in solar maxima than in the minima; and the negative values in the west (lower cloud tops in the solar maxima) are associated with the lower rainfall in solar maxima than in the minima.

These differences in rainfall, vertical motion and outgoing long-wave radiation point to differences between solar maxima and minima in the tropical Walker Cell (vertical, oriented west to east) and in the Hadley Cell (vertical, oriented south to north) as documented in van Loon *et al.* (2004).

8.5 MODELS AND MECHANISMS

Based on observations, the results presented above demonstrate conclusively the existence of a solar cycle in the stratospheric and tropospheric temperatures and heights. There have been many model studies with General Circulation Models (GCMs) to investigate the impact of changes in the solar constant, but the change from solar maxima to minima within the SSC is only about 0.1%, and the influence on the atmosphere (in the models) is very small.

Kodera *et al.* (1991) and Rind and Balachandran (1995) were the first to use GCMs with a better resolution of the stratosphere to study the effects of increases in solar UV. Later, Haigh (1996 and 1999) and Shindell *et al.* (1999) carried out experiments where they imposed in the GCMs realistic changes in the UV part of the solar spectrum and estimates of the resulting ozone changes.

There is general agreement that the direct influence of the changes in the UV part of the spectrum (6–8% between solar maxima and minima) leads to more ozone and warming in the upper stratosphere (around 50 km) in solar maxima (Haigh, 1994; Hood *et al.*, 1993; Hood, 2004). This leads to changes in the thermal gradients and thus in the wind systems, and by this to changes in the vertical propagation of the planetary waves that drive the global circulation. Therefore, the relatively weak, direct radiative forcing of the solar cycle in the stratosphere can lead to a large indirect dynamical response in the lower atmosphere through a modulation of the polar night jet (PNJ), as well as through a change in the Brewer Dobson Circulation (BDC) (Kodera and Kuroda, 2002).

Some of the model results were found to be of similar structure to those seen in the analysis of data in the stratosphere by, for example, van Loon and Labitzke (2000), but the size of the changes were much smaller than observed, especially during summer. This is probably due to the fact that the GCMs do not produce a QBO and that these models are not coupled to the oceans, so that the most important natural forcings (see Table 8.1) are not included in the modelling.

Recently, Matthes *et al.* (2004), using the Freie Universität Berlin–Climate Middle Atmosphere Model (FUB-CMAM), introduced in addition to the realistic spectral UV changes and ozone changes a relaxation towards observed equatorial wind profiles throughout the stratosphere, representing the east and west phases of the QBO, as well as the Semiannual Oscillation (SAO) in the upper stratosphere. The importance of the SAO in the upper stratosphere has been stressed by Gray *et al.* (2001a, b).

During the Arctic winter a realistic poleward–downward propagation of the PNJ anomalies, significantly weaker planetary wave activity and a weaker mean meridional circulation under solar maximum conditions are reproduced in the FUB-CMAM. This confirms the solar signal observed in the upper stratosphere, by, for example, Kodera and Yamazaki (1990) and Kuroda and Kodera (2002). The observed interaction between the Sun and the QBO is captured, and stratospheric warmings occur preferentially in the west phase of the QBO during solar maxima, (see Section 8.4).

It should be pointed out that other GCM studies so far have failed to produce such a good correspondence with the observed magnitude and temporal evolution of the zonal wind anomalies in northern winters.

A complete understanding of the mechanisms which are transferring the direct solar signal from the upper to the lower stratosphere and to the troposphere is still missing, and the final amplitude of the solar signal will only be established after the acquisition of data from more solar cycles (Matthes *et al.*, 2006).

8.6 ACKNOWLEDGEMENTS

We thank the members of the Stratospheric Research Group, FUB, for professional support, and Dipl. Met. Markus Kunze for the computations and graphics. The 10.7-cm solar flux data are from the World Data Center A, Boulder, Colorado. The solar irradiance data in Figure 8.2 (see colour section) are available from the

PMOD/WRC, Davos, Switzerland, *ftp://ftp.pmodwrc.ch/pub/data/irradiance/composite/*.

8.7 REFERENCES

Abbot, C. G., Fowle, F. E., Aldrich, L. B., 1913. *Annals of the Astrophysical Observatory of the Smithsonian Institution*, **3**, Washington.

Angström, A., 1935. Teleconnections of climate changes in present time. *Geogr. Annal.*, **17**, 242–258.

Baldwin, M. P., Dunkerton, T. J., 2001. Stratospheric harbingers of anomalous weather regimes, *Science*, **294**, 581–584.

Baldwin, M. P. *et al.*, 2001. The quasi-biennial oscillation, *Rev. Geophys.*, **39**, 179–229.

Chandra, S., McPeters, R. D., 1994. The solar cycle variation of ozone in the stratosphere inferred from Nimbus 7 and NOAA 11 satellites, *J. Geophys. Res.*, **99**, 20665–20671.

Coughlin, K. T., Tung, K. K., 2004. 11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method, *Adv. Space Res.*, **34**, 323–329.

Damon, P. E., Laut, P., 2004. Pattern of strange errors plagues solar activity and terrestrial climate data, *EOS (AGU)*, **85**, No. 39, Forum, 370 and 374.

Dove, H. W., 1839. Uber die geographische Verbreitung gleichartiger Witterungserscheinungen, Abhandlungen der königlichen Akademie der Wissenschaften, Berlin, 1838, 287–415 (see p. 409).

Friis-Christensen, E., Lassen, K., 1991. Length of the solar cycle: an indicator of solar activity closely associated with climate, *Science*, **254**, 698–700.

Friis-Christensen, E., Lassen, K., 1993. Two and a half centuries of solar activity variations and a possible association with global temperature, Danish Meteorological Institute, Scientific Report, 93–94.

Fröhlich, C., 2000. Observations of irradiance variations, *Space Sci. Rev.*, **94**, 15–24.

Fröhlich, C., 2004. Solar irradiance variability, in *Solar Variability and its Effects on the Earth's Atmosphere and Climate System*, AGU Monograph Series, eds. J. Pap *et al.*, AGU, Washington D.C.

Fröhlich, C., Lean, J., 1998. The sun's total irradiance: cycles, trends and related climate change uncertainties since 1978, *Geophys. Res. Lett.*, **25**, 4377–4380.

Gleisner, H., Thejl, P., 2003. Patterns of tropospheric response to solar variability, *Geophys. Res. Lett.*, **30**(13), 1711, doi: 10.1029/2003GL17129.

Gray, L., Drysdale, E. F., Dunkerton, T. J., Lawrence, B., 2001a. Model studies of the interannual variability of the northern hemisphere stratospheric winter circulation: The role of the Quasi-Biennial Oscillation, *Q. J. Roy. Met. Soc.*, **127**, 1413–1432.

Gray, L., Phipps, S. J., Dunkerton, T. J., Baldwin, M. P., Drysdale, E. F., Allen, M. R., 2001b. A data study of the influence of the equatorial upper stratosphere on northern hemisphere stratospheric sudden warmings, *Q. J. Roy. Met. Soc.*, **127**, 1985–2003.

Haigh, J. D., 1994. The role of stratospheric ozone in modulating the solar radiative forcing of climate, *Nature*, **370**, 544–546.

Haigh, J. D., 1996. The impact of solar variability on climate, *Science*, **272**, 981–984.

Haigh, J. D., 1999. A GCM study of climate change in response to the 11-year solar cycle, *Q. J. Roy. Met. Soc.*, **125**, 871–892.

Holton, J., Tan, H., 1980. The influence of the equatorial Quasi-Biennial Oscillation on the global circulation at 50 mb, *J. Atmos. Sci.*, **37**, 2200–2208.

Hood, L. L., 2003. Thermal response of the tropical tropopause region to solar ultraviolet variations. *Geophys. Res. Lett.*, **30**, No. 23, 2215, doi: 10.1029/2003 GL018364.

Hood, L. L., 2004. Effects of solar UV variability on the stratosphere, in *Solar Variability and its Effect on the Earth's Atmosphere and Climate System*, AGU Monograph Series, eds. J. Pap *et al.*, American Geophysical Union, Washington D.C., 283–304.

Hood, L. L., Soukharev, B., 2003. Quasi-decadal variability of the tropical lower stratosphere: the role of extratropical wave forcing, *J. Atmos. Sci.*, **60**, 2389–2403.

Hood, L. L., Jirikowic, J. L., McCormack, J. P., 1993. Quasi-decadal variability of the stratosphere: influence of long-term solar ultraviolet variations. *J. Atmos. Sci.*, **50**, 3941–3958.

Hoyt, D. V., Schatten, K. H., 1997. *The Role of the Sun in Climate Change*, Oxford University Press, New York.

Kalnay, E., Kanamitsu, R., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Reynolds, R., Jenne, R., Joseph, J., 1996. The NCEP/NCAR 40-year re-analysis project, *Bull. Am. Met. Soc.*, **77**, 437–471.

Kodera, K., Kuroda, Y., 2002. Dynamical response to the solar cycle, *J. Geophys. Res.*, **107**(D24), 4749, doi:10.1029/2002JD002224.

Kodera, K., Yamazaki, K., 1990. Long-term variation of upper stratospheric circulation in the northern hemisphere in December, *J. Met. Soc. Japan*, **68**, 101–105.

Kodera, K., Chiba, M., Shibata, K., 1991. A general circulation model study of the solar and QBO modulation of the stratospheric circulation during northern hemisphere winter, *Geophys. Res. Lett.*, **18**, 1209–1212.

Kristjánsson, J. E., Kristiansen, J., Kaas, E., 2004. Solar activity, cosmic rays, clouds and climate – an update. *Adv. Space Res.*, **34**, 407–415.

Kuroda, Y., Kodera, K., 2002. Effect of solar activity on the polar-night jet oscillation in the northern and southern hemisphere winter, *J. Met. Soc. Japan*, **80**, 973–984.

Labitzke, K., 1982. On the interannual variability of the middle stratosphere during the northern winters, *J. Met. Soc. Japan*, **60**, 124–139.

Labitzke, K., 1987. Sunspots, the QBO, and the stratospheric temperature in the north polar region, *Geophys. Res. Lett.*, **14**, 535–537.

Labitzke, K., 2002. The solar signal of the 11-year sunspot cycle in the stratosphere: Differences between the northern and southern summers, *J. Met. Soc. Japan*, **80**, 963–971.

Labitzke, K., 2003. The global signal of the 11-year solar cycle in the atmosphere: When do we need the QBO? *Meteorolog. Zeitschr.*, **12**, 209–216.

Labitzke, K., 2004a. On the signal of the 11-year sunspot cycle in the stratosphere and its modulation by the Quasi-Biennial Oscillation (QBO), *J. Atmos. Sol.–Terr. Phys.*, **66**, 1151–1157.

Labitzke, K., 2004b. On the signal of the 11-year sunspot cycle in the stratosphere over the Antarctic and its modulation by the Quasi-Biennial Oscillation (QBO), *Meteorolog. Zeitschr.*, **13**, 263–270.

Labitzke, K., 2005. On the solar cycle-QBO relationship: a summary. *J. Atmos. Solar–Terr. Phys.*, **67**, 1–2, 45–54.

Labitzke, K. and collaborators, 2002. *The Berlin Stratospheric Data Series*, CD from Meteorological Institute, Free University Berlin.

Labitzke, K., van Loon, H., 1988. Associations between the 11-year solar cycle, the QBO and the atmosphere. Part I: The troposphere and stratosphere in the northern hemisphere winter, *J. Atmos. Terr. Phys.*, **50**, 197–206.

Labitzke, K., van Loon, H., 1992. Association between the 11-year solar cycle and the atmosphere. Part V: Summer, *J. Clim.*, **5**, 240–251.

Labitzke, K., H. van Loon, 1995. Connection between the troposphere and stratosphere on a decadal scale, *Tellus*, **47**A, 275–286.

Labitzke, K., van Loon, H., 1999. *The Stratosphere (Phenomena, History, and Relevance)*, Springer Verlag, Berlin, Heidelberg, New York.

Labitzke, K., van Loon, H., 2000. The QBO effect on the global stratosphere in northern winter, *J. Atmos. Solar–Terr. Phys.*, **62**, 621–628.

Lean, J. L., Rottman, G. J., Kyle, H. L., Woods, T. N., Hickey, J. R., Puga, L. C., 1997. Detection and parameterisation of variations in solar mid- and near-ultraviolet radiation (200–400 nm), *J. Geophys. Res.*, **102**, 29939–29956.

Matthes, K., Langematz, U., Gray, L. L., Kodera, K., Labitzke, K., 2004. Improved 11-year solar signal in the FUB-CMAM, *J. Geophys. Res.*, **109**, doi: 10.1029/2003JD004012.

Matthes, K., Kodera, K., Kuroda, Y., Langematz, U., 2006: The transfer of the solar signal from the stratosphere to the troposphere: Northern winter, *J. Geophsy. Res.*, doi: 10.1029/2005JD006283.

Naito, Y., Hirota, I., 1997. Interannual variability of the northern winter stratospheric circulation related to the QBO and the solar cycle, *J. Meteor. Soc. Japan*, **75**, 925–937.

Naujokat, B., 1986. An update of the observed Quasi-Biennial Oscillation of the stratospheric winds over the tropics, *J. Atmos. Sci.*, **43**, 1873–1877.

Pittock, A. B., 1983. Solar variability, weather and climate: An update, *Q. J. Roy. Met. Soc.*, **109**, 23–55.

Reid, G. C., 2000. Solar variability and the Earth's climate: Introduction and overview, *Space Science Rev.*, **94**, 1–11.

Rind, D., Balachandran, N. K., 1995. Modelling the effects of UV variability and the QBO on the troposphere–stratosphere systems. Part II: The troposphere, *J. Clim.*, **8**, 2080–2095.

Ruzmaikin, A., Feynman, J., 2002. Solar influence on a major mode of atmospheric variability, *J. Geophys. Res.*, **107**(14), doi: 10.1029/2001JD001239.

Salby, M., Callaghan, P., 2000. Connection between the solar cycle and the QBO: The missing link, *J. Clim.*, **13**, 2652–2662.

Salby, M., Callaghan, P., 2004. Evidence of the solar cycle in the general circulation of the stratosphere, *J. Clim.*, **17**, 34–46.

Shea, D. J., van Loon, H., Labitzke, K., 1992. *Point correlations of geopotential height and temperature at 30 mb and between 500 mb and 30 mb*, NCAR/TN 368 + STR, 6 pp. plus 285 pp. maps. National Center for Atmospheric Research, Boulder, Colorado.

Shepherd, T., 2002. Issues in stratosphere–troposphere coupling, *J. Met. Soc. Japan*, **80**, 769–792.

Shindell, D., Rind, D., Balachandran, N. K., Lean, J., Lonergan, J., 1999. Solar cycle variability, ozone and climate, *Science*, **284**, 305–308.

Soukharev, B., Hood, L. L., 2001. Possible solar modulation of the equatorial quasi-biennial oscillation: Additional statistical evidence, *J. Geophys. Res.*, **106**, 14,855–14,868.

Svensmark, H., 1998. Influence of cosmic rays on climate, *Physical Rev. Lett.*, **81**, 5027.

Svensmark, H., Friis-Christensen, E., 1997. Variation of cosmic ray flux and global cloud coverage – a missing link in solar-climate relationships, *J. Atmos. Solar–Terr. Phys.*, **59**, 1225.

Udelhofen, P. M., Cess, R. D., 2001. Cloud cover variations over the United States: An influence of cosmic rays or solar variability?, *Geophys. Res. Lett.*, **28**, 2617–2620.

van Loon, H., Labitzke, K., 1987. The Southern Oscillation. Part V: The anomalies in the lower stratosphere of the Northern Hemisphere in winter and a comparison with the Quasi-Biennial Oscillation, *Monthly Weather Rev.*, **115**, 357–369.

van Loon, H., Labitzke, K., 1993. Interannual variations in the stratosphere of the Northern Hemisphere: A description of some probable influences. Interactions Between Global Climate Subsystems, The Legacy of Hann, *Geophys. Monograph*, **75**, IUGG 15, 111–122.

van Loon, H., Labitzke, K., 1994. The 10–12 year atmospheric oscillation. Review article in *Meteorologische Zeitschrift*, N.F., **3**, 259–266.

van Loon, H., Labitzke, K., 1998. The global range of the stratospheric decadal wave. Part I: Its association with the sunspot cycle in summer and in the annual mean, and with the troposphere, *J. Clim.*, **11**, 1529–1537.

van Loon, H., Labitzke, K., 1999. The signal of the 11-year solar cycle in the global stratosphere, *J. Atmos. Solar–Terr. Phys.*, **61**, 53–61.

van Loon, H., Labitzke, K., 2000. The influence of the 11-year solar cycle on the stratosphere below 30 km: A review, *Space Sci. Rev.*, **94**, 259–278.

van Loon, H., Shea, D. J., 1999. A probable signal of the 11-year solar cycle in the troposphere of the Northern Hemisphere, *Geophys. Res. Lett.*, **26**, 2893–2896.

van Loon, H., Shea, D. J., 2000. The global 11-year signal in July–August. *Geophys. Res. Lett.*, **27**, 2965–2968.

van Loon, H., Meehl, G. E., Arblaster, J. M., 2004. A decadal solar effect in the tropics in July–August, *J. Atmos. Solar–Terr. Physics*, **66**, 1767–1778.

9

Space weather effects on communications

Louis J. Lanzerotti

In the last century and a half, since the invention and deployment of the first electrical communication system – the electrical telegraph – the variety of communications technologies that can be affected by natural processes occurring on the Sun and in the space environment around Earth have vastly increased. This chapter presents some of the history of the subject of space weather as it affects communications systems, beginning with the earliest electric telegraph systems and continuing to today's wireless communications using satellites and land links. An overview is presented of the present-day communications technologies that can be affected by solar–terrestrial phenomena such as solar and galactic charged particles, solar-produced plasmas, and geomagnetic disturbances in the Earth's magnetosphere and ionosphere.

9.1 INTRODUCTION

The discovery of magnetically confined charged particles (electrons and ions) around Earth by Van Allen (Van Allen *et al.*, 1958) and by Vernov and Chudakov (1960) demonstrated that the space environment around Earth, above the sensible atmosphere, was not benign. Measurements by spacecraft in the five decades since Van Allen's work has demonstrated that Earth's near-space environment – inside the magnetosphere – is filled with particle radiation of sufficient intensity and energy to cause significant problems for satellite materials and electronics that might be placed into it.

Because of the trapped radiation (augmented by trapped electrons from the high-altitude Starfish nuclear explosion on 8 July 1962), the world's first commercial telecommunications satellite, the low-orbit Telstar 1 (launched on 10 July 1962) (*Bell System Technical Journal*, 1963), suffered anomalies in one of its two command lines within a couple of months of its launch, and within five months both command lines had failed. While clever engineering by Bell Laboratories personnel resurrected the satellite for more than a month in early 1963, by the end of February of that year Telstar had fallen silent for good – a victim of the solar–terrestrial radiation environment (Reid, 1963).

It was immediately clear from Van Allen's discovery and then from the Telstar experience that the Earth-orbiting telecommunications satellites that had been proposed by Arthur C. Clark (1945) and by John Pierce (1954) prior to the space age would now have to be designed to withstand the Earth's radiation environment. The semiconductor electronic parts (which were the obvious choice for even the earliest spacecraft and instrument designs) would have to be carefully evaluated and qualified for flight. Furthermore, the space radiation environment would have to be carefully mapped, and time dependencies of the environment would need to be understood if adequate designs were to be implemented to ensure the success of the missions.

9.2 EARLY EFFECTS ON WIRE-LINE TELEGRAPH COMMUNICATIONS

The effects of the solar–terrestrial environment on communications technologies began long before the space age. In 1847, during the eighth solar cycle, telegraph systems that were just beginning to be deployed were found to frequently exhibit 'anomalous currents' flowing in their wires. W. H. Barlow – a telegraph engineer with the Midland Railway in England – appears to be the first to have recognized these currents. Since they were disturbing the operations of the railway's communications system, Barlow (1849) undertook a systematic study of the currents. Making use of a spare wire that connected Derby and Birmingham, Barlow recorded, during a two-week interval (with the exception of the weekend) in May 1847, the deflections in the galvanometer at the Derby station that he installed specifically for his experiment. These data (taken from a Table in his paper) are plotted in Figure 9.1. The galvanometer deflections obviously varied from hour to hour and from day to day by a cause (or causes) unknown to him and his fellow engineers.

The hourly means of Barlow's data for the Derby to Birmingham link, as well as for measurements on a dedicated wire from Derby to Rugby, are plotted in Figure 9.2. A very distinct diurnal variation is apparent in the galvanometer readings: the galvanometers exhibited large right-handed swings during local daytime and left-handed swings during local night-time. The systematic daily change evident in Figure 9.2, while not explicitly recognized by Barlow in his paper, is probably the first measurement of the diurnal component of geomagnetically-induced Earth currents (which, of whatever time scale, were often referred to in subsequent literature in the nineteenth and early twentieth centuries as 'telluric currents'). Such diurnal variations in the telluric currents have been recognized for many decades to be produced by solar-induced effects on the Earth's dayside ionosphere (Chapman and Bartels, 1940).

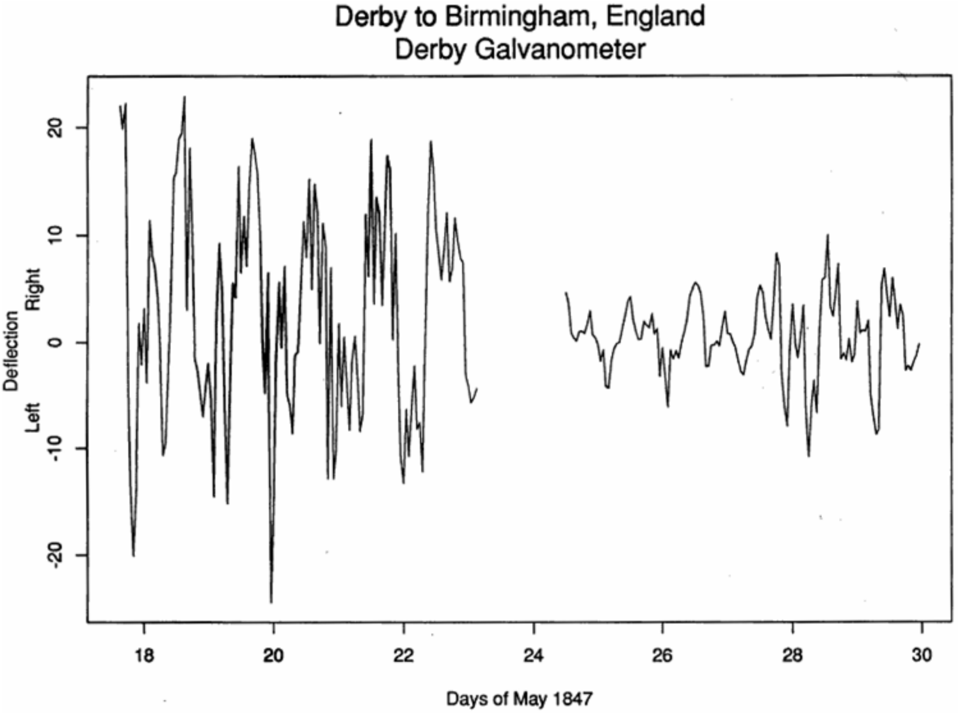


Figure 9.1. Hourly galvanometer recordings of voltage across a cable from Derby to Birmingham, England, in May 1847.

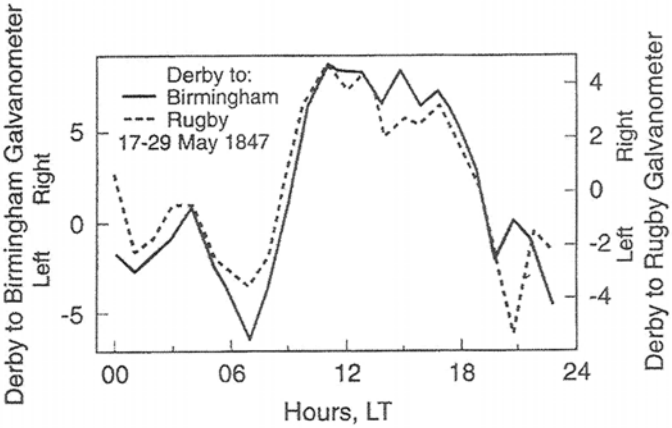


Figure 9.2. Hourly mean galvanometer deflections recorded on telegraph cables from Derby to Birmingham (solid line) and to Rugby (dashed line) in May 1847.

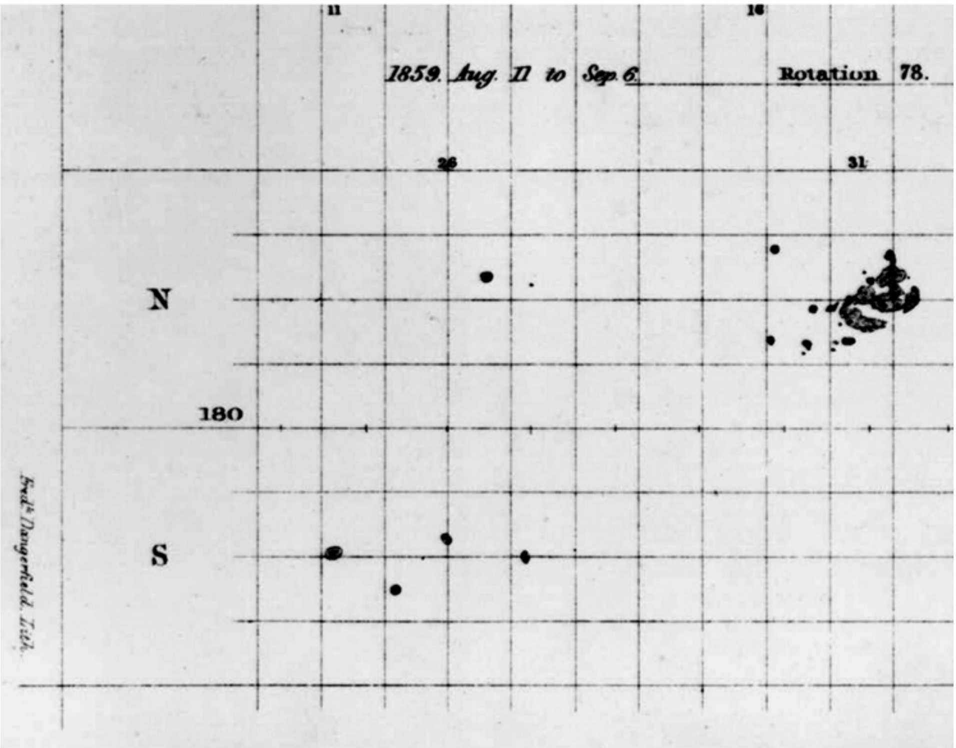


Figure 9.3. Plate 80 from Carrington (1863), showing his sunspot drawings for 11 August to 6 September 1859. The large spot area at about 45°N solar latitude on 31 August is especially notable.

In further discussing his measurements, Barlow noted that 'in every case which has come under [his] observation, the telegraph needles have been deflected whenever aurora has been visible'. Indeed, this was certainly the case during November 1847 as the peak of the sunspot cycle approached, but after Barlow's measurements on the two dedicated Midland Railway wires such monitoring apparently ceased. At that time, large auroral displays over Europe were accompanied by severe disruptions of the Midland Railway telegraph lines, as well as of telegraph lines in other European locations, including the line from Florence to Pisa (Prescott, 1860).

Twelve years after Barlow's pioneering observations (at the end of August 1859 during the tenth solar cycle), while pursuing his systematic programme of observations of spots on the Sun, Richard Carrington, FRS (Fellow Royal Society), recorded an exceptionally large area of spots in the Sun's northern solar hemisphere. Figure 9.3 is a reproduction of Plate 80 from the comprehensive records of his studies, which were carried out over a period of more than seven years around the peak of that sunspot cycle (Carrington, 1863). The large spot area at about 45° N solar latitude on 31 August is especially notable. This observation of an extensive sunspot region on the solar face was more out of the ordinary than Carrington's past research would have originally suggested to him. Quoting from his description of this region: 'At [the observatory at] Redhill [I] witnessed ... a singular outbreak of light which lasted about 5 minutes, and moved sensibly over the entire contour of the spot.' Some hours following this outburst of light from the large dark sunspot region (the first ever reported), disturbances were observed in magnetic measuring instruments on Earth, and the aurora borealis was seen as far south as Rome and Hawaii.

Although Barlow had remarked on the apparent association of auroral displays and the disturbances on his railway telegraph wires, the large and disruptive disturbances that were recorded in numerous telegraph systems within a few hours of Carrington's solar event were nevertheless a great surprise when the many sets of observations and of data began to be compared. (Unlike in the present day, communications between scientists and engineers in the nineteenth century were not nearly instantaneous as now facilitated by the world-wide Internet). Indeed, during the several-day interval that large auroral displays were widely seen, strange effects were measured in telegraph systems across Europe – from Scandinavia to Tuscany. In the eastern United States, it was reported (Prescott, 1860) that on the telegraph line from Boston to Portland (Maine) during 'Friday, September 2nd, 1859 [the operators] continued to use the line [without batteries] for about two hours when, the aurora having subsided, the batteries were resumed.'

The early telegraph systems were also very vulnerable to atmospheric electrical disturbances in the form of thunderstorms, in addition to the 'anomalous' electrical currents flowing in the Earth. As written by Silliman (1850): 'One curious fact connected with the operation of the telegraph is the induction of atmospheric electricity upon the wires ... often to cause the machines at several stations to record the approach of a thunderstorm.' While disturbances by thunderstorms on the telegraph machines could be identified as to their source, the source(s) of the anomalous currents described by Barlow, and as recorded following Carrington's solar event, remained largely a mystery.

The decades that followed the solar event of 1859 prompted significant attention by telegraph engineers and operators to the effects on their systems of Earth's electrical currents. Although little recognized for almost fifty years afterwards, the Sun was indeed seriously affecting the first electrical technology that was employed for communications.

9.3 EARLY EFFECTS ON WIRELESS COMMUNICATIONS

Marconi demonstrated the feasibility of intercontinental wireless communications with his successful transmissions from Poldhu Station, Cornwall, to St John's, Newfoundland, in December 1901. Marconi's achievement (for which he shared the Nobel Prize in Physics with Karl Ferdinand Braun in 1909) was only possible because of the high-altitude reflecting layer, the ionosphere, which reflected the wireless signals. This reflecting layer was subsequently definitively identified by Briet and Tuve (1925) and by Appleton and Barnett (1925). Because wireless

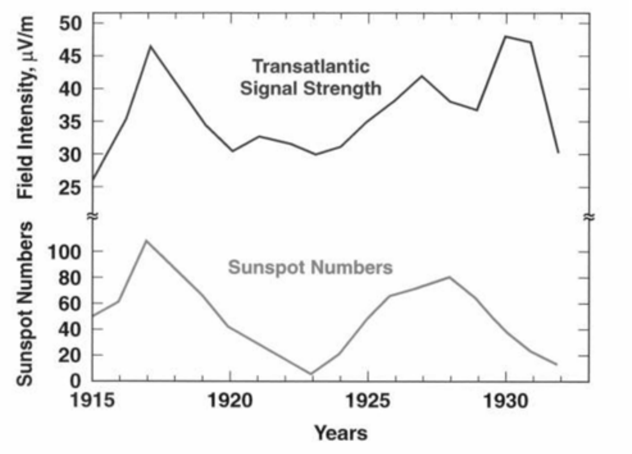


Figure 9.4. Yearly average daylight cross-Atlantic transmission signal strengths and monthly average sunspot numbers for the period 1915–1932. (Fagen, 1975.)

remained the only method for cross-oceanic voice (in contrast to telegraph) communications until the laying of the first transatlantic telecommunications cable, TAT-1 (Newfoundland to Scotland) in 1958, any physical changes in the radio wave-reflecting layer (even before it was 'discovered') were critical to the success (or failure) of reliable transmissions.

The same ionospheric electrical currents that could produce 'spontaneous' electrical currents within the Earth (and thus within the wires of the electrical telegraph) could also affect the reception and fidelity of the transmitted long-distance wireless signals. Indeed, Marconi (1928) commented on this phenomenon when he noted that 'times of bad fading [of radio signals] practically always coincide with the appearance of large sunspots and intense aurora boreali usually accompanied by magnetic storms.' These are 'the same periods when cables and land lines experience difficulties or are thrown out of action'.

An example of the types of studies that were pursued in the early years of long-distance wireless is shown in Figure 9.4. Plotted here (reproduced from Fagen, 1975, which contains historical notes on early wireless research in the old Bell Telephone System) are yearly average daylight cross-Atlantic transmission signal strengths for the years 1915–1932 (upper trace). The intensities in the signal strength curves were derived by averaging the values from about ten European stations that were broadcasting in the $\sim$15–23 kHz band (very long wave lengths), after reducing them to a common base (the signal from Nauen, Germany, was used as the base). Plotted in the lower trace of the figure are the monthly average sunspot numbers per year. Clearly, there is an association between the two plotted quantities, but the physical reason for such an association was very incompletely understood at the time. Nevertheless, this relationship of the received electrical field strengths to the yearly solar activity as represented by the number of sunspots could be used by wireless engineers to provide them some expectation as to transmission quality on a gross, year-to-year, basis – a very early form of prediction of space weather.

The relationship of disturbed long-wavelength radio transmissions and individual incidents of solar activity was first identified in 1923 (Anderson, 1928). The technical literature of the early wireless era showed clearly that solar-originating disturbances were serious assaults on the integrity of these communications during the first decades of the twentieth century. Communications engineers pursued a number of methodologies to alleviate or mitigate the assaults. One of these methodologies that sought more basic understanding is illustrated in the context of Figure 9.4. Another methodology utilized alternative wireless communications routes. As Figure 9.5 (see colour section) illustrates, for the radio electric field strength data recorded during a solar and subsequent geomagnetic disturbance on 8 July 1928 (day 0 on the horizontal axis), the transmissions at long wavelength were relatively undisturbed, while those at the shorter wavelength (16 m) were seriously degraded (Anderson, 1929). Such procedures are still employed by amateur and other radio operators.

The practical effects of the technical conclusions of Figure 9.5 are well exemplified by a headline which appeared over a front page article in the Sunday, 23 January 1938 issue of *The New York Times*. This headline noted that 'Violent magnetic storm disrupts short-wave radio communication'. The subheadline related that 'Transoceanic services transfer phone and other traffic to long wavelengths as sunspot disturbance strikes'. The technical manipulations that shifted the cross-Atlantic wireless traffic from short to longer wavelengths prevented the complete disruption of voice messages during the disturbance.

9.4 THE BEGINNING OF THE SPACE ERA

It should not have been a surprise, to those who may have considered the question, that the space environment (even before Van Allen's discovery) was not likely to be totally benign to technologies. Victor Hess, an Austrian, had demonstrated, from a series of balloon ascents during 1912, that cosmic rays originate outside the Earth's atmosphere. Many authors (for example, Chapman and Bartels, 1941; Cliver, 1994; Siscoe, 2005, for considerable historical perspective) had long discussed the possibility that charged particles, probably from the Sun, played a key role in producing the aurora and geomagnetic activity at Earth. Nevertheless, Van Allen's discovery, and the subsequent race to place instruments and humans in Earth orbit, spurred the need to study the new phenomena by the advent of rockets sent to very high altitudes.

Early in its existence, the US National Aeronautics and Space Administration (NASA, established in 1958) initiated programmes for examining the feasibility of satellite communications. This began with a contract with the Hughes Aircraft Corporation for geosynchronous (GEO) Syncom satellites (the first launched in February 1963) and a low-orbit communications programme (under the name Relay, the first of which was launched in December 1962). NASA also initiated an Applications Technology Satellite (ATS) programme (ultimately six satellites were launched into various orbits, but two were unsuccessful due to launch vehicle failures) to investigate and test technologies and concepts for a number of space applications. In addition to communications, applications included meteorology, navigation, and health delivery, although not all such topics were objectives for each spacecraft.

ATS-1 was launched into a geosynchronous orbit (GEO) in December 1966. Included in the payload were three separate instruments containing charged-particle detectors designed specifically to characterize the space environment at GEO. The three sectors of society – commercial (AT&T Bell Laboratories), military (Aerospace Corporation) and academic (University of Minnesota) – that constructed the three instruments demonstrated the wide-ranging institutional interest in, and scientific importance of, space weather conditions around Earth. The experiments all provided exciting data on such topics as the diurnal variation of the trapped radiation at the geosynchronous orbit (Lanzerotti *et al.*, 1967), the large changes in the radiation with geomagnetic activity (Paulikas *et al.*, 1968; Lezniak and Winckler, 1968), and the ready access of solar-produced particles to GEO (Lanzerotti, 1968; Paulikas and Blake, 1969).

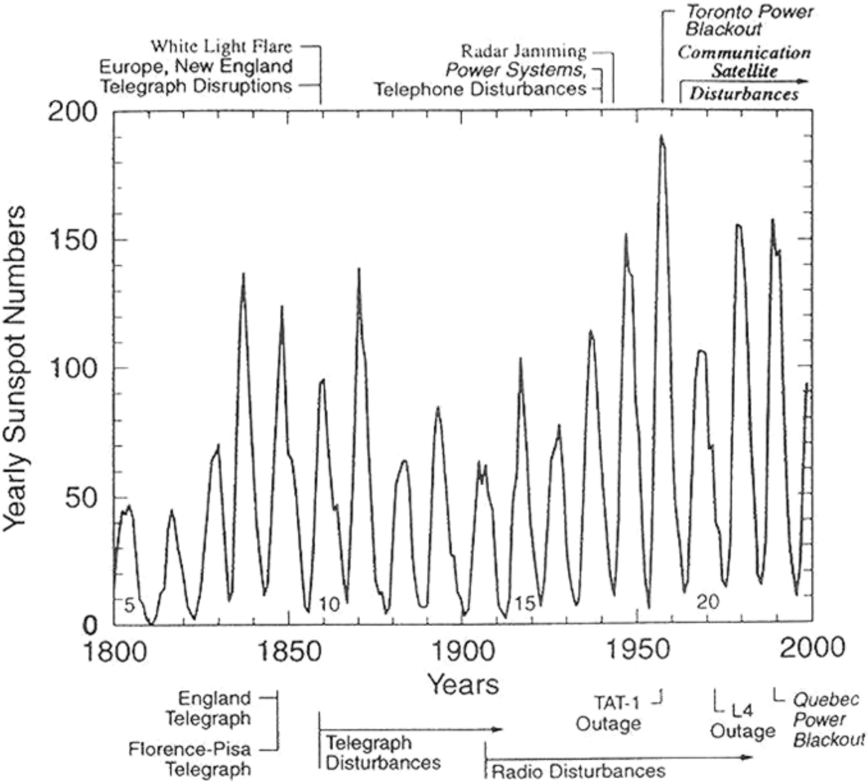


Figure 9.6. Yearly sunspot numbers with indicated times of selected major impacts of the solar–terrestrial environment on largely ground-based technical systems. The numbers just above the horizontal axis are the conventional numbers of the sunspot cycles.

Figure 9.6 shows the times of disturbances on selected communications systems following solar-originating disturbances. Four of the communications disturbances indicated in the figure occurred after the beginning of the space era. The magnetic storm of February 1958 disrupted voice communications on TAT-1, from Newfoundland to Scotland (and also plunged the Toronto region into darkness by the tripping of electrical power company circuits). The outage for nearly an hour of a major continental telecommunications cable (L4), stretching from near Chicago to the west coast, was disrupted between the Illinois and Iowa powering stations by the magnetic storm of August 1972 (Anderson *et al.*, 1974; Boteler and van Beek, 1999).

In March 1989 the entire province of Quebec suffered a power outage for nearly a day as major transformers failed under the onslaught of a large geomagnetic storm (Czech *et al.*, 1992). At the same time the first cross-Atlantic fibre-optic voice cable (TAT-8) was rendered nearly inoperative by the large potential difference that was established between the cable terminals on the coasts of New Jersey and England (Medford *et al.*, 1989).

Point-to-point high-frequency (HF) wireless communications links continue to be affected by ionospheric disturbances caused by solar-produced interactions with the Earth's space environment. Users of such systems are familiar with many anecdotes up to the present day of solar-produced effects and disruptions. For example, in 1979 (near the peak of the 21st solar cycle) a distress signal from a downed commuter plane was received by an Orange County, California, fire department – which responded, only to discover that the signal had originated from an accident site in West Virginia (*Los Angeles Times*, 1979). An Associated Press release posted on 30 October 2003 (during the declining phase of the 23rd solar cycle) noted that airplanes 'flying north of the 57th parallel experienced some disruptions in high-frequency radio communications . . . due to the geomagnetic storm from solar flares'.

As technologies have increased in sophistication, as well as in miniaturization and in interconnectedness, more sophisticated understanding of the Earth's space environment continues to be required. In addition, the increasing diversity of communications systems that can be affected by space weather processes is accompanied by continual changes in the dominance of use of one technology over another for specific applications. For example, in 1988 satellites were the dominant carrier of trans-ocean messages and data, but only about 2% of this traffic was over ocean cables. By 1990, the wide bandwidths provided by fibre-optic cable led to 80% of the trans-ocean traffic utilising ocean cables (Mandell, 2000).

9.5 SOLAR–TERRESTRIAL ENVIRONMENTAL EFFECTS ON COMMUNICATIONS TECHNOLOGIES

Many present-day communications technologies that include considerations of the solar–terrestrial environment in their designs and/or operations are listed in Table 9.1. Figure 9.7 (see colour section) schematically illustrates some of these effects.

Table 9.1. Impacts of solar–terrestrial processes on communications.

- Ionosphere Variations
  - Induction of electrical currents in the Earth
    - Long communications cables
  - Wireless signal reflection, propagation, attenuation
    - Commercial radio and television
    - Local and national safety and security entities
    - Aircraft communications
  - Communication satellite signal interference, scintillation
    - Commercial telecom and broadcast
- Magnetic Field Variations
  - Attitude control of communications spacecraft
- Solar Radio Bursts
  - Excess noise in wireless communications systems
  - Interference with radar and radio receivers
- Charged Particle Radiation
  - Solar cell damage
  - Semiconductor device damage and failure
  - Faulty operation of semiconductor devices
  - Spacecraft charging, surface and interior materials
  - Aircraft communications avionics
- Micrometeoroids and Artificial Space Debris
  - Spacecraft solar cell damage
  - Damage to surfaces, materials, complete vehicles
  - Attitude control of communications spacecraft
- Atmosphere
  - Drag on low-altitude communications satellites
  - Attenuation and scatter of wireless signals

9.5.1 Ionosphere and wireless

A century after Marconi's achievement, the ionosphere remains both a facilitator and a disturber in numerous communications applications. The military, as well as police and fire emergency agencies in many nations, continue to rely on wireless links that make extensive use of frequencies from kHz to hundreds of MHz and that use the ionosphere as a reflector. Commercial air traffic over the north polar regions continues to grow following the political changes of the late 1980s and early 1990s, and this traffic relies heavily on RF communications. Changes in the ionosphere that affect RF signal propagation can be produced by many mechanisms including direct solar photon emissions (solar UV and X-ray emissions), solar particles directly impacting polar region ionospheres, and radiation belt particles precipitated from the trapped radiation environment during geomagnetic storms.

At higher (a few GHz) frequencies the production of 'bubbles' in ionospheric densities in the equatorial regions of the Earth can be a prime source of scintillations in satellite-to-ground signals. Engineers at the COMSAT Corporation discovered these effects after the deployment of the INTELSAT network at geosynchronous orbit (Taur, 1973). This discovery is an excellent example of the surprises that the solar–terrestrial environment can hold for new technologies and for services that are based upon new technologies. A major applications satellite programme (C/NOFS), scheduled for launch in 2006, has been designed by the US Department of Defense to explicitly study the causes and evolutions of the processes that produce equatorial region bubbles, and to examine means of mitigation.

Disturbed ionospheric currents during geomagnetic storms can also be the cause of considerable problems at all geomagnetic latitudes in the use of navigation signals from the Earth-orbiting Global Positioning System (GPS), which provides precise location determination on Earth. These ionospheric perturbations limit the accuracy of positional determinations, thus presently placing limits on some uses of space-based navigation techniques for applications ranging from air traffic control to ship navigation to many national security considerations. The European Galileo Navigation Satellite System (GNSS) will also have to take into account ionospheric disturbances in order to ensure successful operations.

As evidenced by the initiation of the C/NOFS mission, there remain large uncertainties in the knowledge base of the processes that determine the initiation and scale sizes of the ionospheric irregularities that are responsible for the scintillation of radio communications signals that propagate through the ionosphere. Thus, it remains difficult to define mitigation techniques (including multi-frequency broadcasts and receptions) that might be applicable for receivers and/or space-based transmitters under many ionospheric conditions. Further and deeper knowledge from planned research programmes might ultimately yield clever mitigation strategies.

9.5.2 Ionosphere and Earth currents

The basic physical chain of events behind the production of large potential differences across the Earth's surface begins with greatly increased electrical currents flowing in the magnetosphere and the ionosphere. The temporal and spatial variations of these increased currents then cause large variations in the time-rate of change of the magnetic field as seen at Earth's surface. The time variations in the field in turn induce potential differences across large areas of the surface that are spanned by cable communications systems (or any other systems that are grounded to Earth, such as power grids and pipelines). Telecommunications cable systems use the Earth itself as a ground return for their circuits, and these cables thus provide highly conducting paths for concentrating the electrical currents that flow between these newly established, but temporary, Earth 'batteries'. The precise effects of these 'anomalous' electrical currents depend upon the technical system to which the long conductors are connected. In the case of long telecommunications lines, the Earth potentials can cause overruns of the compensating voltage swings that are designed into the power supplies (Anderson *et al.*, 1974) that are used to power the signal repeaters and regenerators (the latter in the case of optical transmissions).

Major issues can arise in understanding in detail the effects of enhanced space-induced ground electrical currents on cable systems. At present, the time variations and spatial dependencies of these currents are not well understood or predicable from one geomagnetic storm to the next. This is of considerable importance, since the induced Earth potentials are very much dependent upon the conductivity structure of the Earth underlying the affected ionospheric regions. Similar electrical current variations in the space/ionosphere environment can produce vastly different Earth potential drops depending upon the nature and orientation of underground Earth conductivity structures in relationship to the variable overhead currents.

Modelling of these effects is becoming advanced in many cases. This is an area of research that involves a close interplay between space plasma geophysics and solid Earth geophysics, and is one that is not often addressed collaboratively by these two very distinct research communities (except by the somewhat limited group of researchers who pursue electromagnetic investigations of the Earth).

9.5.3 Solar radio emissions

Solar radio noise and bursts were discovered more than six decades ago by Southworth (1945) and by Hey (1946) during the early research on radar at the time of the Second World War. Solar radio bursts produced unexpected (and initially unrecognized) jamming of this new technology that was under rapid development and deployment for war-time use for warnings of enemy aircraft (Hey, 1973). Extensive post-war research established that solar radio emissions can exhibit a wide range of spectral shapes and intensity levels (Kundu, 1965; Castelli *et al.*, 1973; Guidice and Castelli, 1975; Barron *et al.*, 1985), knowledge of which is crucial for determining the nature and severity of solar emissions on specific technologies such as radar, radio, satellite ground communications receivers, and civilian wireless communications. Research on solar radio phenomena remains an active and productive field of research today (Bastian *et al.*, 1998; Gary and Keller, 2004).

Some analyses of local noon solar radio noise levels that are routinely taken by the US Air Force, and made available by the NOAA World Data Center, have been carried out in order to assess the noise in the context of modern communications technologies. These analyses show that in 1991 (during the sunspot maximum interval of the 22nd cycle) the average noon fluxes measured at 1.145 GHz and at 15.4 GHz were -162.5 and $-156\,\mathrm{dBW/(m^2\,4\,kHz)}$, respectively (Lanzerotti *et al.*, 1999). These values are only about 6 dB and 12 dB above the 273° K (Earth's surface temperature) thermal noise of $-168.2\,\mathrm{dBW/(m^2\,4\,kHz)}$. Furthermore, these two values are only about 20 dB and 14 dB, respectively, below the maximum flux of $-142\,\mathrm{dBW/(m^2\,4\,kHz)}$ that is allowed for satellite downlinks by the International Telecommunications Union (ITU) regulation RR2566.

Solar radio bursts can have much larger intensities. As an example of an extreme event, that of 23 May 1967 produced a radio flux level (as measured at Earth) of $>10^5$ solar flux units ($1\,\mathrm{sfu} = 10^{-22}\,\mathrm{W/(m^2\,Hz)}$) at 1 GHz, and perhaps much larger

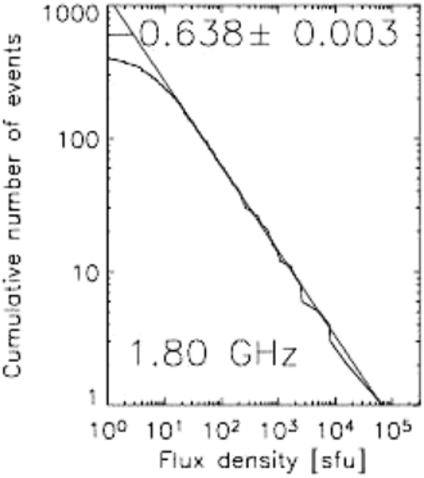


Figure 9.8. Cumulative distribution of intensities of 412 solar radio bursts in 2001–2002 at a frequency of 1.8 GHz at the NJIT Owens Valley Solar Array. (From Nita *et al.*, 2004.)

(Castelli *et al.*, 1973). Such an sfu level corresponds to $-129\,\mathrm{dBW/(m^2\,4\,kHz)}$, or 13 dB above the maximum limit of $-142\,\mathrm{dBW/(m^2\,4\,kHz)}$ noted above, and could cause considerable excess noise in any wireless cell site that might be pointed at the Sun at the time of the burst.

An example of a portion of a study of solar burst events that is directed towards understanding the distributions of events that might produce severe noise in radio receivers is shown in Figure 9.8 (Nita *et al.*, 2004). Plotted here is the cumulative distribution of intensities of 412 solar radio bursts measured in 2001–2002 (during the maximum of the 23rd solar cycle) at a frequency of 1.8 GHz at the NJIT Owens Valley Solar Array. The exponent of a power-law fit to the distribution is shown; the roll-over of the distribution at the lowest flux density is believed to be a result of decreased instrument sensitivities at the very lowest levels. Using such distributions, and taking into account the time interval over which the data were acquired, the probability of a burst affecting a specific receiver can be estimated. Bala *et al.* (2002), in an analysis of forty years of solar burst data assembled by the NOAA National Geophysical Data Center, estimated that bursts with amplitudes $>10^3$ solar flux units (sfu) at $f \sim 1\,\mathrm{GHz}$ could cause potential problems in a wireless cell site on average of once every three to four days during solar maximum, and perhaps once every twenty days or less during solar minimum.

Short-term variations often occur within solar radio bursts, with time variations ranging from several milliseconds to seconds and more (Benz, 1986; Isliker and Benz, 1994). Such short time variations can often be many tens of dB larger than the underlying solar burst intensities upon which they are superimposed. It would be useful to evaluate wireless systems in the context of such new scientific understanding.

9.5.4 Space radiation effects

As related in the Introduction, the discovery of the trapped radiation around Earth immediately implied that the space environment would not be benign for any communications technologies that might be placed within it. Some 200 or so in-use communications satellites now occupy the geosynchronous orbit. The charged particle radiation (over the entire range of energies) that permeates the Earth's space environment remains a difficult problem for the design and operations of these and other space-based systems (Shea and Smart, 1998; Koons *et al.*, 1999). A textbook discussion of the space environment and the implications for satellite design is contained in Tribble (1995).

The low-energy (few eV to few keV) plasma particles in the Earth's magnetospheric plasma can be highly variable in time and in intensity levels, and can produce different levels of surface charging on the materials (principally for thermal control) that encase a satellite (Garrett, 1981). If good electrical connections are not established between the various surface materials, and between the materials and the solar arrays, differential charging on the surfaces can produce lightning-like breakdown discharges between the materials. These discharges can produce electromagnetic interference and serious damage to components and subsystems (Vampola, 1987; Koons, 1980; Gussenhoven and Mullen, 1983).

Under conditions of enhanced geomagnetic activity, the cross-magnetosphere electric field will convect earthward the plasma sheet in the Earth's magnetotail. When this occurs, the plasma sheet will extend earthward to within the geosynchronous (GEO) spacecraft orbit. On such occasions, onboard anomalies from surface charging effects can occur: these tend to be most prevalent in the local midnight-to-dawn sector of the orbit (Mizera, 1983).

While some partial records of spacecraft anomalies exist, there are relatively few published data on the statistical characteristics of charging on spacecraft surfaces, especially from commercial satellites that are used so extensively for communications. Two surface-mounted charge-plate sensors were specifically flown on the former AT&T Telstar 4 GEO satellite to monitor surface charging effects. Figure 9.9 shows the statistical distributions of charging on one of the sensors in January 1997 (Lanzerotti *et al.*, 1998). The solid line in each panel corresponds to the charging statistics for the entire month, while the dashed lines omit data from a magnetic storm event on 10 January (statistics shown by the solid lines). Charging voltages as large as $-800\,\mathrm{V}$ were recorded on the charge-plate sensor during the magnetic storm – an event during which a permanent failure of the Telstar 401 satellite occurred (although the failure has not been officially attributed specifically to the space conditions).

The intensities of higher-energy particles in the magnetosphere (MeV energy protons and electrons to tens of MeV energy protons) can change by many orders of magnitude over the course of minutes, hours and days. These intensity increases occur through a variety of processes, including plasma physics energization processes in the magnetosphere and ready access of solar particles to GEO. Generally it is prohibitively expensive to provide sufficient shielding of all interior spacecraft sub-

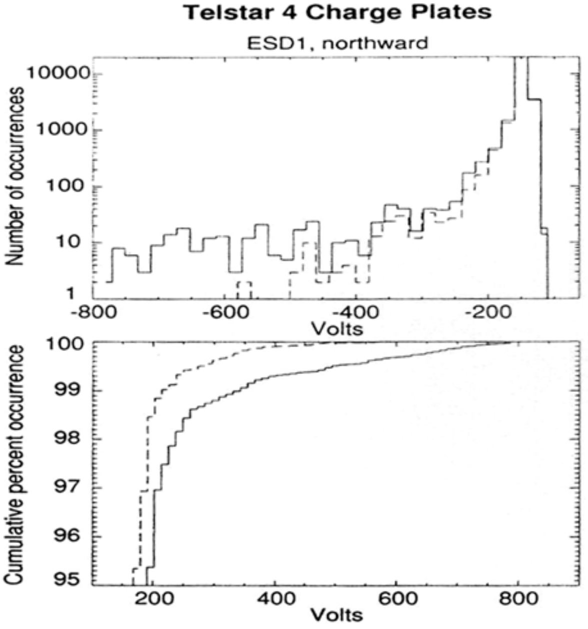


Figure 9.9. Statistical distribution of surface charging recorded on the northward-facing charge plate sensor on the Telstar 4 spacecraft during January 1997 (solid line) and for the same month with data from 10 January (the date of a large magnetic storm) removed (dashed line). The upper panel records (in approximately 25-V bins) the number of voltage occurrences in each voltage bin. The lower panel plots the cumulative percent voltage occurrence above 95% in order to illustrate the extreme events seen by the communications spacecraft.

systems against high-energy particles. Most often, increasing shielding would require a weight trade-off of the benefits of such shielding as compared to flying additional transponders or more orbit control gas, for example.

The range of a 100-MeV proton in aluminium (a typical spacecraft material) is ∼40 mm, while the range of a 3-MeV electron is ∼6 mm. These particles can therefore penetrate deeply into the interior regions of a satellite. In addition to producing transient upsets and latch-ups in signal and control electronics, such particles can also cause electrical charges to build up in interior insulating materials such as those used in coaxial cables. If the charge build-up in interior dielectric materials is sufficiently large, electrical breakdowns will ultimately result, and electromagnetic interference and damage to the electronics will occur.

A number of spacecraft anomalies, and even failures, have been identified as having occurred following many days of significantly elevated fluxes of several MeV energy electrons at GEO (Baker *et al.*, 1987, 1994, 1996; Reeves *et al.*, 1998; see Chapter 6). These enhanced fluxes occurred following sustained interplanetary disturbances called corotating interactions regions. The large solar flare and coronal mass ejection events of October–November 2003 produced anomalies on many spacecraft, as discussed by Barbieri and Mahmot (2004). An adaptation of their

Table 9.2. Summary of space weather impacts on selected spacecraft in October–November 2003. (Adapted from Barbieri and Mahmot, 2004.)

Spacecraft mission	Change in operation status	Electronic errors	Noisy housekeeping data	Solar array degradation	Change in orbit dynamics	High levels of accumulated radiation
Aqua	None	X				
Chandra	Instrument safed					X
CHIPS	Control loss	X				
Cluster	None			X		
Genesis	Auto safed	X				
GOES 9, 10	None		X			
ICESat	None	X				
INTEGRAL	Command safe					
Landsat 7	Instrument safed					
RHESSI	Abs. time seq. stop	X				
SOHO	Instrument safed			X		
Stardust	Auto safed	X				
TDRSS	None	X				
TRMM	Added delta V				X	
WIND	None			X		

listing of some of the affected satellites and the impacts is shown in Table 9.2. They note that, with the exception of the orbit change of the TRMM mission, all of the impacts were caused by 'solar energetic particles ... or similarly accelerated particles in geospace'. The purely communications satellites included in Table 9.2 – the NASA Tracking and Data Relay Satellite System (TDRSS) – suffered electronic errors during the interval of the solar-origin events.

No realistic shielding is possible for most communications systems in space that are under bombardment by galactic cosmic rays (energies ~ 1 GeV and greater). These very energetic particles can produce upsets and errors in spacecraft electronics (as well as in computer chips that are intended for use on Earth (IBM, 1996)). So-called ground-level solar particle events (order of GeV energy) can produce errors in the avionics and communications equipment of an aircraft that might be flying over the polar region at the time of the event.

The significant uncertainties in placing, and retaining, a communications spacecraft in a revenue-returning orbital location has led to a large business in risk insurance and reinsurance for one or more of the stages in a satellite's history. The loss of a spacecraft, or one or more transponders, from adverse space weather conditions is only one of many contingencies that can be insured against. In some years the space insurance industry is quite profitable, and in some years there are serious losses in net revenue after paying claims. For example, Todd (2000) states that in 1998 there were claims totalling more than \$1.71 billion after salvage – an amount just less than about twice that received in premiums. These numbers vary by large amounts from year to year.

9.5.5 Magnetic field variations

Enhanced solar wind flow velocities and densities, such as those that can occur in coronal mass ejection events, can easily distort the dayside magnetopause and push it inside its normal location at about ten Earth radii distance. During large solar wind disturbances, the magnetopause can be pushed inside the geosynchronous orbit. At such times, the magnetic field at GEO increases to as much as twice its quiescent value. In addition, the magnetic field outside the magnetopause will have a polarity that is predominantly opposite to that inside the magnetosphere.

The high variations in magnitude, space and time of magnetic fields that occur at the boundary and outside the magnetosphere can seriously disrupt the stabilization of any GEO satellite that uses the Earth's magnetic field for attitude control. Such magnetically stabilized GEO communications spacecraft must take into account the high probability that the satellite will on occasion, during a large magnetic disturbance, find itself near and even outside the magnetosphere on the sunward side of the Earth. Thus, appropriate GEO satellite attitude control designs must be implemented in order to cope with highly fluctuating magnetopause magnetic fields, and even the complete 'flipping' of the field when the magnetopause is crossed.

9.5.6 Micrometeoroids and space debris

The impacts on communications spacecraft of solid objects – such as micrometeoroids, and debris left in orbit from space launches and from satellites that break up for whatever reason – can seriously disorient a satellite and even cause a total loss (Beech *et al.*, 1997; McBride, 1997). The US Air Force systematically tracks thousands of items of space debris, most of which are in low-altitude orbits.

9.5.7 Atmosphere: low-altitude spacecraft drag

The ultraviolet emissions from the Sun change by more than a factor of two at wavelengths ≤ 170 nm during a solar cycle (Hunten *et al.*, 1991). This is significantly more than the $\sim 0.1\%$ changes that are typical of the visible radiation. The heating of the atmosphere by the increased solar UV radiation causes the atmosphere to expand. The heating is sufficient to raise the 'top' of the atmosphere by several hundred km during solar maximum. The greater densities at the higher altitudes result in increased drag on both space debris and on communications spacecraft in low Earth orbits (LEO). Telecommunications spacecraft that fly in LEO have to plan to use some amount of their orbit control fuel to maintain orbit altitude during the build-up to, and in, solar maximum conditions (Picholtz, 1996).

9.5.8 Atmosphere: water vapour

At frequencies in the Ka (18–31 GHz) band that are planned for high bandwidth space-to-ground applications (as well as for point-to-point communications between ground terminals), water vapour in the neutral atmosphere is the most significant natural phenomenon that can seriously affect the signals (Gordon and Morgan, 1993). It would appear that, in general, the space environment can reasonably be ignored when designing around the limitations imposed by rain and water vapour in the atmosphere.

A caveat to this claim would certainly arise if it were definitely to be shown that there are effects of magnetospheric and ionospheric processes (and thus effects of the interplanetary medium) on terrestrial weather. It is well recognized that even at GHz frequencies the ionized channels caused by lightning strokes, and possibly even charge separations in clouds, can reflect radar signals. Lightning and cloud charging phenomena may produce as yet unrecognized noise sources for low-level wireless signals. Thus, if it were to be learned that ionospheric electrical fields influenced the production of weather disturbances in the troposphere, the space environment could be said to affect even those wireless signals that might be disturbed by lightning. Much further research is required in this area of speculation.

9.6 SUMMARY

In the 150 years since the advent of the first electrical communication system – the electrical telegraph – the diversity of communications technologies that are embedded within space-affected environments have vastly increased. The increasing sophistication of these communications technologies, and how their installation and operations may relate to the environments in which they are embedded, requires ever more sophisticated understanding of natural physical phenomena. At the same time, the business environment for most present-day communications technologies that are affected by space phenomena is very dynamic. The commercial and national security deployment and use of these technologies do not wait for optimum knowledge of possible environmental effects to be acquired before new technological embodiments are created, implemented, and marketed. Indeed, those companies that might foolishly seek perfectionist understanding of natural effects can be left behind by the marketplace. A well-considered balance is needed between seeking ever deeper understanding of physical phenomena and implementing ‘engineering’ solutions to current crises. The research community must try to understand, and operate in, this dynamic environment.

9.7 ACKNOWLEDGEMENTS

This chapter relies heavily on past research and engineering conducted over nearly four decades at Bell Laboratories. Some of these activities are recorded in several of the papers referenced in the text, as well as in overviews presented; for example, in Lanzerotti (2001a, b). I also sincerely thank numerous colleagues for vigorous discussions on this topic over many years, including C. G. Maclennan, D. J. Thomson, G. Siscoe, J. H. Allen, J. B. Blake, G. A. Paulikas, A. Vampola, H. C. Koons and L. J. Zanetti.

9.8 REFERENCES

Anderson, C. N., Correlation of radio transmission and solar activity, *Proc. I. R. E.*, **16**, 297, 1928.

Anderson, C. N., Notes on the effects of solar disturbances on transatlantic radio transmissions, *Proc. I. R. E.*, **17**, 1528, 1929.

Anderson, C. W., III, L. J. Lanzerotti, and C. G. Maclennan, Outage of the L4 system and the geomagnetic disturbance of 4 August 1972, *The Bell Sys. Tech. J.*, **53**, 1817, 1974.

Appleton, E. V. and M. A. F. Barnett, Local reflection of wireless waves from the upper atmosphere, *Nature*, **115**, 333, 1925.

Baker, D. N., R. D. Balian, P. R. Higbie, *et al.*, Deep dielectric charging effects due to high energy electrons in Earth's outer magnetosphere, *J. Electrost.*, **20**, 3, 1987.

Baker, D. N., S. Kanekal, J. B. Blake, *et al.*, Satellite anomaly linked to electron increase in the magnetosphere, *Eos Trans. Am. Geophys. Union*, **75**, 401, 1994.

Baker, D. N., An assessment of space environment conditions during the recent Anik E1 spacecraft operational failure, *ISTP Newsletter*, **6**, 8, 1996.

Baker, D. N., J. H. Allen, S. G. Kanekal, and G. D. Reeves, Disturbed space environment may have been related to pager satellite failure, *Eos Trans. Am. Geophys. Union*, **79**, 477, 1998.

Bala, B., L. J. Lanzerotti, D. E. Gary, and D. J. Thomson, Noise in wireless systems produced by solar radio bursts, *Radio Sci.*, **37**(2), 10.1029/2001RS002488, 2002.

Barbieri, L. P., and R. E. Mahmot, October–November 2003's space weather and operations lessons learned, *Space Weather*, **2**, No. 9, S0900210.1029/2004SW000064, 2004.

Barlow, W. H., On the spontaneous electrical currents observed in the wires of the electric telegraph, *Phil. Trans. R. Soc.*, **61A**, 61, 1849.

Barron, W. R., E. W. Cliver, J. P. Cronin, and D. A. Guidice, Solar radio emission, in *Handbook of Geophysics and the Space Environment*, ed. A. S. Jura, Chap. 11, AFGL, USAF, 1985.

Bastian, T. S., A. O. Benz, and D. E. Gary, Radio emission from solar flares, *Ann. Rev. Astron. Astrophys.*, **36**, 131, 1998.

Beech, M., P. Brown, J. Jones, and A. R. Webster, The danger to satellites from meteor storms, *Adv. Space Res.*, **20**, 1509, 1997.

Bell System Technical Journal, Special Telstar Issue, **42**, Parts 1, 2, 3, July 1963.

Benz, A. O., Millisecond radio spikes, *Solar Phys.*, **104**, 99, 1986.

Boteler, D. H., and G. Jansen van Beek, August 4, 1972 revisited: A new look at the geomagnetic disturbance that caused the L4 cable system outage, *Geophys. Res. Lett.*, **26**, 577, 1999.

Breit, M. A., and M. A. Tuve, A test of the existence of the conducting layer, *Nature*, **116**, 357, 1925.

Carrington, R. C., *Observation of the Spots on the Sun from November 9, 1853, to March 24, 1863, Made at Redhill*, William and Norgate, London and Edinburgh, 167, 1863.

Castelli, J. P., J. Aarons, D. A., Guidice, and R. M. Straka, The solar radio patrol network of the USAF and its application, *Proc. IEEE*, **61**, 1307, 1973.

Chapman, S., and J. Bartels, *Geomagnetism*, 2 vols, Oxford University Press, 1940.

Clark, Arthur C., Extra-Terrestrial Relays – Can Rocket Stations Give World-Wide Radio Coverage? *Wireless World*, 305, October 1945.

Cliver, E. W., Solar activity and geomagnetic storms, *Eos Trans. Am. Geophys. Union*, **75**, 569, 1994.

Czech, P., S. Chano, H. Huynh, and A. Dutil, The Hydro-Quebec system blackout of 13 March 1989: System response to geomagnetic disturbance, *Proc. EPRI Conf. Geomagnetically Induced Currents*, EPRI TR-100450, Burlingame, CA, 19, 1992.

Fagen, M. D., *A History of Science and Engineering in the Bell System*, Bell Tel. Labs., Inc., Murray Hill, NJ, 1975.

Garrett, H. B., The charging of spacecraft surfaces, *Revs. Geophys.*, **19**, 577, 1981.

Gary, D. E., and C. U. Keller, eds., *Solar and Space Weather Radiophysics*, Springer, Heidelberg, 2004.

Gordon, G. D., and W. L. Morgan, *Principals of Communications Satellites*, John Wiley, New York, 178–192, 1993.

Guidice, D. A., and J. P. Castelli, Spectral characteristics of microwave bursts, in *Proc. NASA Symp. High Energy Phenomena on the Sun*, Goddard Space Flight Center, Greenbelt, MD, 1972.

Gussenhoven, M. S., and E. G. Mullen, Geosynchronous environment for severe spacecraft charging, *J. Spacecraft Rockets*, **20**, 26, 1983.

Hey, J. S., Solar radiations in the 4–6-metre radio wavelength band, *Nature*, **158**, 234, 1946.

Hey, J. S., *The Evolution of Radio Astronomy*, Neale Watson Academic Pub. Inc., New York, 1973.

Hunten, D. M., J.-C. Gerard, and L. M. Francois, The atmosphere's response to solar irradiation, in *The Sun in Time*, ed. C. P. Sonett, M. S. Giampapa, and M. S. Matthews, Univ. Arizona Press, Tucson, 463, 1991.

IBM Journal of Research and Development, **40**, 1–136, 1996.

Isliker, H., and A. O. Benz, Catalogue of 1–3-GHz solar flare radio emission, *Astron. Astrophys. Suppl. Ser.*, **104**, 145, 1994.

Koons, H. C., Characteristics of electrical discharges on the P78-2 satellite (SCATHA), *18th Aerospace Sciences Meeting*, AIAA 80-0334, Pasadena, CA, 1980.

Koons, H., C., J. E. Mazur, R. S. Selesnick, J. B. Blake, J. F. Fennel, J. L. Roeder, and P. C. Anderson, *The Impact of the Space Environment on Space Systems*, Engineering and Technology Group, The Aerospace Corp., Report TR-99(1670), El Segundo, CA, 1999.

Kundu, M. R., *Solar Radio Astronomy*, Interscience, New York, 1965.

Lanzerotti, L. J., C. S. Roberts, and W. L. Brown, Temporal Variations in the Electron Flux at Synchronous Altitudes, *J. Geophys. Res.*, **72**, 5893, 1967.

Lanzerotti, L. J., Penetration of solar protons and alphas to the geomagnetic equator, *Phys. Rev. Lett.*, **21**, 929, 1968.

Lanzerotti, L. J., C. Breglia, D. W. Maurer, and C. G. Maclennan, Studies of spacecraft charging on a geosynchronous telecommunications satellite, *Adv. Space Res.*, **22**, 79, 1998.

Lanzerotti, L. J., C. G. Maclennan, and D. J. Thomson, Engineering issues in space weather, in *Modern Radio Science*, ed. M. A. Stuchly, Oxford, 25, 1999.

Lanzerotti, L. J., Space weather effects on technologies, in *Space Weather*, ed. P. Song, H. J. Singer and G. L. Siscoe, American Geophysical Union, Washington, 11, 2001a.

Lanzerotti, L. J., Space weather effects on communications, in *Space Storms and Space Weather Hazards*, ed. I. A. Daglis, Kluwer, Holland, 313, 2001b.

Lezniak, T. W. and J. R. Winckler, Structure of the Magnetopause at 6.6 R_E in Terms of 50- to 150-kev Electrons, *J. Geophys. Res.*, **73**, 5733, 1968.

Los Angeles Times, Sunspots playing tricks with radios, Metro Section, p. 1, 13 February 1979.

Mandell, M., 120,000 leagues under the sea, *IEEE Spectrum*, 50, April 2000.

Marconi, G., Radio communication, *Proc. IRE*, **16**, 40, 1928.

McBride, N., The importance of the annual meteoroid streams to spacecraft and their detectors, *Adv. Space Res.*, **20**, 1513, 1997.

Medford, L. V., L. J. Lanzerotti, J. S. Kraus, and C. G. Maclennan, Trans-Atlantic Earth potential variations during the March 1989 magnetic storms, *Geophys. Res. Lett.*, **16**, 1145, 1989.

Mizera, P. F., A summary of spacecraft charging results, *J. Spacecraft Rockets*, **20**, 438, 1983.

Nita, G. M., D. E. Gary, and L. J. Lanzerotti, Statistics of solar microwave burst spectra with implications for operations of microwave radio systems, *Space Weather*, **2**, S11005, doi:10.1029/2004SW000090, 2004.

Paulikas, G. A., J. B. Blake, S. C. Freden, and S. S. Imamoto, Boundary of Energetic Electrons during the January 13–14, 1967, Magnetic Storm, *J. Geophys. Res.*, **73**, 5743, 1968.

Paulikas, G. A. and J. B. Blake, Penetration of Solar Protons to Synchronous Altitude, *J. Geophys. Res.*, **74**, 2162, 1969.

Picholtz, R. L., Communications by means of low Earth orbiting satellites, in *Modern Radio Science 1996*, ed. J. Hamlin, Oxford University Press, 133, 1996.

Pierce, John R., Orbital Radio Relays, *Jet Propulsion*, p. 153, April 1955.

Prescott, G. B., *Theory and Practice of the Electric Telegraph*, fourth edn., Tichnor and Fields, Boston, 1860.

Reeves, G. D., The relativistic electron response at geosynchronous orbit during January 1997 magnetic storm, *J. Geophys. Res.*, **103**, 17559, 1998.

Reid, E. J., How can we repair an orbiting satellite?, in *Satellite Communications Physics*, ed. R. M. Foster, Bell Telephone Laboratories, 78, 1963.

Shea, M. A., and D. F. Smart, Space weather: The effects on operations in space, *Adv. Space Res.*, **22**, 29, 1998.

Silliman, Jr., B., *First Principals of Chemistry*, Peck and Bliss, Philadelphia, 1850.

Siscoe, G. L., Space weather forecasting historically through the lens of meteorology, in *Space Weather, Physics and Effects*, ed. V. Bothmer and I. A. Daglis, Springer–Praxis, 2006.

Southworth, G. C., Microwave radiation from the sun, *J. Franklin Inst.*, **239**, 285, 1945.

Taur, R. R., Ionospheric scintillation at 4 and 6 GHz, *COMSAT Technical Review*, **3**, 145, 1973.

Todd, D., Letter to *Space News*, p. 12, 6 March 2000.

Tribble, A. C., *The Space Environment, Implications for Spacecraft Design*, Princeton University Press, Princeton, NJ, 1995.

Vampola, A., The aerospace environment at high altitudes and its implications for spacecraft charging and communications, *J. Electrost.*, **20**, 21, 1987.

Van Allen, J. A., *Origins of Magnetospheric Physics*, Smithsonian Institution, Washington, 1983.

Vernov, S. N., and A. E. Chudakov, Terrestrial corpuscular radiation and cosmic rays, *Adv. Space Res.*, **1**, 751, 1960.

10

Space weather effects on power grids

Risto Pirjola

10.1 INTRODUCTION

The solar wind formed by charged particles emitted by the Sun interacts with the geomagnetic field, producing the comet-like magnetosphere around the Earth. The magnetosphere is in a complex plasmaphysical coupling with the ionosphere. During a space weather storm the magnetosphere–ionosphere system becomes highly disturbed, containing intense and rapidly-varying currents. Regarding effects on the ground: particularly important are auroral electrojets with accompanying currents in the high-latitude ionosphere. At the Earth's surface, varying space currents manifest themselves as disturbances or storms in the geomagnetic field, and as expressed by Faraday's law of induction, a geoelectric field is also induced. The electric field produces currents, known as geomagnetically induced currents (GIC), in ground-based technological networks, such as electric power transmission systems, oil and gas pipelines, telecommunication cables and railway equipment. GIC thus constitute the ground end of the space weather chain. In power grids, GIC can saturate transformers, which may cause different problems ranging from harmonics in the electricity to a blackout of the whole system and permanent damage of transformers. In pipelines, GIC can enhance corrosion and interfere with protection systems and control surveys. Telecommunication equipment and railways may suffer from overvoltages due to GIC, possibly leading to failures.

The first observations of space weather effects on technological systems were made in early telegraph devices about 150 years ago (Prescott, 1866; Boteler *et al.*, 1998) (see Figure 10.1, colour section). At times the systems became completely inoperative, and at other times the operators were able to work without batteries by using space-weather-induced voltages. The situation between impossible and favourable operation conditions, of course, depended whether the induced voltage was in the same or in the opposite direction with the operational voltage. Today's optical-fibre cables are not directly affected by space weather, but GIC may flow in metallic wires used for powering repeaters. Trans-oceanic submarine communication cables are a special category, since the distances imply large end-to-end voltages (Lanzerotti *et al.*, 1995; Chapter 9).

Statistically GIC phenomena, as well as other space weather processes, follow the eleven-year sunspot cycle (Figure 10.1, colour section). However, large storms can also occur during sunspot minima, so concern about possible GIC problems should be continuous.

GIC problems in power grids, which result from saturation of transformers by the dc-like GIC, were fully realized in North America during the large magnetic storm in March 1940 (Davidson, 1940). The most famous and serious GIC event occurred in the Hydro-Québec power system in Canada in March 1989, resulting in a province-wide blackout that lasted nine hours, with many economic and social impacts (Kappenman and Albertson, 1990; Czech *et al.*, 1992; Bolduc, 2002). During the same magnetic storm, a transformer in New Jersey, USA, suffered heating damage and had to be replaced. A recent harmful event produced by GIC was the blackout in southern Sweden in October 2003 (Pulkkinen *et al.*, 2005).

Corrosion problems in pipelines caused by pipe-to-soil voltages associated with GIC became famous in connection with the construction of the long oil pipeline across Alaska in the 1970s (Campbell, 1978). GIC in pipelines have been investigated thoroughly in an international several-year study carried out some years ago (Boteler, 2000). Research on GIC impacts on railway equipment has not been very extensive so far. A well-documented event occurred in Sweden in July 1982 when railway traffic signals unexpectedly turned red (Wallerius, 1982). This was due to the fact that a voltage accompanying GIC made the automatic safety equipment interpret that a train was short-circuiting the rails. It is probable that railways, as well as other technological systems, have suffered from GIC problems in the past more often than documented, but the actual reason for the failures has remained obscure due to poor knowledge of space weather.

The significance of GIC is evidently increasing as modern society is becoming more and more dependent on reliable technology. Since GIC are mainly (but not only) a high-latitude phenomenon, most of the research of the topic is carried out in North America and in Nordic countries (Kappenman and Albertson, 1990; Elovaara *et al.*, 1992; Viljanen and Pirjola, 1994; Petschek and Feero, 1997; Bolduc *et al.*, 1998; Bolduc *et al.*, 2000; Boteler, 2000; Bolduc, 2002; Molinski, 2002; Pulkkinen, 2003). Problems due to GIC may be avoided by trying to block the flow of GIC in the network or by aiming at a GIC-insensitive design of systems whenever possible. Another approach to minimise GIC harm is to develop forecasting techniques of space weather events.

The horizontal geoelectric field at the Earth's surface is the key physical quantity which should be known to be able to calculate, estimate or forecast GIC magnitudes in a system. The 'complex image method' (CIM) has been shown to be an applicable technique for determining the geoelectric field, since it enables fast and accurate computations (Boteler and Pirjola, 1998a; Pirjola and Viljanen, 1998). Recent studies by Viljanen *et al.* (2004), however, indicate that the simple 'plane wave method' obviously provides the best tool for practical GIC computation purposes.

As concerns the determination of GIC produced by the geoelectric field, discretely-earthed networks (power systems) and continuously-earthed systems (buried pipelines) require different treatments (Lehtinen and Pirjola, 1985; Pulkkinen *et al.*, 2001b).

In Section 10.2, we summarize the effects GIC may have on power systems. Modelling techniques of the geoelectric field and of GIC in a power network are discussed in Section 10.3, and Section 10.4 is devoted to research on GIC carried out in Finland for almost thirty years.

10.2 GIC PROBLEMS IN POWER SYSTEMS

The basic physical principle of the flow of geomagnetically-induced currents in a technological network is easy to understand based on fundamental electromagnetic laws. A space weather storm produces intense and rapidly varying currents in the magnetosphere and ionosphere, which, as described by the Biot-Savart law, cause time-dependent magnetic fields seen as geomagnetic disturbances or storms. As expressed by Faraday's law of induction, a time variation of the magnetic field is always accompanied by an electric field. The geomagnetic disturbance and the geoelectric field observed at the Earth's surface not only depend on the (primary) space currents but are also affected by (secondary) currents driven by the electric field within the conducting Earth (Ohm's law). In particular for the electric field, the secondary contribution is essential. The horizontal geoelectric field drives ohmic currents (GIC) in technological conductor networks. Their flows obey Kirchhoff's laws.

Characteristic times of geomagnetic variations range from seconds and minutes to hours and days, and 1 Hz may be regarded as the upper limit of frequencies important regarding geoelectromagnetic fields and GIC. Thus, GIC flowing in electric power transmission systems are dc-like currents in comparison with the 50/60-Hz frequency used for the electricity. Figure 10.2 schematically shows the paths which GIC have when entering a three-phase power grid from the ground. The figure depicts three different types of transformer: a 'delta' transformer, not earthed at all; a 'normal' transformer, with an earthed neutral; and an 'autotransformer', containing common windings for the higher and lower voltages. Due to symmetry, GIC are equally divided between the three phases.

In normal conditions, the ac exciting current needed to provide the magnetic flux for the voltage transformation in a power transformer is only a few amperes, and the transformer operates within the range where the dependence of the flux on the exciting current is linear (Kappenman and Albertson, 1990; Molinski, 2002). However, the presence of a dc-like GIC produces an offset to the non-linear part of the operation curve. This means that the transformer is saturated during one half of an ac cycle and the exciting current becomes very large (even some hundreds of amperes). Besides, the exciting current is asymmetric with respect to the ac half-cycles, and is thus distorted by even and odd harmonics. The resulting

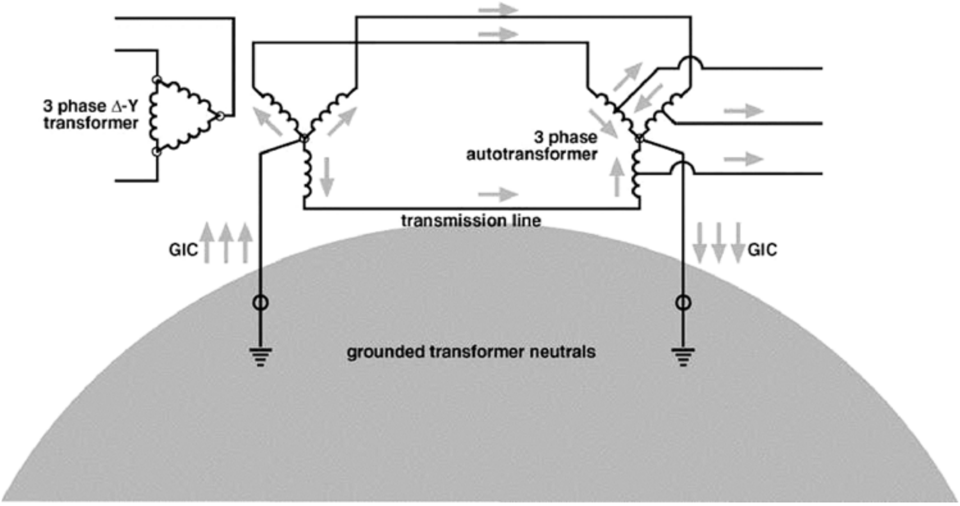


Figure 10.2. Schematic diagram of the flow of GIC in a three-phase power system. (Molinski, 2002.)

problems – which may constitute a fast cascading chain of processes in a power system – can be listed as follows:

- Production of harmonics in the electricity.
- Unnecessary relay trippings.
- Increased reactive power demands.
- Voltage fluctuations.
- Unbalanced network; even a collapse of the entire system.
- Magnetic stray fluxes in transformers.
- Hot spots in transformers; even permanent damage.

The best-known GIC failure occurred in the Hydro-Québec power system on 13 March 1989, at about 2.45 am local time (Kappenman and Albertson, 1990; Czech *et al.*, 1992; Bolduc, 2002). The problems were started by harmonics created by transformer saturation due to GIC. The harmonics affected static voltampere reactive compensators, which should rapidly regulate the voltage in order to keep the system stable. Activation of protective systems tripped seven compensators with a consequence that a generated power of 9,500 MW – 44% of Québec's total power consumption at the particular time – lost the voltage regulation. Combined with increased reactive power demands, serious voltage problems resulted, which implied that a 735-kV line was tripped, interrupting the 9,500-MW generation entirely. The frequency and the voltage decreased in the rest of the system. There was also a great imbalance between the load connected to the Hydro-Québec system and the generated power. All this caused the whole network to collapse, and a major part of Québec and six million people experienced a power blackout. In total, 21,500 MW of load and generation were lost. These cascading phenomena occurred in some tens of seconds, and the time between the onset of the magnetic storm and the collapse of the network was not more than about one-and-a-half minutes. After nine hours, 17% of the load was still out of service. The estimated costs of material damage amounted to 6.5 million Canadian dollars (CAD), and the net costs of the failure to 13.2 million CAD (Bolduc, 2002). Furthermore, a long blackout also creates many consequences with harmful impacts that cannot be measured in monetary units. The March 1989 storm created a thorough and intensive GIC investigation in the Hydro-Québec system, and several protective and corrective measures against GIC have been taken, so that today the system is considered safe (Bolduc and Langlois, 1995; Bolduc *et al.*, 1998; Bolduc *et al.*, 2000; Bolduc, 2002).

During the same geomagnetic storm in March 1989, overheating produced by GIC permanently damaged a transformer in New Jersey, USA, causing a cost of several million US dollars together with replacement energy costs of roughly $400,000 per day (Kappenman and Albertson, 1990). It has been stated that the total costs of a GIC failure in the north-eastern USA, during a slightly more severe storm than that of March 1989, would be $3–6 billion (Kappenman, 1996).

GIC problems also occurred in the UK during the March 1989 storm, as reported by Erinmez *et al.* (2002a, b). In addition, they list the storms in July 1982, October 1989 and November 1991 as times when GIC disturbed the UK power system significantly.

Pulkkinen *et al.* (2005) provide a detailed physical and technological summary about the power blackout produced by GIC in southern Sweden on 30 October 2003, leaving 50,000 people without electricity for 20–50 minutes.

The March 1989 space weather storm with its GIC consequences, caused a significant increase in GIC research efforts – especially in North America, where a project called SUNBURST was started to understand more about space weather and GIC processes and to prepare techniques for mitigating GIC problems (Petschek and Feero, 1997; Molinski, 2002).

As GIC are practically dc currents, a simple means to block their flow (but enable the 50/60-Hz currents) would be the installation of series capacitors in transmission lines or in earthing leads of transformer neutrals. However, the use of capacitors is not a cheap solution. The design of appropriate capacitors is not straightforward, as they should not disturb the operation of the system or decrease the level of safety. Furthermore, by considering the effect of blocking capacitors on GIC in the entire high-voltage power grid of England and Wales, Erinmez *et al.* (2002a) show that, contrary to what might be expected, the capacitors may even increase average GIC flowing through transformers. The same conclusion is drawn from an investigation of an idealized system (Pirjola, 2002a). Another study about the effect of neutral point reactors, which imply additional resistances in the earthing leads of transformer neutrals, on GIC in the Finnish 400-kV system, supports the observation that increasing the resistance experienced by GIC, thus decreasing GIC, at some sites tends to increase GIC at other sites, so that the overall situation may get worse (Pirjola, 2005b). Recent theoretical investigations about the effects of neutral-point reactors and series capacitors on GIC magnitudes also emphasise the complexity of the whole problem (Arajärvi *et al.*, 2006).

(It should be noted that neutral-point reactors and series capacitors are used in power systems for reasons other than GIC. The former are applied to increase the 50/60-Hz impedance, and thus decrease possible Earth-fault currents, which is important for safety; whereas the latter decrease the reactances, so the lines become 'electrically shorter', and the ability of a stable transmission of electric energy is improved.)

Attempts to avoid GIC harm also include the development of forecasting and warning methods of large prospective GIC events (Kappenman *et al.*, 2000). Recommendations for actions to be taken when a GIC event occurs or is forecasted are given by Molinski (2002) and by Bolduc (2002). The actions contain, for example, restoring out-of-service lines, stopping maintenance works, adjusting protective relay settings and reducing key transformer loadings. Procedures, whose purpose is to mitigate possible GIC impacts, are often complicated and expensive, so the forecasting should be reliable and without too many unnecessary alerts. System design aspects – such as trying to choose appropriate transformer types whenever possible – also provide countermeasures against GIC problems.

10.3 MODELLING OF GIC IN A POWER SYSTEM

The appearance of GIC in a power system depicted in Figure 10.2 can be interpreted by stating that the presence of the geoelectric field implies a voltage between the groundings of two transformer neutrals, and so a current will flow. This can also be equivalently understood by stating that the transformers and the power line between them provide a by-pass for currents flowing within the Earth. It is, however, important to note that GIC may also occur in systems not grounded at all. This is due to the fact that, in general, the geoelectric field is rotational, thus producing currents in horizontal loops. The rotational nature also means that there is no single-valued geopotential (or Earth-surface potential), and so the voltages obtained by integrating the geoelectric field are path-dependent (Boteler and Pirjola, 1998b; Pirjola, 2000).

A theoretical calculation of GIC in a technological system is conveniently divided into two separate steps:

(1) The determination of the horizontal geoelectric field at the Earth's surface.
(2) The computation of GIC driven by this field.

The first – the 'geophysical step' – is the same for all networks; whereas the second – the 'engineering step' – depends on the type of network. Generally, the geophysical step is more difficult, since the input data – the ionospheric–magnetospheric currents as functions of space and time, and the Earth's conductivity distribution – are poorly known, and certainly so complicated that an exact calculation of the geoelectric field is impossible in practice. The engineering step is basically a straightforward application of Ohm's law, Kirchhoff's laws, and Thévenin's theorem. The fact that geoelectromagnetic phenomena are slow, permitting a dc treatment (at least as the first approximation) simplifies the engineering step.

10.3.1 Calculation of the geoelectric field

The calculation of the geoelectric field requires information or assumptions about the current distribution in the magnetosphere and ionosphere produced by a space storm, and about the Earth's conductivity structure. The ionospheric–magnetospheric currents create the primary contributions to the electric and magnetic fields occurring at the Earth's surface. The ground conductivity distribution affects the currents driven by the electric field within the Earth. These currents and the accumulating charges create a secondary contribution to the surface fields. Because the Earth is a good conductor, the primary and secondary parts of the horizontal electric field are almost equal but opposite in sign at the surface. Thus, it is important to take into account both contributions to the geoelectric field, and their numerical computations have to be accurate enough to obtain the total horizontal field correctly. In general, GIC are a regional phenomenon, and so the Earth can be treated as a half-space with a flat surface without referring to global spherical calculations (even though the curvature of the Earth is indicated in the schematic diagram in Figure 10.2). Usually in geoelectromagnetic studies, the Earth's surface is the xy plane of a Cartesian coordinate system, with the z axis pointing downwards and x and y being (geographically) northwards and eastwards, respectively.

The simplest model for determining the geoelectric field from magnetic data at the Earth's surface utilises the assumptions that the primary field originating from space currents is a plane wave propagating vertically downwards, and that the Earth is uniform (Pirjola, 1982). Considering a single angular frequency ω, a horizontal electric field component E_y can be expressed in terms of the perpendicular horizontal magnetic field component B_x:

$$E_y = -\sqrt{\frac{\omega}{\mu_0 \sigma}} e^{i\frac{\pi}{4}} B_x \tag{10.1}$$

where μ_0 is the vacuum (and Earth) permeability, and σ is the conductivity of the Earth. Equation (10.1) presumes that the displacement currents can be neglected, which is acceptable in geoelectromagnetics because the relevant frequencies are very small (below 1 Hz). The ratio of the electric field to the magnetic field increases with an increasing frequency and with a decreasing conductivity. Equation (10.1) is the 'basic equation of magnetotellurics', since it indicates how electric (or 'telluric') and magnetic data measured at the surface can be used to determine the Earth's conductivity. This equation may be directly generalized to the case of a layered Earth, and σ is then replaced by a frequency-dependent 'apparent conductivity'.

Inverse Fourier transforming equation (10.1) to the time (t) domain gives

$$E(t) = -\frac{1}{\sqrt{\pi \mu_0 \sigma}} \int_0^{\infty} \frac{g(t-u)}{\sqrt{u}} \, du \tag{10.2}$$

where the subscript 'y' has been omitted from the electric field and the time derivative of $B_x(t)$ is denoted by $g(t)$. Equation (10.2) provides a simple means to estimate the electric field by using magnetic data. It can be seen that $E(t)$ is affected by past values of $g(t)$ but that their weight decreases with time (the square-root factor in the denominator).

The plane wave equations (10.1) and (10.2) do not strictly require that the primary field is a vertically-propagating plane wave. It is sufficient that the primary field does not vary much in the horizontal direction, which much widens the applicability of the equations. The validity of the plane wave model, however, becomes questionable at high latitudes due to the vicinity of localized ionospheric primary currents. In magnetotelluric studies this is associated with the so-called source effect distortion (Pirjola, 1992). The conductivity of a real Earth changes both vertically and horizontally. In a mathematical treatment the former is much more easily taken into account than the latter. Recently, Viljanen *et al.* (2004) have proved that a 'local' application of the plane wave relation between electric and magnetic fields at the Earth's surface is sufficiently accurate for practical GIC computation purposes.

Häkkinen and Pirjola (1986) derive exact equations for the surface electric and magnetic fields created by a general current system model above a layered Earth. The derivation is based on a straightforward use of Maxwell's equations and boundary conditions. The current system model is three-dimensional, and consists of a horizontal sheet current distribution, which may have any time and space dependencies, and of geomagnetic-field-aligned currents (FAC). The direction of the FAC is given by the inclination (I) and declination (D) angles, and the magnitudes of the FAC are adjusted to carry the charges possibly accumulated by the sheet current, so the total current is non-divergent at each time and point. The equations for the surface fields are, however, laborious in practical numerical computations (Pirjola and Häkkinen, 1991), and are therefore not suitable for time-critical purposes such as GIC forecasting.

The complex image method (CIM) already discussed by Wait and Spies (1969) and introduced by Thomson and Weaver (1975) to geophysical applications offers a good alternative of an exact (but laborious) computation of the surface electric and magnetic fields (Boteler and Pirjola, 1998a; Pirjola and Viljanen, 1998). CIM appears to be accurate enough, and it also enables fast computations. Furthermore, it should be noted here that the exact equations also lead to approximate field values, because the models of space currents and of the Earth's structure are necessarily idealizations of the real situation.

The basic idea of CIM is to calculate the secondary contribution (due to Earth currents) to the surface fields by fictitiously replacing the real (layered) Earth by a perfect conductor which lies at $z = p$ with the 'complex skin depth' p defined by

$$p = \frac{Z}{i\omega\mu_0} \tag{10.3}$$

The complex skin depth is a function of the angular frequency ω. The (plane wave) surface impedance $Z = Z(\omega)$ is determined by the Earth's real conductivity structure. Knowing p, the Earth's contribution to the surface fields is obtained by setting an image of the primary space current source at the (complex) mirror location. The advantage is that the secondary fields have mathematical expressions similar to those of the fields created by the primary source. Since CIM operates in the complex space it should just be regarded as a mathematical concept. However, the real and imaginary parts of p also have physical interpretations, as they represent the central depths of Earth currents oscillating in phase and out of phase with the primary current system (Pirjola and Viljanen, 1998).

As CIM works in the frequency domain, an inverse Fourier transform is needed to obtain the surface fields (and GIC) as functions of time. The present CIM formulation requires that the Earth's conductivity structure is layered. In particular, near ocean–continent boundaries the assumption need not be well satisfied. Therefore, possibilities of extensions to non-layered cases could be examined in the future. Investigations have been performed about the possible applicability of the 'exact image theory' (EIT) to geoelectromagnetics (Hänninen *et al.*, 2002). However, no breakthrough has been achieved – at least, not yet.

CIM and EIT are obviously not the only methods and not necessarily the best approaches applicable to the computation of electric and magnetic fields at the Earth's surface in a fast and accurate manner. Without giving precise details, Kappenman *et al.* (2000) refer to an efficient technique used for GIC studies in the British power system. Artificial intelligence and neural networks offer a different possible means to be considered in connection with space weather forecasting (Lundstedt, 1999; Boberg *et al.*, 2000).

10.3.2 Calculation of GIC

As mentioned above, the geoelectric field is generally rotational, and so it is not possible to talk about 'potentials' of different points at the Earth's surface. Thus, (geo)voltages producing GIC in a system have to be calculated by integrating the geoelectric field along the paths of the conductors. In other words, the voltage (or electromotive force) V_{ji}^0 affecting a line between points j and i in a network is obtained by integrating the geoelectric field E_0 along the conductor denoted by s_{ji} between j and i:

$$V_{ji}^0 = \int_{s_{ji}} E_0 \cdot ds \tag{10.4}$$

Concerning the 'engineering step' in the case of a power grid (a discretely-earthed system) with N nodes (= earthing points), Lehtinen and Pirjola (1985) have derived the following matrix formula for the $N \times 1$ column matrix I_e consisting of the earthing currents (with the positive direction into the Earth) at the nodes of the network:

$$I_e = (U + Y_n Z_e)^{-1} J_e \tag{10.5}$$

where U is an $N \times N$ unit (or identity) matrix, and Y_n and Z_e are the network admittance matrix and the earthing impedance matrix, respectively (Pirjola, 2005a). The matrix Z_e couples the earthing currents with the voltages between the earthing points and a remote Earth, while the matrix Y_n is associated with currents in the network conductors (power grid transmission lines). The derivation by Lehtinen and Pirjola (1985) refers to Ohm's and Kirchhoff's laws and Thévenin's theorem. Since problems caused by GIC in a power system arise from transformer saturation GIC flowing through, transformers, called 'earthing GIC', are more important in practice than GIC in transmission lines. (Compared to the usual power industry terminology, our 'earthing GIC' indicates GIC from, or to, a transformer neutral to, or from, the ground.)

Because geoelectromagnetic and GIC phenomena are slow, the matrices Y_n and Z_e are real in GIC computations (at least as the first approximation). If the earthing points are distant enough to render the influence of one earthing current on the voltage at another earthing point negligible, Z_e is simply diagonal, with the elements equalling the earthing resistances (R_i^e, $i = 1 \dots N$). The network admittance matrix is defined by

$$(i \neq j) : Y_{n,ij} = -\frac{1}{R_{ij}^n}, (i = j) : Y_{n,ij} = \sum_{k \neq i} \frac{1}{R_{ik}^n} \tag{10.6}$$

where R_{ij}^n is the conductor line resistance between nodes i and j. Equation (10.6) directly shows that Y_n is a symmetric $N \times N$ matrix. It can be shown that Z_e is a symmetric $N \times N$ matrix, too, even in the general case when Z_e is not diagonal. The elements of the $N \times 1$ column matrix J_e are given by

$$J_{e,i} = \sum_{j \neq i} J_{n,ji} \tag{10.7}$$

where

$$J_{n,ji} = \frac{V_{ji}^0}{R_{ji}^n} \tag{10.8}$$

It is seen from equation (10.5) that assuming perfect earthings ($Z_e = 0$), the earthing currents are given by the elements of the matrix J_e.

In a power system, GIC is equally divided between the three phase conductors. Therefore, all three phases are usually treated as one conductor with a resistance of one-third of that of a single phase. GIC values thus have to be divided by three to obtain the per-phase magnitudes. Such a single-line description is acceptable based on symmetry between the three phases (Mäkinen, 1993; Pirjola, 2005a). In GIC calculations, the 'earthing resistance' $R_i^e (i = 1 \dots N)$ can be conveniently understood to contain the actual earthing resistance of the station, the transformer resistance, and the resistance of a possible neutral point reactor in the earthing lead of the transformer, all in series. However, stations with several transformers in parallel, or autotransformers, at which two different voltage levels are in a metallic connection permitting the flow of GIC from one level to another, require a special careful modelling, as explained by Mäkinen (1993) and Pirjola (2005a) (see also Figure 10.2).

We now consider a theoretical example in which GIC are assumed to be created in the Finnish 220-kV and 400-kV power systems (Figure 10.3, see colour section) by an ionospheric 'westward travelling surge' (WTS) – an ionospheric current vortex that propagates to the west (Pirjola *et al.*, 2000; Pirjola, 2002b). The height and

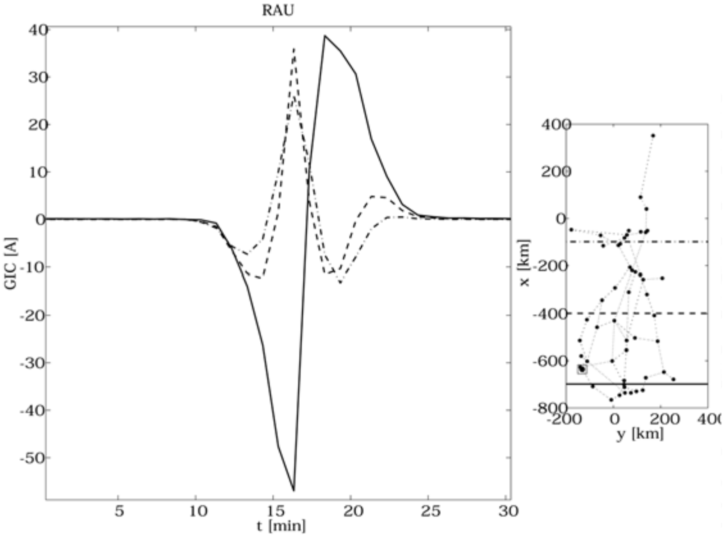


Figure 10.4. GIC as a function of time, flowing into the Earth through the Rauma 400-kV transformer in south-western Finland. GIC is calculated theoretically by assuming that it is produced by an ionospheric WTS event propagating over the Finnish high-voltage power system along three different paths indicated in the right-hand map, which also shows the location of Rauma. (For further details of the calculation see the text.) (Pirjola *et al.*, 2000; Pirjola, 2002b.)

(westward) velocity of the WTS are 110 km and 5 km/s, respectively. The WTS is accompanied by vertical currents (nearly field-aligned currents at high latitudes) that ensure the current continuity. The Earth is assumed to have a six-layer structure with layer thicknesses and resistivities 12, 22, 16, 50, 50, ∞ km, and 30,000, 3000, 50, 1000, 5000, 1 Ωm, approximately valid in Central Finland. Three different paths of the propagation of the WTS system indicated in the right-hand map of Figure 10.4 are investigated. The map also shows the location of the Rauma 400-kV station at which GIC are calculated and presented as functions of time for the three paths in Figure 10.4. The 'geophysical step' is carried out by applying CIM (equation (10.3)), and the 'engineering step' by using the matrix formalism (equations (10.4)–(10.8)). Figure 10.4 shows that the position of the WTS has a clear effect on GIC, and as expected, the largest GIC at Rauma is obtained for the closest path to this particular station.

GIC due to other ionospheric events – such as an electrojet, a Harang discontinuity, an omega band and a pulsation – can also be easily studied by applying CIM (Viljanen *et al.*, 1999).

10.4 GIC RESEARCH IN THE FINNISH HIGH-VOLTAGE POWER GRID

Together with the high-latitude location of Finland, the GIC problems that the large magnetic storm in August 1972 caused in North American power systems (Albertson and Thorson, 1974) were a motivation to start GIC studies in the Finnish high-voltage system in the 1970s. Another reason was the rapid growth of the power grid in Finland at that time. (When GIC research began, the network was much simpler than that shown in Figure 10.3 (colour section).) The works on GIC in the Finnish power grid have been and are carried out as a collaboration between the power company (Imatran Voima Oy, Fingrid Oyj) and the Finnish Meteorological Institute (FMI). The research has mostly concerned the highest voltage level (400 kV), but the 220 kV grid has also been considered. At lower voltages, GIC are probably not a potential source of problems.

The studies were started in 1977 by installing a GIC recording system in the earthing lead of the Huutokoski 400-kV transformer neutral (Elovaara *et al.*, 1992) (Figure 10.3, see colour section). At that time, Huutokoski was the end station of a roughly west–east 400-kV line, and it was thought (correctly) that such a station would experience large GIC and that modelling of GIC would be easy there. Although Finnish GIC recordings now cover more than two eleven-year sunspot cycles, drawing statistical conclusions from the data is not straightforward. This is because changes in the power system configuration may have large effects on the GIC distribution and values, and so observations of GIC at different times are not necessarily comparable. The recordings at Huutokoski were stopped in the beginning of the 1990s because changes that had occurred in the power network configuration implied that magnitudes of GIC flowing into (or from) the Earth at Huutokoski were drastically decreased. Consequently, the Huutokoski data were no longer interesting nor significant.

At the present time, GIC are monitored in the neutral earthing leads of the Pirttikoski, Rauma and Yllikkälä 400-kV transformers (Figure 10.3, colour section). The largest GIC measured in the Finnish 400-kV power system is as high as (−) 201 A (as a 1-minute mean value) at Rauma on 24 March 1991 (67 A per phase) (uppermost panel of Figure 10.5). GIC values are sensitive to changes of resistances and the configuration of the network, which explains why the largest current recorded at Rauma during the great magnetic storm at the end of October 2003 was only 41.5 A (as a 10-second mean value) (upper panel of Figure 10.6). Of course, the storms in March 1991 and October 2003 have several physical differences affecting the geoelectric field magnitudes and GIC but, obviously, changes in the power system that occurred between 1991 and 2003 play an important role in the decrease of GIC at Rauma.

The largest documented and confirmed GIC value anywhere, and ever measured, is probably 320 A in a transformer neutral lead (107 A per phase) in the Swedish power system during the geomagnetic storm in April 2000 (Erinmez *et al.*, 2002b). The comment mentioned by Stauning (2002) that the particular GIC value in Sweden had been even 600 A is an overestimation probably based on a misunderstanding (private communication with a Swedish power engineer). We may believe that larger GIC than 320 A have occurred somewhere, at some time, but they have not been recorded or reported.

Besides GIC at Rauma, Figure 10.5 also shows the recordings of the north component of the geomagnetic field and its time derivative at the Nurmijärvi

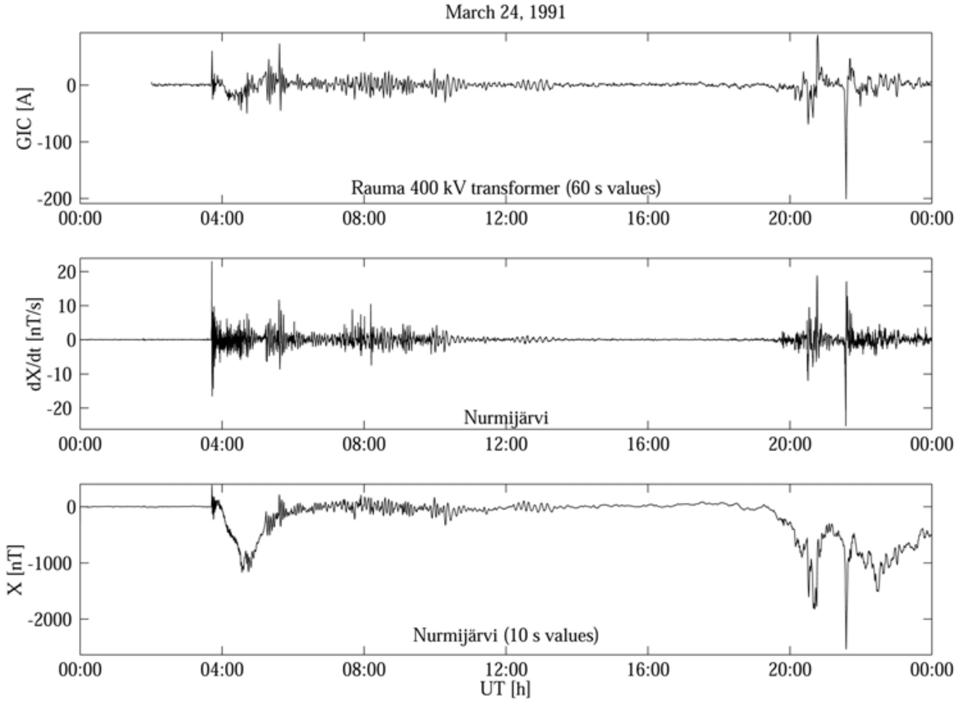


Figure 10.5. GIC (uppermost panel) recorded in the earthing lead of the Rauma 400-kV transformer neutral in south-western Finland on 24 March 1991. The north component (X) (lowermost panel) of the geomagnetic field and its time derivative (middle panel) at the Nurmijärvi Geophysical Observatory in southern Finland are also shown. The value of 201 A seen in the uppermost panel is the largest GIC measured in the Finnish 400-kV system. (Pirjola *et al.*, 2003; Pirjola *et al.*, 2005.)

Geophysical Observatory in southern Finland. In Figure 10.6 we compare GIC with the magnetic time derivative. A clear correlation between GIC and the time derivative is seen in both figures (being in agreement with the inductive nature of GIC). Nevertheless, it should be noted that the horizontal geoelectric field is not proportional to the magnetic time derivative as shown, for example, by equation (10.2). Depending on the conductivity structure of the Earth, the time dependence of the geoelectric field may in some cases resemble that of the magnetic field rather than of its time derivative. Because geoelectromagnetic variations are slow, making the network (approximately) purely resistive, the temporal behaviour of GIC follows that of the geoelectric field.

Possible power system problems arise from GIC flowing through transformers, so most recordings of GIC concern earthing leads of transformer neutrals. Such measurements are also easier to perform in practice than monitoring GIC in high-voltage transmission lines (Elovaara *et al.*, 1992). In Finland, however, GIC have also been recorded in a 400-kV line, which happened in 1991 to 1992 during a special

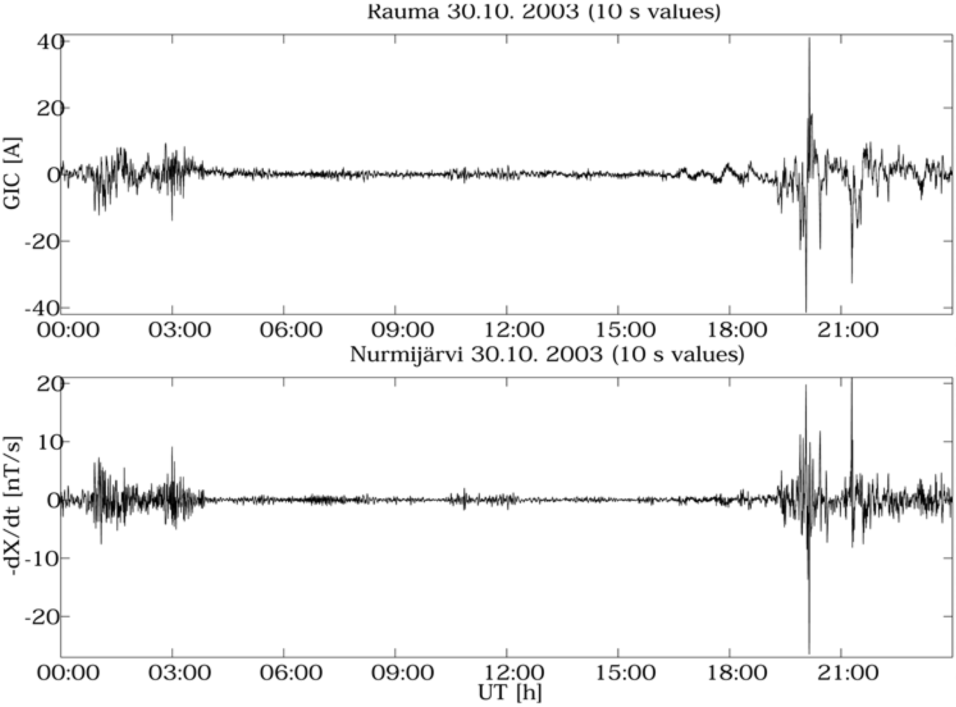


Figure 10.6. GIC (upper panel) recorded in the earthing lead of the Rauma 400-kV transformer neutral in southwestern Finland on 30 October 2003. The (negative) time derivative of the north component (X) of the geomagnetic field (lower panel) at the Nurmijärvi Geophysical Observatory in southern Finland is also shown. (Pirjola *et al.*, 2005).

GIC project. The measurement was accomplished by observing the magnetic field produced by GIC with two magnetometers – one lying below the line and the other located farther away, thus representing the reference field (Mäkinen, 1993; Viljanen and Pirjola, 1994). (A similar method is being used in measurements of GIC in the Finnish natural gas pipeline started in 1998 (Pulkkinen *et al.*, 2001a).) On average, GIC in power transmission lines are larger than those through transformers (Pirjola, 2005b).

In order to obtain a general idea of the distribution of GIC in a grid and to locate the sites that most probably will experience large GIC, it is appropriate to assume that the horizontal geoelectric field is spatially constant (Pirjola and Lehtinen, 1985). More realistic spatially-varying fields are also easily manageable, which was demonstrated, for example, by the WTS study (end of Section 10.3.2).

For the evaluation of the GIC risk in the Finnish high-voltage power system, FMI has performed several statistical studies (Viljanen and Pirjola, 1989; Mäkinen, 1993; Pulkkinen *et al.*, 2000). A particular effort was the 'GIC project' (mentioned above) in 1991–1992, when GIC were recorded simultaneously at four 400-kV transformers and in a transmission line (Elovaara *et al.*, 1992). To know precisely the possible GIC impacts on Finnish high-voltage transformers, Fingrid Oyj has carried out field tests (in 1979 and in 1999) in which dc currents were injected into transformers and extensive and detailed measurements and investigations about the consequences were performed (Pesonen, 1982; Lahtinen and Elovaara, 2002). The tests indicate that GIC that would last long enough to cause problems in Finland are extremely rare. In spite of large GIC magnitudes in transformers during some magnetic storms, the Finnish power system has practically not experienced GIC problems and is thus generally considered secure. This is obviously due to the transformer type and structure used in the country. Nevertheless, research is continuing to concentrate especially on extreme and exceptional space weather events, whose ionospheric characteristics and GIC impacts should still be investigated. Such research is also useful for geophysical and space physical science.

10.5 CONCLUSION

Space weather is a very important research topic today. It refers to particle and electromagnetic conditions in the Earth's near space, and is a source of problems for technological systems in space and on the Earth's surface. Furthermore, space weather constitutes a radiation risk to people onboard spacecraft as well as in high-altitude aircraft. Although ground effects of space weather (GIC) were already observed in early telegraph systems about 150 years ago, the term 'space weather' is new. The importance of space weather phenomena is continuously increasing as society becomes more and more dependent on reliable technology. Space weather risk also begins to be of interest to the insurance business (Jansen *et al.*, 2000), and space weather also plays an economical role in electricity markets (Forbes and St. Cyr, 2002, 2004).

GIC can flow in electric power transmission systems, oil and gas pipelines, telecommunication cables and railway equipment. Power grids are obviously today's systems most vulnerable to GIC (Kappenman, 2004b), and a great deal of research has been carried out about GIC impacts on power systems, in which problems result from saturation of transformers produced by the dc-like GIC. The troubles extend from minor relaying problems to a collapse of a large network and to permanent damage of transformers, implying large economic losses and other harm. The most famous and worst GIC event is the several-hour blackout in Québec, Canada, in March 1989. Another recent example is the shorter blackout in southern Sweden in October 2003. Although the reason was other than GIC in most cases, the blackouts that occurred in Scandinavia, Italy and the USA in 2003 clearly demonstrated the vulnerability of electric transmission systems and the chaos created in a modern society (Kappenman, 2004a). A large power system may be collapsed due to a set of cascading domino-effect phenomena originating in a failure in a limited area.

The greatest geomagnetic storms occur in high-latitude auroral regions, so the GIC problem is evidently largest in the same areas. However, GIC magnitudes and, in particular, their effects on the system do not only depend on the intensity of the magnetic disturbance but are also affected by the configuration and structure of the network, which means that lower latitudes may also experience GIC problems. Finland is a high-latitude country, in which GIC studies have been actively carried out for almost thirty years. Serious GIC problems have, however, not arisen in Finland.

In principle, there are two ways to avoid GIC problems. The first is to try to block the flow of GIC and to design the system and its equipment to be GIC-insensitive. In a power system, the blocking can in principle be accomplished by installing capacitors which block GIC but do not disturb the flow of 50/60-Hz currents. Such a solution is, however, not straightforward and simple. Therefore, the second alternative – forecasting GIC events – is obviously more feasible. After receiving a GIC alert, system operators can take actions to minimise problems in their systems. However, it is of utmost importance that GIC predictions are reliable, since false alarms resulting in unnecessary countermeasures are both expensive and harmful.

The horizontal geoelectric field induced at the Earth's surface during a space weather storm is the key parameter, as it is the driving force of GIC. Consequently, efforts should be concentrated on developing prediction methods of the electric field. Satellites continuously monitor the solar wind, and the Sun is also observed. These data provide a good basis for predictions of space weather and GIC. However, without a sufficient understanding of the plasmaphysical coupling between the solar wind, the magnetosphere and the ionosphere, the forecasting of the geoelectric field may remain just qualitative. Quantitative and accurate forecasts still require new achievements in modelling research. The computations used for forecasting purposes must be fast. Thus, instead of, or in parallel with, physical modelling of space weather and GIC phenomena, artificial intelligence and neural network techniques should also be investigated. The best solution today might be a suitable combination of these two approaches.

10.6 ACKNOWLEDGEMENTS

The author wishes to thank his colleagues Dr Ari Viljanen and Dr Antti Pulkkinen (Finnish Meteorological Institute), and Dr David Boteler and Dr Larisa Trichtchenko (Natural Resources Canada) for fruitful collaboration on GIC research during many years. The great interest in, and support for GIC studies by the Fingrid Oyj and Gasum Oy companies, Finland, is also gratefully acknowledged.

10.7 REFERENCES

Albertson, V. D., and J. M. Thorson, Jr. (1974), Power system disturbance during a K-8 geomagnetic storm, August 4, 1972, *IEEE Transactions on Power Apparatus and Systems*, PAS-93, 1025–1030.

Arajärvi, E., R. Pirjola, and A. Viljanen (2006), Effects of neutral point reactors and series capacitors on geomagnetically induced currents in a high-voltage electric power transmission system; under preparation.

Boberg, F., P. Wintoft, and H. Lundstedt (2000), Real Time Kp Predictions from Solar Wind Data using Neural Networks, *Physics and Chemistry of the Earth*, C, **25**, 4, 275–280.

Bolduc, L. (2002), GIC observations and studies in the Hydro-Québec power system, *Journal of Atmospheric and Solar-Terrestrial Physics*, **64**, 16, 1793–1802.

Bolduc, L. and P. Langlois (1995), Hydro-Québec follow-up of the March 1989 event, Paper presented at Session GA 5.06 'Geomagnetic Applications' at the *XXI General Assembly of the International Union of Geodesy and Geophysics*, Boulder, Colorado, USA, July 2–14 1995, 20 pp.

Bolduc, L., P. Langlois, D. Boteler, and R. Pirjola (1998), A study of geoelectromagnetic disturbances in Québec, 1. General results, *IEEE Transactions on Power Delivery*, **13**, 4, 1251–1256.

Bolduc, L., P. Langlois, D. Boteler, and R. Pirjola (2000), A Study of Geoelectromagnetic Disturbances in Québec, 2. Detailed Analysis of a Large Event, *IEEE Transactions on Power Delivery*, **15**, 1, 272–278.

Boteler, D. H. (2000), Geomagnetic effects on the pipe-to-soil potentials of a continental pipeline, *Advances in Space Research*, **26**, 1, 15–20.

Boteler, D. H. and R. J. Pirjola (1998a), The complex-image method for calculating the magnetic and electric fields produced at the surface of the Earth by the auroral electrojet, *Geophysical Journal International*, **132**, 1, 31–40.

Boteler, D. H. and R. J. Pirjola (1998b), Modelling Geomagnetically Induced Currents Produced by Realistic and Uniform Electric Fields, *IEEE Transactions on Power Delivery*, **13**, 4, 1303–1308.

Boteler, D. H., R. J. Pirjola, and H. Nevanlinna (1998), The effects of geomagnetic disturbances on electrical systems at the earth's surface, *Advances in Space Research*, **22**, 1, 17–27.

Campbell, W. H. (1978), Induction of Auroral Zone Electric Currents Within the Alaska Pipeline, *Pure and Applied Geophysics (PAGEOPH)*, **116**, 1143–1173.

Czech, P., S. Chano, H. Huynh, and A. Dutil (1992), The Hydro-Québec system blackout of 13 March 1989: system response to geomagnetic disturbance, EPRI Report, TR-100450, in *Proceedings of Geomagnetically Induced Currents Conference*, Millbrae, California, USA, November 8–10 1989, 19.1–19.21.

Davidson, W. F. (1940), The magnetic storm of March 24, 1940 – Effects in the power system, *Edison Electric Institute Bulletin*, July 1940, 365–366 and 374.

Elovaara, J., P. Lindblad, A. Viljanen, T. Mäkinen, R. Pirjola, S. Larsson, and B. Kielén (1992), Geomagnetically induced currents in the Nordic power system and their effects on equipment, control, protection and operation, *CIGRE General Session 1992* (CIGRE – International Conference on Large High Voltage Electric Systems), Paris, August 31–September 5, 1992, No. 36-301, 10 pp.

Erinmez, I. A., J. G. Kappenman, and W. A. Radasky (2002a), Management of the geomagnetically induced current risks on the national grid company's electric power transmission system, *Journal of Atmospheric and Solar-Terrestrial Physics*, **64**, 5–6, 743–756.

Erinmez, I. A., S. Majithia, C. Rogers, T. Yasuhiro, S. Ogawa, H. Swahn, and J. G. Kappenman (2002b), Application of modelling techniques to assess geomagnetically induced current risks on the NGC transmission system, *CIGRE, Session – 2002*, No. 39-304, 10 pp.

Forbes, K. F. and O. C. St. Cyr (2002), Space Weather and the Electricity Market: A Preliminary Assessment, in *CD ROM Proceedings of the XXVIIth General Assembly of the International Union of Radio Science (URSI)*, Maastricht, The Netherlands, August 17–24, Paper Number 0233, 4 pp.

Forbes, K. F. and O. C. St. Cyr (2004), Space weather and the electricity market: An initial assessment, *AGU Space Weather Journal*, **2**, S10003, doi:10.1029/2003SW000005, 28 pp.

Häkkinen, L. and R. Pirjola (1986), Calculation of electric and magnetic fields due to an electrojet current system above a layered earth, *Geophysica*, **22**, 1–2, 31–44.

Hänninen, J. J., R. J. Pirjola, and I. V. Lindell (2002), Application of the exact image theory to studies of ground effects of space weather, *Geophysical Journal International*, **151**, 2, 534–542.

Jansen, F. and R. Pirjola (2004), Space Weather Research Elucidates Risks to Technological Infrastructure. *EOS, Transactions, American Geophysical Union*, **85**, 25, 241 and 245–246.

Jansen, F., R. Pirjola, and R. Favre (2000), *Space Weather, Hazard to the Earth?*. Swiss Re Publishing, Swiss Reinsurance Company, Zurich, Switzerland, R&R, 6/00, 4000 en, 40 pp.

Kappenman, J. G. (1996), Geomagnetic Storms and Their Impact on Power Systems, *IEEE Power Engineering Review*, **16**, 5–8.

Kappenman, J. G. (2004a), Systemic Failure on a Grand Scale: The 14 August 2003 North American Blackout, *Space Weather*, **1**, 1011, doi: 10.1029/2003SW000027, 1, 2, 4.

Kappenman, J. G. (2004b), The Evolving Vulnerability of Electric Power Grids, *Space Weather*, **2**, S01004, doi:10.1029/2003SW000028, 1, 2, 10–12.

Kappenman, J. G. and V. D. Albertson (1990), Bracing for the geomagnetic storms, *IEEE Spectrum*, March 1990, 27–33.

Kappenman, J. G., W. A. Radasky, J. L. Gilbert, and I. A. Erinmez (2000), Advanced Geomagnetic Storm Forecasting: A Risk Management Tool for Electric Power System Operations, *IEEE Transactions on Plasma Science*, **28**, 6, 2114–2121.

Lahtinen, M. and J. Elovaara (2002), GIC Occurrences and GIC Test for 400 kV System Transformer, *IEEE Transactions on Power Delivery*, **17**, 2, 555–561.

Lanzerotti, L. J., L. V. Medford, C. G. Maclennan and D. J. Thomson (1995), Studies of Large-Scale Earth Potentials Across Oceanic Distances, *AT&T Technical Journal*, May/June 1995, 73–84.

Lehtinen, M. and R. Pirjola (1985), Currents produced in earthed conductor networks by geomagnetically-induced electric fields, *Annales Geophysicae*, **3**, 4, 479–484.

Lundstedt, H. (1999), The Swedish space weather initiatives, in *Proceedings of the ESA Workshop on Space Weather*, ESTEC, Noordwijk, The Netherlands, November 11–13 1998, European Space Agency, WPP-155, 197–205.

Mäkinen, T. (1993), Geomagnetically induced currents in the Finnish power transmission system, *Geophysical Publications*, No. 32, Finnish Meteorological Institute, Helsinki, Finland, 101 pp.

Molinski, T. S. (2002), Why utilities respect geomagnetically induced currents, *Journal of Atmospheric and Solar-Terrestrial Physics*, **64**, 16, 1765–1778.

Pesonen, A. J. (1982), Discussion on the paper 'Characteristics of geomagnetically induced currents in the B. C. Hydro 500 kV system' by D. H. Boteler, T. Watanabe, R. M. Shier and R. E. Horita, *IEEE Transactions on Power Apparatus and Systems*, PAS-101, **6**, 1453–1454.

Petschek, H. E. and W. E. Feero (1997), Workshop Focuses on Space Weather's Impact on Electric Power, *EOS, Transactions, American Geophysical Union*, **78**, 21, 217–218.

Pirjola, R. (1982), Electromagnetic induction in the earth by a plane wave or by fields of line currents harmonic in time and space, *Geophysica*, **18**, 1–2, 1–161.

Pirjola, R. (1992), On magnetotelluric source effects caused by an auroral electrojet system, *Radio Science*, **27**, 463–468.

Pirjola, R. (2000), Geomagnetically Induced Currents During Magnetic Storms, *IEEE Transactions on Plasma Science*, **28**, 6, 1867–1873.

Pirjola, R. (2002a), Fundamentals about the flow of geomagnetically induced currents in a power system applicable to estimating space weather risks and designing remedies, *Journal of Atmospheric and Solar–Terrestrial Physics*, **64**, 18, 1967–1972.

Pirjola, R. (2002b), Geomagnetic Effects on Ground-Based Technological Systems, in W. Ross Stone (ed.) *Review of Radio Science 1999–2002*, International Union of Radio Science (URSI), Commission E, edited by P. Degauque, IEEE Press, John Wiley & Sons, Inc., Publication, Chapter 21, 473–496.

Pirjola, R. (2005a), Effects of space weather on high-latitude ground systems, *Advances in Space Research*, **36**, doi:10.1016/j.asr.2003.04.074, 2231–2240.

Pirjola, R. (2005b), Averages of geomagnetically induced currents (GIC) in the Finnish 400 kV electric power transmission system and the effect of neutral point reactors on GIC, *Journal of Atmospheric and Solar-Terrestrial Physics*, **67**, 7, 701–708.

Pirjola, R. J. and L. V. T. Häkkinen (1991), Electromagnetic Field Caused by an Auroral Electrojet Current System Model, in H. Kikuchi (ed.) *Environmental and Space Electromagnetics, Proceedings of the URSI International Symposium on Environmental and Space Electromagnetics*, Tokyo, Japan, September 4–6 1989, Springer-Verlag, Chapter 6.5, 288–298.

Pirjola, R. and M. Lehtinen (1985), Currents produced in the Finnish 400 kV power transmission grid and in the Finnish natural gas pipeline by geomagnetically-induced electric fields, *Annales Geophysicae*, **3**, 4, 485–491.

Pirjola, R. and A. Viljanen (1998), Complex image method for calculating electric and magnetic fields produced by an auroral electrojet of finite length, *Annales Geophysicae*, **16**, 11, 1434–1444.

Pirjola, R., A. Pulkkinen, and A. Viljanen (2003), Studies of space weather effects on the Finnish natural gas pipeline and on the Finnish high-voltage power system, *Advances in Space Research*, **31**, 4, 795–805.

Pirjola, R., A. Viljanen, A. Pulkkinen, and O. Amm (2000), Space Weather Risk in Power Systems and Pipelines, *Physics and Chemistry of the Earth*, C, **25**, 4, 333–337.

Pirjola, R., K. Kauristie, H. Lappalainen, A. Viljanen, and A. Pulkkinen (2005), Space weather risk, *AGU Space Weather Journal*, **3**, 2, S02A02, doi:10.1029/2004SW000112, 11 pp.

Prescott, G. B. (1866), *History, Theory, and Practice of the Electric Telegraph*, Ticknor and Fields, Boston, 508 pp.

Pulkkinen, A. (2003), Geomagnetic induction during highly disturbed space weather conditions: studies of ground effects, *Contributions*, No. 42, Finnish Meteorological Institute, Helsinki, Finland, 129 pp.

Pulkkinen, A., S. Lindahl, A. Viljanen, and R. Pirjola (2005), Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system, *AGU Space Weather Journal*, **3**, S08C03, doi:10.1029/2004SW000123, 19 pp.

Pulkkinen, A., A. Viljanen, K. Pajunpää, and R. Pirjola (2001a), Recordings and occurrence of geomagnetically induced currents in the Finnish natural gas pipeline network, *Journal of Applied Geophysics*, **48**, 219–231.

Pulkkinen, A., A. Viljanen, R. Pirjola, and BEAR Working Group (2000), Large geomagnetically induced currents in the Finnish high-voltage power system, *Reports*, No. 2000:2, Finnish Meteorological Institute, Helsinki, Finland, 99 pp.

Pulkkinen, A., R. Pirjola, D. Boteler, A. Viljanen, and I. Yegorov (2001b), Modelling of space weather effects on pipelines, *Journal of Applied Geophysics*, **48**, 233–256.

Stauning, P. (2002), High-voltage power grid disturbances during geomagnetic storms, in *Proceedings of the Second Solar Cycle and Space Weather Conference (SOLSPA 2001)*, Vico Equense, Italy, 24–29 September 2001, European Space Agency, SP-477, 521–524.

Thomson, D. J. and J. T. Weaver (1975), The Complex Image Approximation for Induction in a Multilayered Earth, *Journal of Geophysical Research*, **80**, 1, 123–129.

Viljanen, A. and R. Pirjola (1989), Statistics on geomagnetically-induced currents in the Finnish 400 kV power system based on recordings of geomagnetic variations, *Journal of Geomagnetism and Geoelectricity*, **41**, 411–420.

Viljanen, A. and R. Pirjola (1994), Geomagnetically induced currents in the Finnish high-voltage power system, a geophysical review, *Surveys in Geophysics*, **15**, 4, 383–408.

Viljanen, A., O. Amm, and R. Pirjola (1999), Modeling geomagnetically induced currents during different ionospheric situations, *Journal of Geophysical Research*, **104**, A12, 28059–28071.

Viljanen, A., A. Pulkkinen, O. Amm, R. Pirjola, T. Korja, and BEAR Working Group (2004), Fast computation of the geoelectric field using the method of elementary current systems and planar Earth models, *Annales Geophysicae*, **22**, 1, 101–113.

Wait, J. R. and K. P. Spies (1969), On the Representation of the Quasi-Static Fields of a Line Current Source above the Ground, *Canadian Journal of Physics*, **47**, 2731–2733.

Wallerius, A. (1982), Solen gav Sverige en strömstöt (The Sun gave Sweden a current impact), (in Swedish), *Ny Teknik – teknisk tidskrift*, **29**, p. 3.

11

Space weather impacts on space radiation protection

Rainer Facius and Günther Reitz

11.1 INTRODUCTION

Apart from the optical radiation of our home star, the Sun, whose energy input sustains the terrestrial biosphere, Earth, together with other planets, is embedded in a most complex mixture of ionizing radiation – radiation with energy sufficient to ionize the atoms or molecules of the matter which it penetrates. This chapter focuses on the interplay of space weather parameters and the field of ionizing space radiation as far as it is a source of occupational radiation exposure of humans – space or air crew – and examines the related problems of radiation protection [R109]. The still, in large part, unresolved questions about unique radiobiological aspects of irradiation by heavy ions – galactic as well as from terrestrial accelerators – as they were outlined time and again in [R68, R69, R50, R51, R52], have recently been summarized once more in [R13]. Though most pertinent for the assessment of radiation risks – especially during long-term space missions outside the shielding provided by the terrestrial magnetosphere – a discussion of these significant radiobiological particularities of galactic heavy ions would exceed the scope of a discourse on space weather effects and radiation protection. We also omit a discussion of indirect 'health' risks which might ensue from malfunctioning – mainly electronic – equipment that might become overexposed during space weather events (e.g., Chapter 12). Of course, a comprehensive appraisal of space radiation risks must also include these space radiation related risks.

We review the physical features of the radiation fields as far as they are relevant for radiation protection, and present the quantitative measures needed to assess exposure to these fields in such terms as required by radiation protection. In addition to a proper discussion of the physical features of space radiation, the extraordinary complexity of this radiation field as well as the intricacies of radiobiological effects in man enforce a somewhat detailed discussion of the underlying dosimetric and radiobiological principles. After the description of radiobiological health effects in man we discuss the exposure limits which for the terrestrial environment have been developed to either prevent adverse health effects altogether, or to at least keep them below levels which can be accepted as tolerable. These limits will be juxtaposed to radiation exposures which have been or will be incurred during previous or future planned missions to the Moon or Mars. Returning to Earth, we will describe and discuss radiation exposures and health effects in civil air flight which takes place in the fringes of space radiation fields, as far as they can penetrate into our atmosphere down to cruising altitudes. In contrast to the small numbers of exposed astronauts, here we have already a sufficient epidemiological database to measure, with some reliability, health effects directly instead of theoretical predictions which we have to rely upon for space missions. Finally, a cursory survey of the effects of space weather and space radiation on the biosphere in general will complete this chapter.

11.2 RADIATION FIELDS

This section presents the salient features of the field of ionizing space radiation as far as it is known to be relevant for the biosphere and in particular for radiation protection of humans. Primary electromagnetic ionizing radiation, for example, from solar Röntgen flares such as of 4 November 2003, 19:29 UTC, or from conspicuous extreme gamma-ray bursts such as the most recent, of 27 December 2004, 21:30:26.55 UTC, at least presently do not contribute measurably to this exposure and hence are omitted, although on a geological time scale their impact on the biosphere might have been significant. Secondary electromagnetic radiation, of course, contributes as Bremsstrahlung emitted from charged particles upon penetration through matter and as gamma-rays from the decay of π^0 pions created in atmospheric showers. So, we will focus on the particulate component of space radiation and here again on ions only. Electrons might become relevant if manned activities in the outer radiation belts were an issue – which they are not, for the foreseeable future.

11.2.1 Primary fields

Like all celestial bodies equipped with a magnetic moment, Earth is surrounded by toroidal belts of particulate radiation which becomes trapped by the magnetic field [w1]. Energetic charged particles emitted by the Sun constitute one of the primary sources which constantly replenish these radiation belts, but which of course constitute an important primary source of ionizing radiation itself. Galactic cosmic radiation constitutes the third source of ionizing radiation, which in contrast to the other two is not restricted to the heliosphere. These primary sources interact by various mechanisms and jointly they determine, at any given time and location within the heliosphere, the actual field of ionizing radiation, the complexity of which is unrivalled by anything we know from terrestrial experience.

Figure 11.1 (see colour section) illustrates these three major sources of ionizing space radiation, their respective spatial scales, and the dominant role our Sun plays in modifying its composition. The highest energies measured for galactic cosmic ray (GCR) particles (Figure 11.14) are too large to be compatible with their postulated acceleration and containment by intragalactic magnetic fields, thereby giving rise to speculation about extragalactic sources for this part, and extending the spatial scales even further. Yet the corresponding intensities are too low to contribute measurably to radiation exposures.

In addition to their variation with location in space, the intensity and particulate composition in these fields are subject to temporal variations. As far as space radiation is concerned, two temporal scales of space weather events are relevant. Similar to the 'smooth' annual alternation between summer and winter of ordinary weather we have to deal with a rather regular change of solar activity between phases of maximal ('summer') and minimal ('winter') solar activity [R22]. The solar 'year' in this case is the Schwabe cycle – a period of about 11 terrestrial years, the duration of which (presently) varies due to so far unknown mechanisms between 9 and 13.6 years [w4, R21]. One measure of this activity, for which a continuous observational record exists since 1755, is the Zürich sunspot number [R3]. Apparently, the maximum of solar activity is inversely associated with the length of the cycle, which in turn appears to be also associated with climate variability [R4, R5]. In addition to the regular solar cycle, during episodes of extreme solar activity, as characterized by explosive releases of magnetic energy [R15], giant masses of charged particles are ejected from the corona into the interplanetary magnetic field (coronal mass ejection, CME). After further acceleration in this field, particle energies up to several GeV can be attained. The impact on the space radiation field of these solar particle events (SPE) can last for days to some weeks.

Further observed solar periodicities – such as the 22-year magnetic Hale cycle, the $\sim$88-year Gleisberg cycle or the $\sim$210-year De Vries or Suess cycle – have not yet been identified to modulate radiation exposures, although their impact on the biosphere probably outclasses by far that of radiobiological mechanisms, as a recent study on glacial climate cycles discloses [R14].

11.2.1.1 Radiation belts

Apart from very low-altitude (less than 200-km) and low-inclination orbits (less than, say, 25°), radiation exposure of space crew in low Earth orbits (LEO) is dominated by energetic charged particles trapped by the geomagnetic field.

In 1956, James Alfred Van Allen, from the University of Iowa, proposed to the IGY (International Geophysical Year) directorate that a simple but globally comprehensive cosmic-ray investigation should be implemented in one of the early US satellites. By virtue of preparedness and good fortune, his Iowa cosmic-ray instrument was selected as the principal element of the payload of the first successful orbiting satellite of the United States. Measurements taken by a single Geiger–Müller tube after the launch of Explorer I on 31 January 1958, at 10:48 pm-EST, provided the first empirical evidence of trapped radiation as it had been theoretically anticipated due to the well known interaction of charged particles with magnetic fields (Figure 11.2).

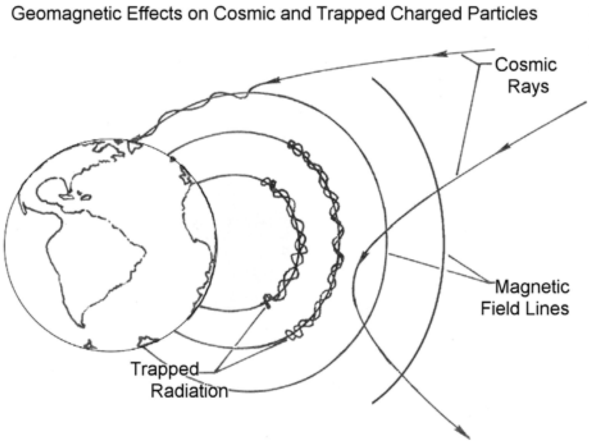


Figure 11.2. Deflection and trapping of charged particles by the geomagnetic field.

During the subsequent decade, extensive measurements with more advanced and dedicated instrumentation on several satellites in well-coordinated orbits yielded the main quantitative database which finally became integrated in the *AP8 Trapped Proton Model* [w2], and which provides energy spectra of average proton fluxes during quiet magnetospheric conditions. A major application for which these models have been designed is the assessment of the radiation exposure from trapped radiation during manned low Earth orbit (LEO) missions such as presently on the International Space Station. The *AE8 Trapped Electron Model* [w3] serves the same purpose of predicting radiation doses from trapped electrons. Their direct contribution to radiation exposure is restricted to lightly shielded spacecraft. In heavier shielded vehicles they can contribute indirectly by means of Bremsstrahlung.

Figure 11.3 reveals the major qualitative features detected by these quantitative observations of the radiation belts. Inner and outer belts exist for both electrons and protons where the outer belts are populated by solar particles with lower energies – or rather, magnetic stiffness (rigidity). In particular, the inner proton belt is replenished by protons which originate from the decay of 'albedo-' neutrons, which in turn are the spallation products from interactions of GCR particles with atoms of the atmosphere. Energy loss by cyclotron radiation and by penetration into the upper atmosphere near the geomagnetic mirror points constitutes the major loss mechanisms for the trapped particle population.

Figure 11.4 displays the spatial distribution of electron flux for electron energies above 0.5 MeV (right) and of proton flux for proton energies above 34 MeV (left), at which energy the latter are able to penetrate about $2\,\mathrm{g\,cm^{-2}}$ of aluminium – the shielding provided by lighter spacecraft. Proton fluxes in the inner belt are sufficiently intense and reach sufficient energies to penetrate the shielding provided by walls and equipment of spacecraft, so that primarily their energy spectra, as shown in Figure 11.5, have to be known in order to assess radiation exposures of space crew.

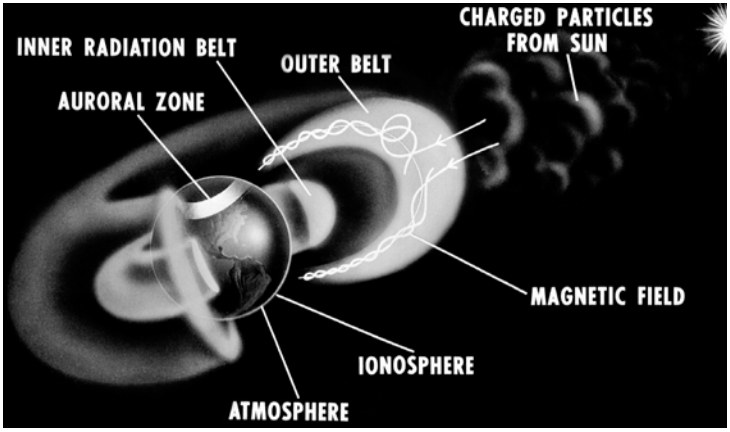


Figure 11.3. Inner and outer belts of magnetically trapped charged particles.

The data in the figure are the results of measurements of trapped proton spectra in the early 1960s. As a ‘natural’ coordinate system to specify the satellite position within the geomagnetic field, the (B, L) coordinates are used [R10, R6]. Here, B denotes the magnetic field strength at a given point, and L the altitude in Earth radii at which the magnetic field line through this point intersects the plane through the geomagnetic equator.

Figures 11.6 and 11.7 show the fluxes of trapped electrons and protons averaged over the orbit of the Hubble Space Telescope. Electron fluxes during solar maximum

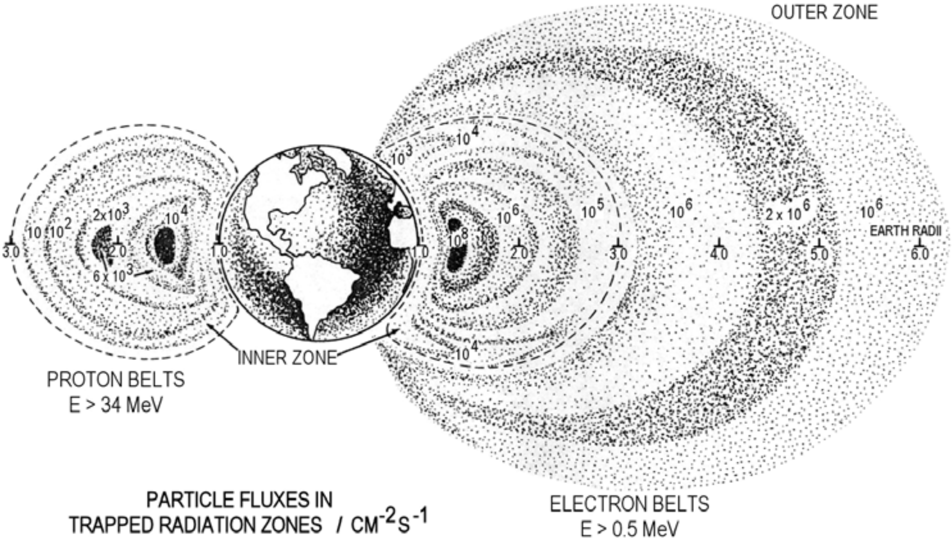


Figure 11.4. Integral flux densities in inner and outer terrestrial radiation belts for trapped protons and electrons in particles $cm^{-2}\,s^{-1}$.

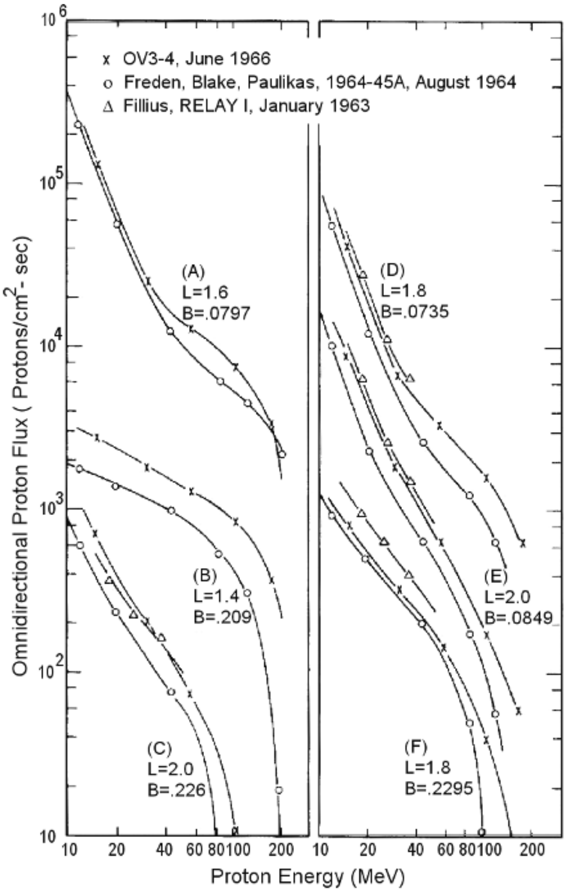


Figure 11.5. Energy spectra of inner belt protons measured at various locations within the belts as expressed in the B, L–coordinate system.

are greater than during solar minimum, pointing to the Sun as the dominant primary source which feeds the trapped electron population. In contrast, the trapped proton fluxes reflect the (Forbush) modulation of the GCR intensity by the solar wind, which results in higher intensities during solar minimum conditions. (Further details on the terrestrial trapped radiation belts and on their linkage to space weather are given in [R6, R7].)

The fluxes and spectra shown pertain to quiet magnetic conditions of the terrestrial and interplanetary magnetic field during the minimum and maximum of solar activity. In addition to the regular solar cycle variation, both magnetic storms and intensive fluxes from energetic solar particle events (SPE) significantly shift positions and energies of trapped particle populations, so that even additional though transient radiation belts can be created. Especially in the outer electron belt, flux fluctuations by orders of magnitude can occur. Until the dedicated

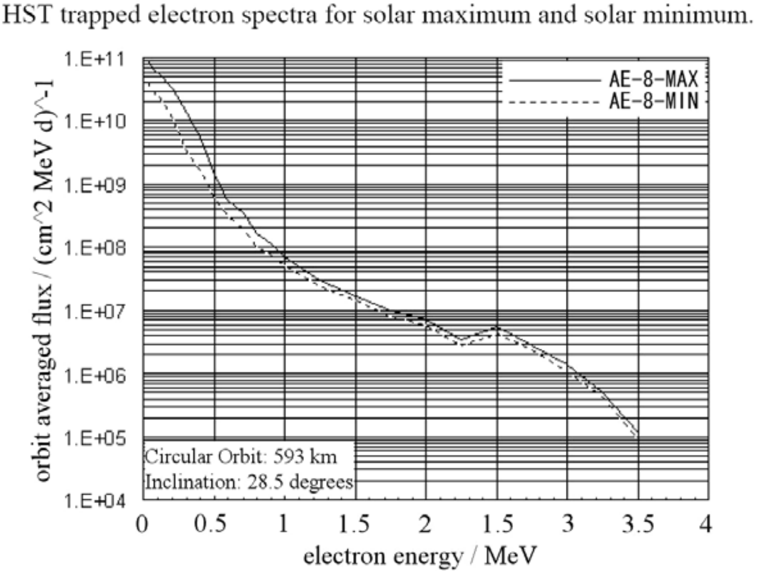


Figure 11.6. AE-8 model differential trapped electron fluxes at solar minimum and maximum conditions for the Hubble Space Telescope.

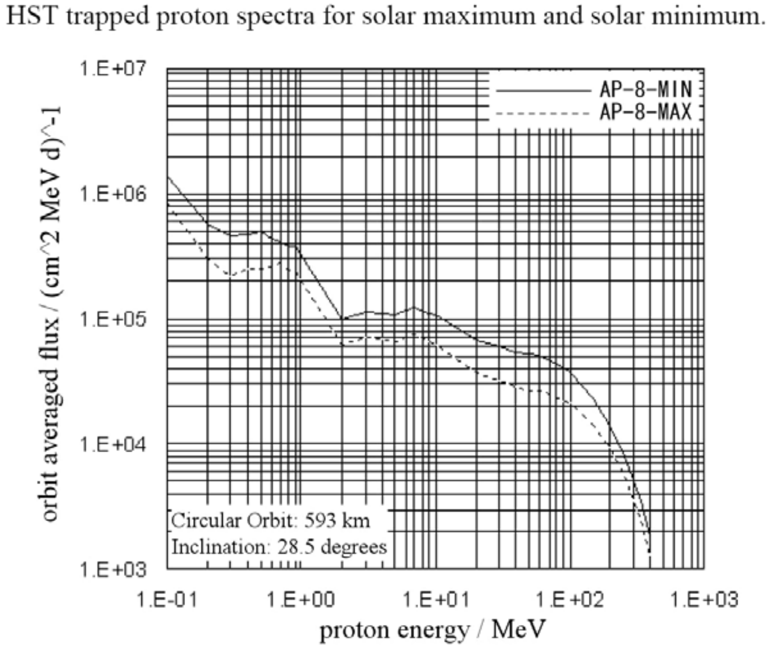


Figure 11.7. AP-8 model differential trapped proton fluxes at solar minimum and maximum conditions for the Hubble Space Telescope.

14-month mission of the Combined Release and Radiation Effects Satellite (CRRES) [R91] between 25 July 1990 to 12 October 1991, only isolated fragmentary data existed regarding the short-term dynamic behaviour of the trapped populations in response to space weather events. A very large SPE starting on 23 March 1991 and lasting until 25 March [W6] created two new radiation belts. At 03:43 UTC on 24 March, another high-energy proton belt, with energies above 80 MeV with a centre at $L = 2.25$ Earth radii and a third electron belt centred at $L = 2.15$ Earth radii, showed up in CRRES [R92, R93] and also in other satellite measurements [R94]. Whereas until then no magnetospheric electrons with energies above ≈ 10 MeV had been observed, electron energies exceeding 30 MeV were measured during this event. Although in these regions – apart from transits – no manned missions are considered, these changes also affected the radiation dose received in LEO. But, the AE8 and AP8 radiation belt models, apart from the regular change between solar minimum and solar maximum, do not model such stochastic space weather-related changes of radiation exposures of space crews in LEO.

As far as radiation protection is concerned, these models also ignore the quite significant directional anisotropy of the trapped particle flux [R112]. In contrast to irradiation by GCR and SPE particles, which for practical purposes can be treated as isotropic fields – the latter at least after the initial phase and in LEO, apart from Earth's shadow – the anisotropy of trapped radiation can be exploited for optimization of shield mass effectiveness by arranging vehicle masses so that maxima in vehicle mass distribution are oriented in the direction of peak proton fluxes.

11.2.1.2 Solar particle radiation

In addition to its electromagnetic radiation, the sun emits a continuous stream of particulate radiation: the solar wind. At 1 AU, the intensities of these low-energy particles vary by 2 orders of magnitude between around some 10^{10} and 10^{12} particles $\text{cm}^{-2}\,\text{s}^{-1}\,\text{sr}^{-1}$. In terms of velocities, this particle stream is characterized by velocities between about $300\,\text{km}\,\text{s}^{-1}$ and $800\,\text{km}\,\text{s}^{-1}$ and more (in CMEs) [W5]. The corresponding proton energies of some 100 to a few thousand eV are far too low to affect radiation exposures of man (at $800\,\text{km}\,\text{s}^{-1}$, a proton has an energy of 3.34 keV). They will be stopped within the first few hundred Ångströms of skin. However, the temporal variation of the solar wind is the major driver which determines radiation exposure from GCR in space – at least within the inner heliosphere. The heliosphere itself can be defined as the domain of interstellar space which the solar wind can fill out. The magnetic field carried along with the solar wind exerts a similar shielding influence as does the geomagnetic field. In this case the shielding strength can be simulated in terms of a pseudo-electrostatic heliocentric potential against which the GCR ions have to work when entering the heliosphere from the local interstellar medium. This potential modifies the GCR energy spectra to the same degree as does the interplanetary magnetic field [R12].

On top of that rather smoothly varying low-energy particle stream which only indirectly affects radiation exposures, occasional energetic SPEs have the potential to

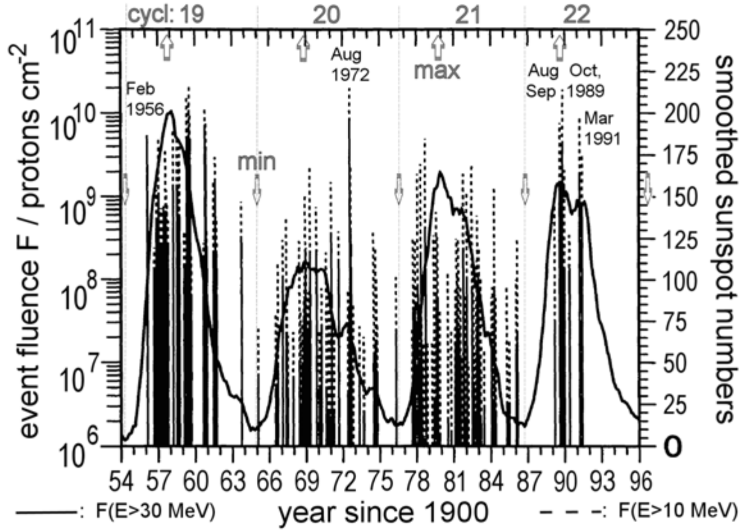


Figure 11.8. Occurrence of major and extreme solar particle events in solar cycles 19–22.

expose space crews to life-threatening doses. Apart from the fact that magnetic field configurations in sunspots – and hence the sunspots themselves – are a mandatory prerequisite for such energetic SPEs to occur, our present limited knowledge of the underlying mechanisms [R49] is insufficient to predict the precise occurrence and even less the energies and intensities of SPEs. Figure 11.8 shows that frequencies and intensities of SPEs vary in parallel to the sunspot number as the common measure of solar activity. Rare extreme events occur about once per Schwabe cycle, and usually near the maximum of solar activity, whereas during minimal solar activity even low-intensity and low-energy SPEs are rare. For the vast majority of SPEs, proton energies stay well below a few hundred MeV. Figure 11.9 demonstrates the enormous variability of the corresponding energy spectra. However, even such events as these can induce adverse skin reactions in astronauts if they become caught outside a habitat, since at energies above about 10 MeV protons can penetrate spacesuits and reach the skin or the lens of the eye. Depending on their intensities, they may induce erythema or trigger late radiation cataracts in the lens of the eye. While the latter take several years to develop and hence pose no threat to a safe mission completion, severe erythema may well induce performance decrements which could compromise mission success.

Since the onset of the space age, five SPEs with intensities and energies large enough to jeopardize crew health behind normal or even enhanced spacecraft shielding have so far been observed. Figure 11.10 displays integral energy spectra for these large events as they have been measured by satellite instruments. For a sixth event – that of 23 February 1956 – the energy spectrum has been inferred from an analysis of the count rates of terrestrial neutron monitors which at sea level recorded the flux of induced secondary neutrons. Such enhancements of neutron count rates are monitored in a world-wide net of neutron monitor stations, a selected subset of

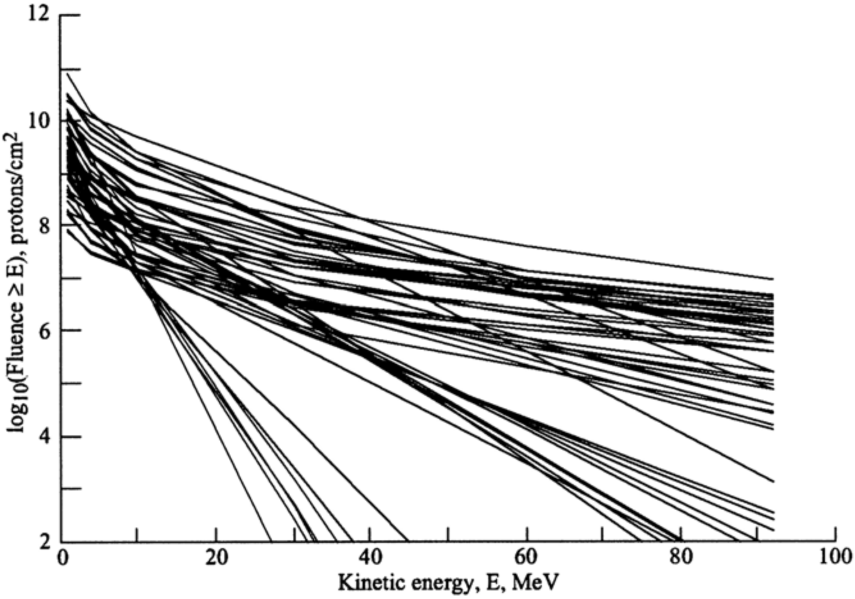


Figure 11.9. Compilation of SPE proton energy spectra, highlighting their extreme variability regarding intensities and spectral shape.

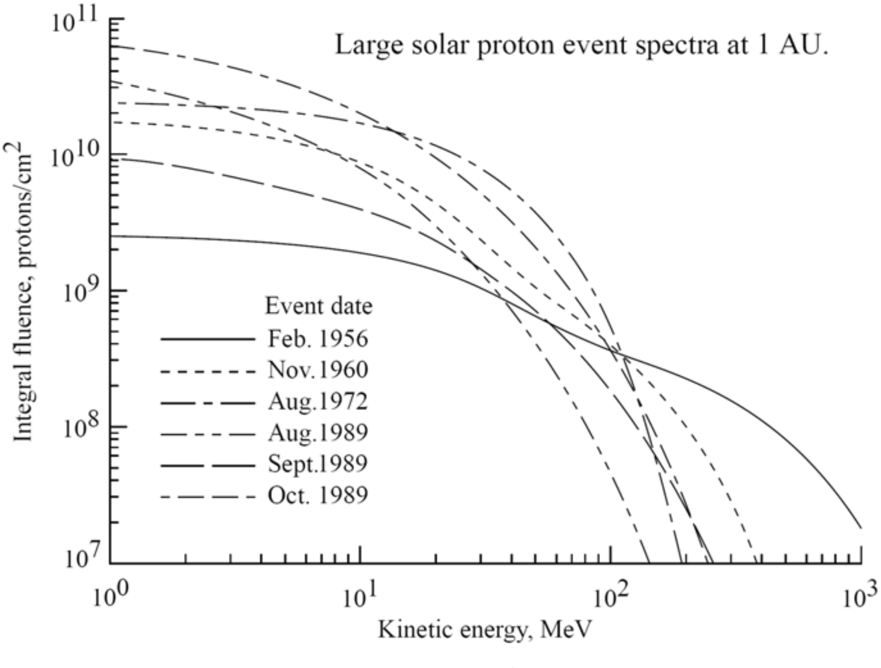


Figure 11.10. Integral energy spectra of extreme solar particles events.

which is forming the so-called Spaceship Earth [R96]. So-called ground level enhancement (GLE) events indicate that in the associated SPE proton, energies above about 450 MeV were sufficiently numerous to raise the neutron flux at sea level by at least 5%. A comprehensive list of GLEs observed since 23 February 1956 (GLE No. 5) and 14 July 2000 (GLE No. 59), together with all neutron monitor stations where these events were observed, is provided by the Australian Antarctica Data Centre [W7]. Among all these GLEs, the enhancement by GLE No. 5 in Leeds (latitude 53.83° N, longitude 358.42° E, altitude 100 m, cut-off rigidity Pc = 2.20 GV) reached a peak value of 4,581% above the pre-event count rate, whereas for other SPEs the enhancement very rarely exceeds a 100% increase.

The energy spectrum that has been reconstructed for the SPE associated with GLE No. 5 is much harder than so far measured by satellites. In shielding design calculations for long-term space missions – particularly outside the terrestrial magnetosphere, such as missions to Mars – this spectrum until now is taken as the worst-case reference spectrum [R16]. However, recent isotope analyses of ice cores [R17, R18] suggest that an even more intensive and harder SPE was associated with the flare which R. C. Carrington observed with his bare eyes on 1 September 1859 [R19]. According to this analysis, this SPE appears to be the most energetic event during the last 500 years. Accordingly, it has most recently been proposed [R20] to adopt the event associated with the Carrington flare as the historically established worst-case SPE for space mission design and radiation protection assessments. However, the absence of any spectral information for this event places rather serious uncertainties on such appraisals. At present we can only guess the spectral shape of the SPE associated with this flare. From analyses of nitrate concentrations in Arctic ice cores [R17, R18], its total event fluence was estimated as 1.88×10^{10} per cm^2 for protons with energies above 30 MeV. Combining this fix point with the spectral shapes of the large measured SPE spectra in Figure 11.10 yields the range of 'feasible' spectra for the Carrington SPE as shown in Figure 11.11. It has recently been demonstrated [R24] that a Weibull distribution function in energy provides a more satisfactory approximation over a wider energy range of empirical SPE spectra than the previously adopted representations as exponentials in rigidity or energy. With the parameters determined from such approximations, the spectra in Figure 11.11 were established. Before September 1989, the spectrum of the 4 August 1972 SPE had been considered as a probable example of a worst-case event, since in the lower energy range it exhibits by far the largest intensities (Figure 11.10). For the comparatively lightly shielded Apollo vehicles, this August 1972 event – squarely happening amid the Apollo 16 and Apollo 17 missions (see Figure 11.34) – could have resulted in dire mission termination had the launch of one of the latest two lunar missions been shifted by a few months, as could easily have happened. Later, when manned missions became more frequent and mass shields became heavier, the fluxes at higher energies became more important for risk assessment. Therefore, the change to the much harder 23 February 1956 SPE as a prototype of a worst case became indicated. In actuality, its poorly determined spectral shape was mimicked by the spectrum of the 29 September 1989 event. The total flux was scaled by a factor of 10, which corresponds to the ratio of the ground level enhancement profiles of

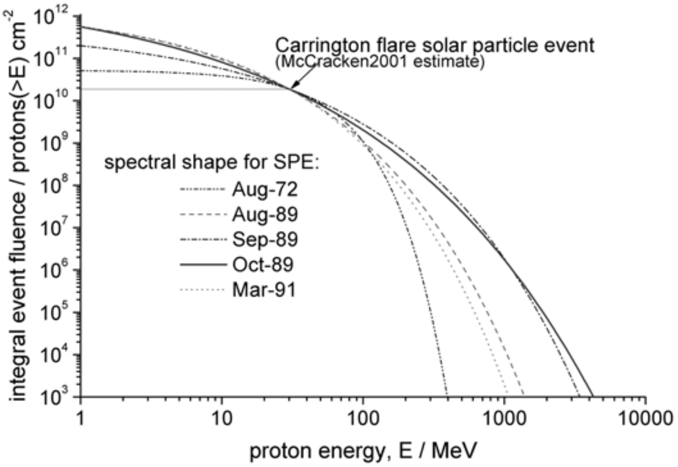


Figure 11.11. Spectra of known extreme solar particle events shifted so as to match the estimated integral event fluence of the SPE associated with the Carrington flare, 1859.

neutron monitor count rates. The change to the Carrington flare event SPE as a still more conservative worst-case solar particle event which still has some empirical basis would just represent the choice of another flux scaling factor since – as Figure 11.11 demonstrates – the 1989 spectra remain the hardest among the so-far measured SPE spectra.

For long-term mission planning, in addition to the magnitude that a worst-case event can attain, the frequencies by which they can occur also become important. The cumulative frequency of modern events above a given total event flux for proton energies greater than 30 MeV, F, can be approximated by a power function with exponent -0.4 [R95] (Figure 11.12). If this dependence were to continue into higher fluxes, then quite substantial probabilities would result to run into 10 or 100 times larger fluxes than so far observed. Fortunately, the steeper decrease (with an exponent of -0.9) which had been previously derived from isotope analyses of lunar surface material has recently been corroborated by analyses of NO_3 concentration profiles in polar ice cores [R17, R18]. In fact, the ice-core data in Figure 11.12 would even be compatible with a decrease by a power of -2.9, as it has been estimated from ^{14}C determinations in tree rings [R95].

11.2.1.3 Galactic radiation

Historically, space radiation was detected as Höhenstrahlung by Victor Hess's balloon measurements in 1911–1913, which won him the Nobel Prize in 1936. After several other discoveries made in this Höhenstrahlung and awarded this honour, in 1948 Phyllis Freier and her colleagues Lofgren, Ney and Oppenheimer [R56, R57] accomplished the latest significant discovery: the detection of tracks of galactic heavy ions in nuclear emulsions and in a Wilson chamber flown in high-altitude balloons. A decade later the space age dawned, and apart from the discovery

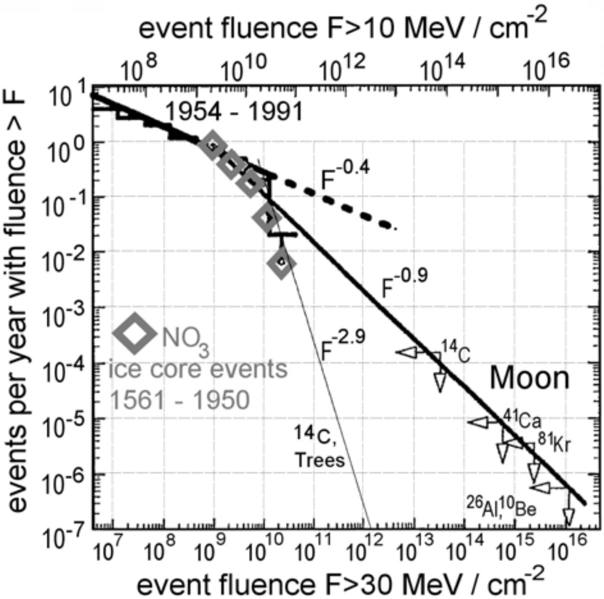
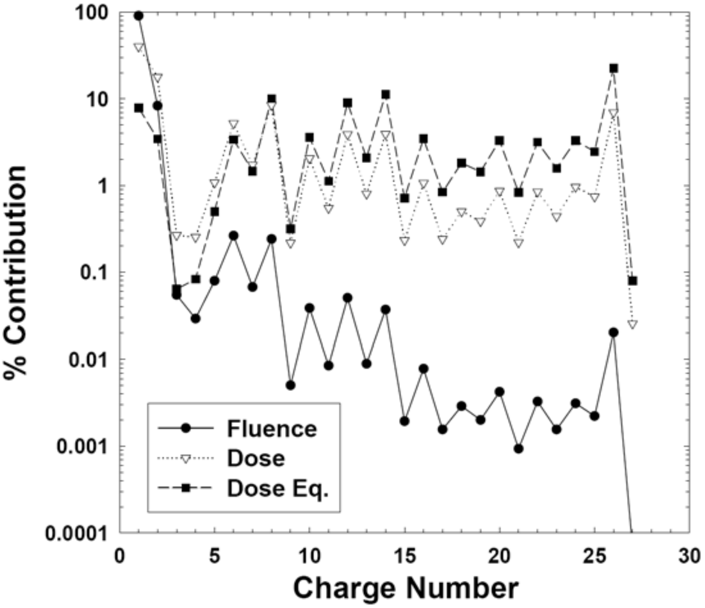


Figure 11.12. Event frequency distribution of total proton event fluences, estimated from various sources.

of the Van Allen belts and their detailed mapping and spectral probing, a major effort went into the elucidation by satellite measurements of the particulate and energy spectra of this component of cosmic radiation. In part this commitment arose from early conjectures regarding the potentially inordinate radiobiological importance of this radiation [R58, R59]; and, of course, manned space flight was already seriously envisioned at that time. In fact, one of the conjectures has been [R59] that astronauts might 'see' these heavy ions with their closed dark-adapted eyes; and indeed, about 15 years later that prediction came true. Already before the discovery of the galactic heavy ions, findings from radiobiological high-altitude experiments had provoked the speculation that the cosmic radiation is radiobiologically a 'radiation sui generis' [R60]. Results from subsequent balloon experiments in the 1950s [R61, R62, R63, R64] tended to substantiate this wariness. When Apollo astronauts reported that they experienced light flashes with their closed eyes when trying to fall asleep, this was immediately apprehended as a confirmation of the earlier prediction. The uneasiness that beyond this phenomenon other unknown radiobiological properties of this radiation component might pose disproportionate health risks to manned spaceflight, unleashed considerable means to address this problem. As a first step the particle and energy spectra of these GCR particles had to be established with sufficient accuracy.

Figure 11.13 presents the particulate composition of the GCR, which by now has been fairly well established. Although the atomic numbers of all elements have been demonstrated, the intensities beyond iron are too low to be of any concern for



Percent contributions from individual galactic cosmic ray elements for the particle flux, dose, dose equivalent at solar minimum.

Figure 11.13. Elemental composition of galactic heavy ions and their weighted contribution to absorbed and equivalent dose.

protection concerns. However, for radiobiological applications the raw distribution by number frequency is misleading. In a zero-order approximation the energy deposited by these GCRs, and hence the dose (see Section 11.3.1) is relevant for the biological effect. Since the energy deposited (dE/dx) is roughly proportional to the square of their atomic number, Z^2, the correspondingly weighted number distribution in Figure 11.13 raises the importance of the heavier ions to nearly the same level as that of the light ones. Further weighting with the proper radiobiological effectiveness (again, see Section 11.3.1) finally causes the heavy ions to become more important than the light ones.

As regards the energy range which is covered by GCR as shown in Figure 11.14, only a negligible fraction of that range – energies between about 10^7 and 10^{12} eV – is relevant for space radiation protection. The deviation in the lowest energy range of the energy spectrum from the line marked ISF (interstellar flux), however, reveals a feature of major importance for radiation protection in space. It reflects the modulation by the solar wind of the GCR intensity inside the heliosphere. This is displayed in more detail in Figure 11.15 which covers the energies and atomic numbers relevant for radiation protection needs. Between solar minimum and solar maximum conditions, the energy of the peak intensity is typically shifted by 500 MeV/n or more to higher energies. The peak intensities are attenuated by factors of 6 for iron ions to about 10 for protons. The flux modulation by the solar activity affects energies up to

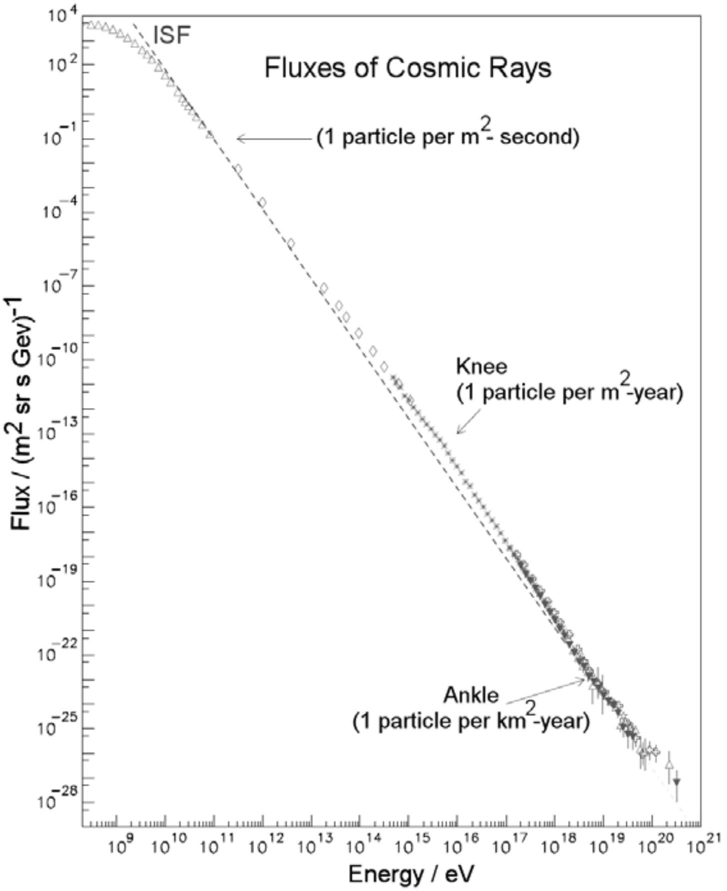


Figure 11.14. Differential energy spectrum of galactic cosmic rays at 1 AU, reaching up to the largest energies observed.

about 30 GeV. Phenomenologically, the modulation of GCR fluxes can be simulated by a heliocentric pseudo-electrostatic potential against which the heavy ions have to work their way from the border of the heliosphere into its interior. A convenient data source for deriving the value of this potential is provided by the count rates of neutron monitors. They continuously monitor the flux of secondary neutrons which near sea level constitute the remains of the cascade of secondary reaction products set in motion by the impact of primary cosmic radiation on the top of the atmosphere. The upper curve of Figure 11.16 (see colour section) shows, for the neutron monitor in Apatity (longitude 33.33° E, latitude 67.55° N, altitude 177 m, cut-off-rigidity P_c: 0.57 GV), the temporal profile of the neutron count rates as they reflect the variation of the GCR flux impinging on the atmosphere. The lower curve displays the variation of the corresponding heliocentric potential as it has been derived for the CARI code [W11], which needs this input to calculate dose rates

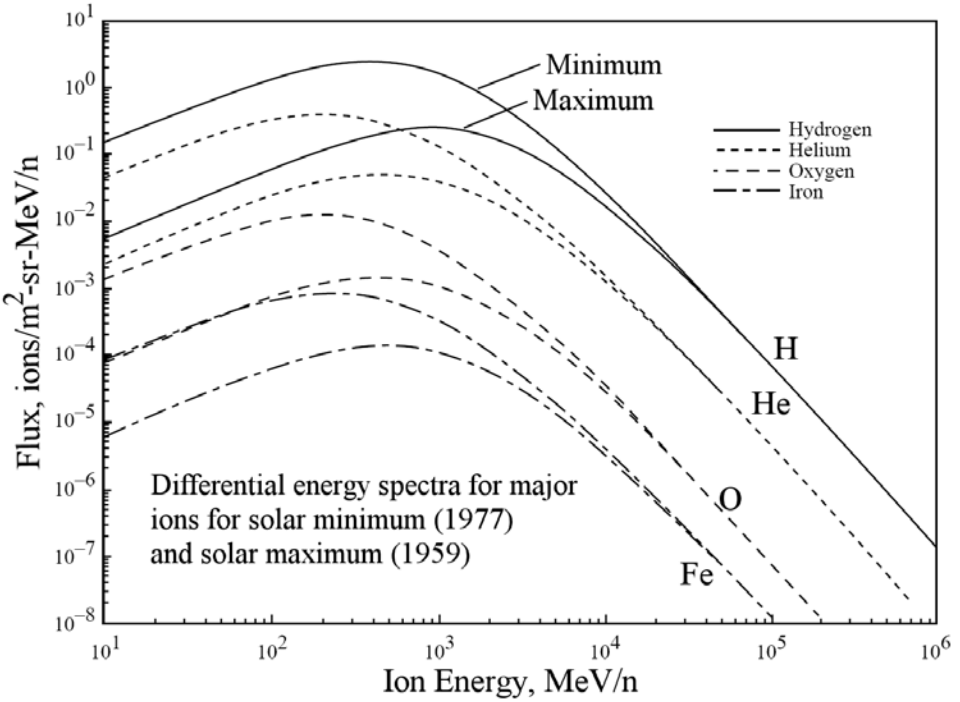


Figure 11.15. Modification by solar activity of galactic heavy ion energy spectra at 1 AU, as relevant for radiation protection.

for air-crew exposure. The full profile of the heliocentric potential covering the era of commercial jet air flight is shown in Figure 11.17. During solar minima the heliocentric potential settles around 350–400 MV. The amplitudes attained during solar maxima scatter widely between about 750 MV and more than 1,800 MV. The largest two values for the monthly average value of the heliocentric potential of 1,872 MV and 1,812 MV occurred in June and July 1991. In June a series of two ground-level enhancement events were observed in the neutron count rates of the terrestrial monitor stations (GLE 51, GLE 52 on 11 and 15 June), indicating that significant fluxes of solar protons above about 450 MeV were present in the associated SPEs. However, the enhancements were moderate or even minor as compared to the enhancements in August–September–October 1989, when two of the largest GLEs of the space era were observed and when the heliocentric potential reached a maximum value of only 1,339 MV. Obviously, the GCR flux responds to the low-energy solar wind, with its much larger intensities, rather than to the high-energy coronal mass ejection (CME) and SPE fluxes (Figure 11.17).

11.2.1.4 Synoptic view of primary components

Figure 11.18 (see colour section) provides a synoptic view of the three major primary sources of space radiation in terms of their integral energy distribution. Together

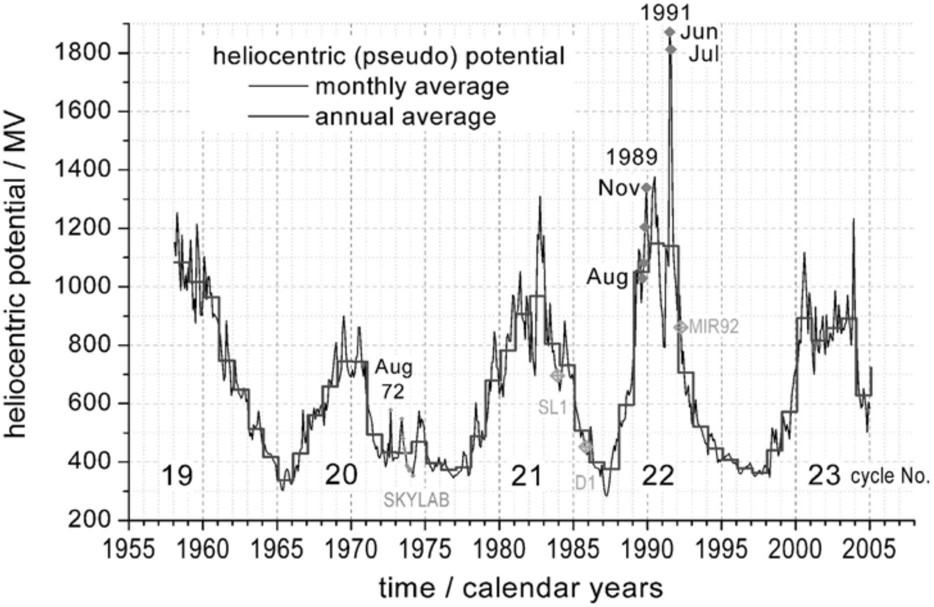


Figure 11.17. Variation with the solar cycle of the heliocentric potential as a modifier of the GCR flux at 1 AU. Several events which are discussed in the text are also marked.

with the scheme of their spatial distributions in Figure 11.1 (see colour section) and their elemental composition in Figure 11.13, this figure characterizes the physical features of the fields of ionizing space radiation as they are relevant for space radiation protection. The aspects of their regular temporal variation with the solar cycle are schematically indicated by the variation of the solar wind intensity by more than 2 orders of magnitude. Despite a discrepancy in energies between solar wind and GCR protons by more than 5 orders of magnitude, the energetically inferior solar wind protons imprint their behaviour on the full energy range of GCRs which is relevant for space radiation protection. The variability of trapped protons fluxes comprises both their variation with the position in the belts and their modulation by solar activity. The irregular stochastic behaviour of SPE spectra is reflected in the huge spread of energies and intensities that have been observed for this radiation source. However, events as represented by the right-most spectra in Figure 11.18, approaching energies and intensities of GCRs, are extremely rare.

The typical spectral fluxes for the other components – auroral electrons, trapped electrons and trapped protons of the outer Van Allen belt – can be ignored as far as radiation protection is concerned. Figure 11.19 (see colour section) provides the reason for this. Since man in space will always be surrounded by some minimal material shielding – at least a space suit during extravehicular activities – only particles can evoke radiobiological effects with energies that enable them to penetrate at least this shielding. In Figure 11.19 the range of shield thicknesses between 0.2 and 5 g cm^{-2} aluminium is marked corresponding to a minimal shield of spacesuit and the thickness of a 'normally' shielded vehicle, including its equipment. In addition to the energy spectra of Figure 11.19, the range energy dependence of electrons – protons as the lightest, and Fe as the heaviest ion – are given. Neglecting Bremsstrahlung photons, only components with energies to the right of the intersection of the range energy curves with a given shield thickness can reach at least the skin of astronauts. This finally justifies the restriction to inner trapped protons, SPE protons and GCR heavy ions as the components that are relevant for space radiation protection. It should, however, be kept in mind that as far as trapped radiation is concerned, this could change should man consider to sojourn in orbits around planets with much stronger magnetic fields than the terrestrial.

11.2.2 Magnetic and material shielding

Above, the minimal energies have been determined which primary space radiation components must have in order to penetrate shielding material so that they can pose direct radiation health risks for man. In orbits around planets or moons with magnetic fields, an additional shielding effect arises from the deflection of charged particles by the Lorentz force from locations inside the field, as indicated in Figure 11.2.

11.2.2.1 Transport through shielding matter

Upon penetration of primary components through shielding material, the loss of projectile energy is not the only modification of the radiation field. In particular at the higher energies of the GCR heavy ions, nuclear reactions with the nuclei of the shield material lead to fragmentation of the projectile as well as to spallation of the target nuclei, giving rise to numerous secondary projectiles. As a result, behind lower shield thicknesses even an increase of the flux of energetic particles with an associated increase of radiation dose can occur instead of the intended attenuation. This build-up of radiation doses upon penetration of high-energy radiation into absorbing material is quite a general phenomenon. In the Höhenstrahlung this build-up was detected by G. Pfotzer [R98, R99] when he measured ionization density as a function of altitude. Between 20 and 25 km altitude (corresponding to about 56 g cm^{-2} residual air layer), a maximum was attained, with a further steady decrease towards space. Figure 11.20 (see colour section) shows a schematic view of the cascade of secondary reactions which develops from the top of the atmosphere when a high-energy GCR proton hits an atmospheric nucleus. For radiation protection the identification of the radiation quality of these components is important. Electrons, photons and μ-mesons are sparsely ionizing components, whereas protons, π-mesons and especially neutrons engender high spatial densities of ionized atoms in the irradiated material and are therefore more effective in damaging biological systems (see below). For the transport of these high-energy radiation fields, various computer codes are available. It is interesting to observe the 'evolution' of results computed by one of the earlier codes, HZETRN [R26], as

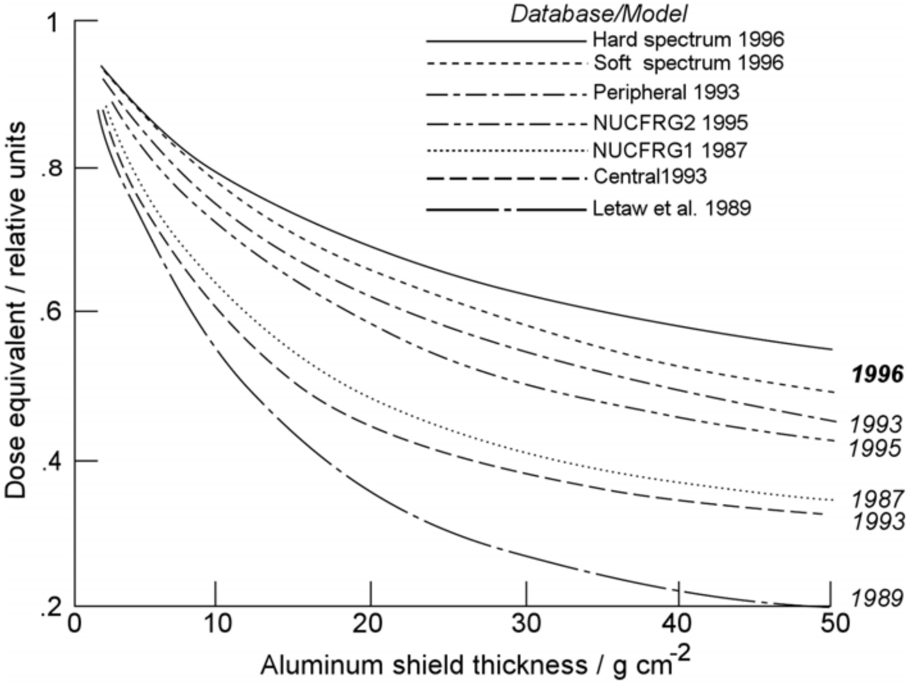


Figure 11.21. Development of important databases for transport codes for heavy ions through shielding matter.

displayed in Figure 11.21. The underlying code tackles the transport problem as a numerical solution of a Boltzmann equation into which the problem can be cast. In the early 1970s the first attempts towards numerical solutions for realistic shield geometries and compositions were made, with significant advances regarding the numerical algorithms. However, the rate at which the results converged towards the 'true' solution was dominated by the truthfulness by which the energy dependence of the copious inelastic reaction cross-sections was modelled. It is essentially the progress in increasingly more realistic cross-sections which is reflected in Figure 11.21. The improvement of the cross-section data of course depended on the availability of experimental data from appropriate accelerator beams. A similar solution for the less taxing problem of transporting energetic trapped and solar protons is provided by the BRYNTRN code [R25], which does not have to cope with the fragmentation of the projectiles. Today, alternative approaches rely on Monte Carlo codes which simulate, by random number generators, the reaction cascade depicted in Figure 11.20. Of course, these brute force solutions – in contrast to the HZETRN approach – also depend on the availability of accurate cross-section data. At the Centre Européenne pour la Recherche Nucléaire (CERN) in Geneva, one of the most widely used high-energy physics transport Monte Carlo simulation packages, FLUKA [W9], has been developed since the early 1960s. On top of this core, also at CERN, a toolkit, GEANT4 [W8, R100], has been developed for applications from medical exposures to space physics studies. Another widely used tool, similar to but somewhat less rigorous than FLUKA, is the LUIN code developed in the late 1970s [R101]. Toolkits which transport the GCR component down through the atmosphere to aviation cruising altitudes especially rely on such Monte Carlo codes. The European programme EPCARD [W10] builds on FLUKA-generated data, whereas the US programme CARI uses the latest LUIN version, LUIN2000 [W11]. Despite the progress already made, the application to long-term missions to, for example, Mars, needs further improvements in the reaction cross-section database [R97].

Finally, for the assessment of radiation doses received by human space or air crews not only the transport through the shielding of the spacecraft or the atmosphere, but also the self-shielding of the human body, has to be taken into account if, for example the exposure of the blood-forming organs (the bone marrow) has to be determined. Still widely in use for providing the required self-shielding distributions for internal organs by ray-tracing is the computerized anatomical man (CAM) model which has been developed for space applications [R27]. The cascade of high-energy nuclear reactions depicted in Figure 11.20 (see colour section) also takes place in the irradiated tissue. The cell in whose interior such a nuclear interaction has taken place experiences a peculiar spatial and temporal distribution of ionization events of its atoms or molecules which in its natural environment it never encounters. When these events occur in nuclear emulsions, the well known multi-pronged ‘nuclear reaction stars’ emerge after development of the film.

In the case of thicker shields, the neutrons – which, as primary components of space radiation, are negligible – can become a noticeable source of radiation exposure by themselves. This occurs not only in heavier shielded spacecrafts but also on those planetary or lunar surfaces which lack an atmosphere thick enough to attenuate this radiation source to the level which at sea level we enjoy due to protection by the $1\,\mathrm{kg\,cm^{-2}}$ air layer. On the surface of Mars, and even more so on the Moon, this secondary ‘albedo’ neutron component emerging from the ground contributes significantly to the overall exposure – in particular so, since, radiobiologically, neutrons belong to the more damaging types of (indirectly) ionizing radiation.

The same holds for the radiobiological effectiveness of the nuclear reaction stars as another type of densely – though secondary – ionizing space radiation. Among all components of space radiation, their radiobiological properties are probably the least understood.

11.2.2.2 Geomagnetic shielding (transmission function)

In LEO as well as in aviation, in addition to transport through shielding material a second shielding mechanism has to be incorporated into the transport of the primary GCR or SPE ions. Whereas the geomagnetic field on the one hand is responsible for the added radiation exposure in LEO from trapped radiation, on the other hand it causes a quite substantial reduction of radiation exposure, at least near the geomagnetic equator (which differs from the geographical equator). This stems from the

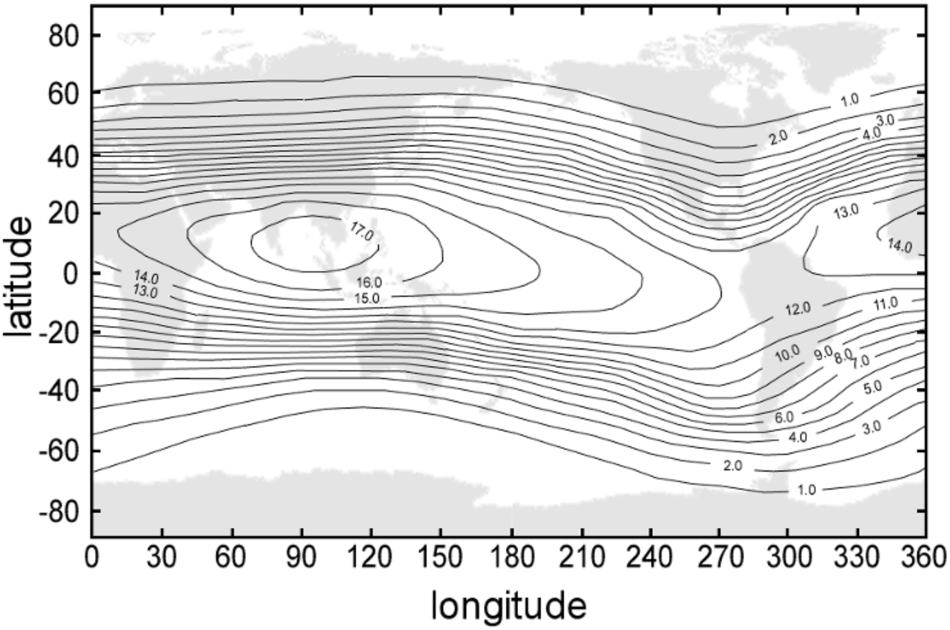


Figure 11.22. Map of vertical cut-off rigidities in GV for the geomagnetic field model of Epoch 2000.

deflection due to the Lorentz force of charged particles by the geomagnetic field as, illustrated in Figure 11.2. As indicated in that figure, this shielding is minimal for locations near the magnetic poles and maximal near the equator. The geomagnetic shielding effect can be approximately expressed in terms of the vertical cut-off rigidity, which is the minimum rigidity a charged particle must have in order to reach a given point inside the magnetosphere. Figure 11.22 provides a global map of the vertical cut-off rigidities for the geomagnetic field model of 2000. For a homogeneous dipole field, the iso-rigidity lines would be parallel to the (geomagnetic) equator. The marked asymmetry with a peak above 17 GV of the cut-off rigidity in the Indian ocean (longitude 90° E, latitude 10° N) reflects the offset from the geographical centre of the magnetic centre by about 450 km in this direction. At the opposite side, in the South Atlantic this offset results in the corresponding subsidence of the lower fringes of the inner proton belt, thereby creating the so-called South Atlantic Anomaly (SAA). This is the reason for the already mentioned fact that the bulk of radiation exposure in most LEOs is accumulated there.

For a given orbit the shielding due to this effect is expressed by the so-called geomagnetic transmission function or factor, which specifies the fraction of the GCR or solar particle flux of a given energy (or momentum) which has access to this orbit. Figure 11.23 demonstrates the dependence of the geomagnetic transmission function for a circular orbit at 700 km altitude. For an orbit of 28.5° inclination, which for a large fraction evades the SAA, on average GCR ions with a momentum below about

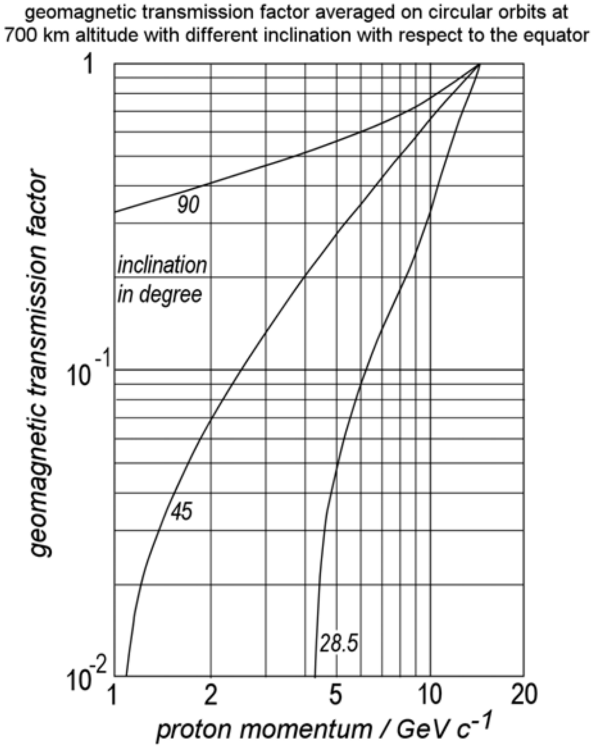


Figure 11.23. Dependence of the geomagnetic shielding on the orbit inclination during quiet solar/geomagnetic conditions.

4.2 GeV c^{-1} do not reach a spacecraft. For 45° inclination this threshold drops to about 1.1 GeV c^{-1}, whereas for polar orbits at least 20% of the lowest energies always have access to this altitude. On the other hand the shielding effect vanishes for ions with a momentum above about 15 GeV c^{-1}, where at all inclinations 100% of the flux reaches this orbit. The transmission functions in Figure 11.23 do not, however, include the shadow effect of the Earth itself. When this is included, the geomagnetic transmission functions of Figure 11.24 result. They show the geomagnetic shielding effect for the Hubble Space Telescope (HST), which intentionally was also given the best shielded LEO at 28.5-degree inclination, with the effect that no ions below an energy of about 3 GeV can reach it. The shadow effect of the Earth reduces the flux of even the most energetic GCR ions by about 30%. An Earth observation satellite such as TERRA, on the other hand, must use a near polar orbit, and can therefore be accessed by particles of all energies. Its higher altitude also slightly reduces the shielding by the Earth's shadow. The high inclination of the International Space Station (ISS) – 51.6° – makes this manned spacecraft quite accessible to even lower energetic SPE ions, as demonstrated in Figure 11.25. To compensate for this drawback, the ISS crew needs protection by heavier mass shielding. This is particularly important, since in the case of geomagnetic disturb-

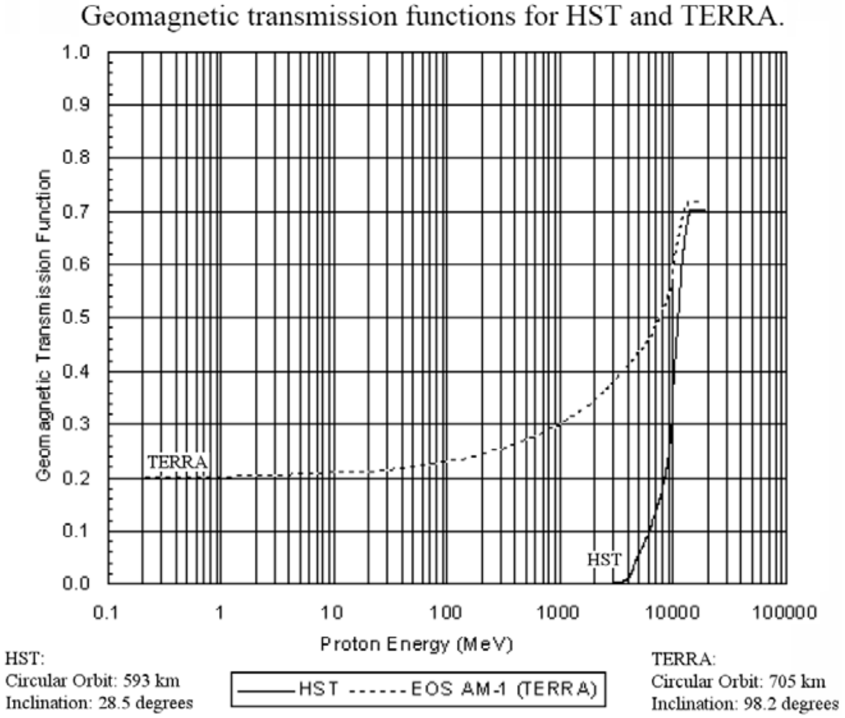


Figure 11.24. Shielding by the geomagnetic field for the TERRA satellite and Hubble Space Telescope, including Earth's shadow.

ances which often are associated with solar events, this geomagnetic shielding is further reduced. Figure 11.26 demonstrates this loss of geomagnetic shielding for the ISS for storms as characterized by the Kp index of global geomagnetic activity, which can vary between 0 and 9. Under such conditions a much larger fraction of SPE ions can reach the orbit of the ISS.

By now, all external factors determining radiation exposures in manned spaceflight and their modification by space weather have been presented and – at least cursorily – been discussed. Next the task arises to convert these data into a quantity which somehow can be taken as a measure of the biological effects that such exposures can engender.

11.3 RADIATION DOSIMETRY

The primary physical data, regarding the composition of the space radiation environment, are given in particle fluxes and energy spectra, and the obvious choice for a measure of exposure would therefore be the particle fluence itself. However, mainly for historical reasons the necessary radiobiological data to relate

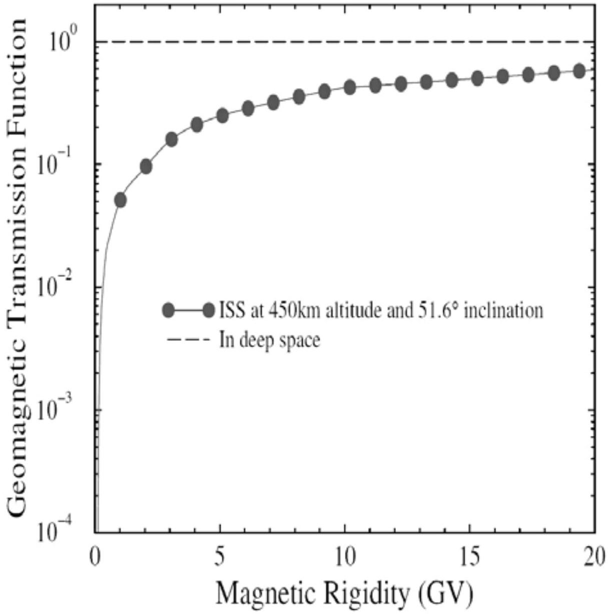
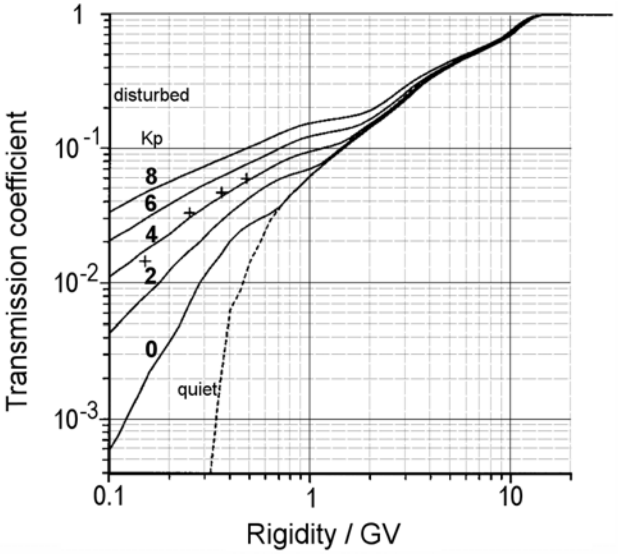


Figure 11.25. Geomagnetic shielding, including Earth's shadow, for the orbit of the International Space Station.



Calculated transmission functions for the ISS orbit.
Dotted Line for quiet time geomagnetic field (IAGA, 1992).
Solid lines for disturbed fields correspond to given Kp levels.

Figure 11.26. Variation of the geomagnetic shielding for the ISS orbit in disturbed geomagnetic conditions as expressed by the Kp index.

an exposure to its corresponding biological effect have been established in terms of another quantity into which the primary flux data have to be converted.

11.3.1 Measures of exposure

Conventionally, the amount of ionizing radiation to which a body is exposed is quantified by the amount of energy absorbed by the body divided by the body's mass. The unit for this 'energy dose' or 'absorbed dose' is the Gray (Gy), which corresponds to an absorbed energy of 1 Joule per kg (J/kg). The SI (Système International) unit of energy, 1 J, corresponds to the energy consumed by a 60-W incandescent lamp within no more than 1/60 of a second. The (heat) energy of 4.2 J is needed to raise the temperature of a spoonful of water (1 cm^3) by 1° C. A typically lethal homogeneous whole-body radiation dose of about 3 Gy (3 J/kg) would raise the temperature of the exposed person by no more than 0.0007° C. In contrast to this heat energy, which is evenly spread out among all molecules of the body, the absorbed radiation energy is localized around the single absorbing atoms. This is the basic explanation for the drastically different consequences of an otherwise utterly negligible amount of absorbed energy. Conceptually, the absorbed dose is essentially a surrogate for the spatial density of ionized atoms or molecules produced in the irradiated matter. In fact, the earliest measure of ionizing radiation had been the 'ionization dose', with the Röntgen (R) as its unit. It was defined as the amount of radiation which produces, in 1 cm^3 of dry air at standard conditions (NTP), ions with the total charge of 1 esu (electrostatic charge unit in the cgs system) which corresponds to 2.58×10^{-4} Coulomb/kg.

11.3.2 Relative biological effectiveness (RBE), equivalent dose

As soon as radiation sources other than the initially studied photons (X-ray and gamma-ray sources) became available to radiobiologists, it transpired that the absorbed dose is, after all, not an appropriate predictor of the ensuing biological effects. As long as dose effect curves were linear, a corrective factor – the relative biological effectiveness (RBE) – could be defined to account for the different effectiveness of different types of radiation or different radiation 'qualities'. The RBE with respect to a reference radiation – usually X-rays – of a radiation quality q is defined as the ratio of the respective absorbed doses $D_X(E)$ and $D_q(E)$, which yield the same biological effect E: RBE $= D_X(E)/D_q(E)$. For a given radiation quality q, such RBE values depend on the biological organism and the effect investigated. When new biological test objects other than biomolecules or microorganisms were studied, most of the dose effect curves turned out to be non-linear, with the result that the correction factor RBE became a dose-dependent correction function.

At least for radiobiological effects with linear dose effect curves, the product of the absorbed dose with the relevant RBE value yields an 'equivalent dose' which represents a biological weighted measure of exposure for a given radiobiological effect from a given radiation quality. In radiation protection the separate SI unit Sievert (Sv) is defined for the biologically weighted equivalent dose.

11.3.3 Ionization density, LET

Basically, the RBE is a purely empirical quantity, which in principle has to be determined experimentally for each pair of radiation qualities and biological effects. However, the unsurpassed complexity of space radiation renders such an approach unfeasible. Phenomenologically, the various forms of ionizing radiation differ with respect to the spatial (and temporal) density or concentration of ionization events; that is, ionized atoms or molecules. X-rays (Röntgen rays) and gamma-rays as examples of electromagnetic radiation, or electrons as particulate ionizing radiation, usually produce comparatively low concentrations of ionizations and are therefore classified as sparsely ionizing radiation. Energetic protons, neutrons, α particles and heavier atoms are examples of particulate radiation constituting densely ionizing radiation, among which neutrons can ionize only indirectly. In general, higher concentrations of ionization confer larger molecular and hence biological damage. This is the basis for the attempt to relate RBE to the spatial density of ionizations engendered in the irradiated medium. As an easily available surrogate for the spatial ionization density, the energy deposited along an infinitesimal path, the linear energy transfer ($\text{LET} = -dE/dx$) was taken as a physical quantity to systematize the dependence of empirical RBE data on the radiation quality. Figure 11.27 illustrates the various RBE(LET) functions determined experimentally for different radiobiological effects in one single test organism: seeds of *Arabidopsis thaliana*. The form of this dependence reflects the shape of most, if not all, radiobiological effects, although both the LET values, where the maximum RBE occurs as well as the amplitude, can vary largely.

11.4 RADIATION EFFECTS ON MAN

One classification of radiobiological health effects in man – as in animals in general – assigns them to one of two categories: early or late radiation effects. Late effects materialize years to even decades after exposure, while early effects can arise within some hours and may extend to several weeks. At extreme doses they can appear within minutes after exposure. The principal late effect in man is carcinogenesis; more specifically, mortality from late radiation-induced cancers. Late cancer mortality is the reference risk utilized in radiation protection to derive limits of exposure which might be considered acceptable.

The initially rather phenomenological distinction in the final analysis arises from a fundamental biological principle: the capability of living systems to counteract the assault of deleterious environmental agents by a manifold of defence and repair systems. Ionizing radiation is just one of the hosts of exogenous and endogenous deleterious factors which biological systems are able to counteract on the molecular, cellular, tissue and whole-system level. Only when the protective capacity of this hierarchy of defence systems becomes overwhelmed, detrimental effects can become manifest.

In particular, early effects of acute more-or-less instantaneous exposures arise

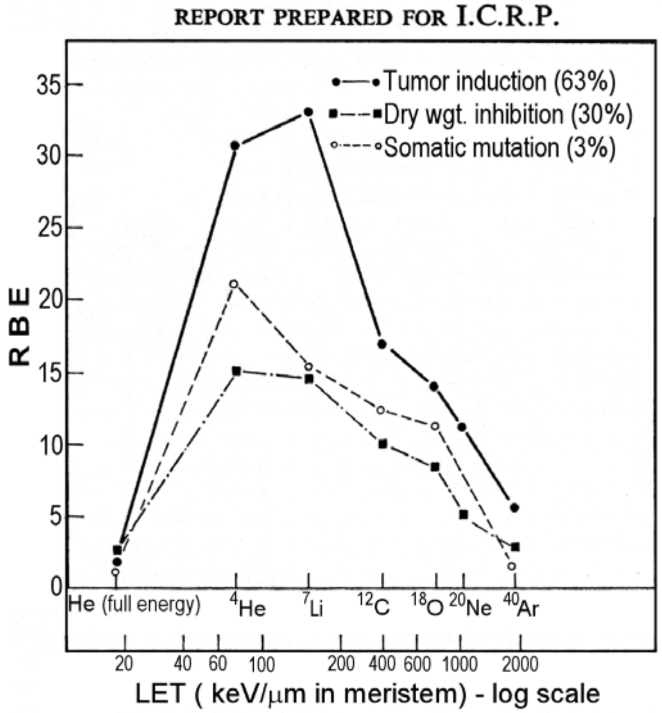


Figure 11.27. RBE values for accelerated heavy ions and 250-kVp X-rays at different LETs (shown on the abscissa) with respect to tumour induction, growth inhibition, and the induction of somatic mutations in seeds of *Arabidopsis*.

when the rate of cell killing or debilitation surpasses the rate of cell recovery and of tissue repopulation, and when the remaining cells can no longer maintain the minimum required tissue functionality. Historically, such effects which occur after a dose threshold is surpassed have been called ‘deterministic’, to denote the concept that beyond that threshold the severity of the damage increases with dose. Of course, the level of severity again is a stochastic event and depends on the defence capacity of the affected individual.

In contrast, the term ‘stochastic’ radiation effect is meant to reflect the idea that only the probability that an effect will arise as a function of dose, whereas the severity of the effect, such as cancer, is largely independent of the dose. Whereas early ‘deterministic’ effects occur only after exposures to rather large doses at large dose rates, stochastic effects constitute the relevant late health effects after chronic (low dose rate) exposures to low doses.

The distinction between deterministic and stochastic radiation effects constitutes another frequently used classification of radiation effects, although the intersection with early and late effects respectively is rather large. In the context of radiation protection, irradiation at dose rates below 50 mSv/a and doses below 200 mSv are considered chronic low dose exposure where stochastic effects prevail. The high end above dose rates of 3 Sv/h and doses above 1.5 Sv constitutes the realm of deterministic effects where a dose above 3 Sv marks the border to ultra-high exposures where early mortality becomes likely.

11.4.1 Radiation weighting factors and quality factors

In radiation protection, the different radiation quality of different types of ionizing radiation is accounted for by radiation weighting factors or quality factors instead of the RBE. In contrast to the empirical nature of RBE, radiation quality factors are determined by consensus of expert committees which attempt to consider a large number and widely scattering values of RBE for effects that are relevant for cancer induction, promotion and progression. Numerical values assigned to these radiation weighting factors, w_R, range from 1 for electron-, X- or gamma radiation, 5–20 for neutrons of different energies, and up to 30 for heavy ions of the primary galactic radiation. In terms of LET, Figure 11.28 displays the currently accepted dependence of the quality factor as it had been agreed upon by the ICRP in its currently valid 1991 recommendations [R8] as well as the dependence recommended in the earlier 1977 report [R9]. The difference between the two representations mainly reflects changes based on microdosimetric theory regarding the interpretation of RBE values rather than new empirical data. The product of the radiation weighting factor for a given radiation component, R, with the absorbed dose, $D_{R,T}$, deposited in a given organ by this component, yields the equivalent dose for this organ, $H_{R,T} = w_R * D_{R,T}$, which corresponds to the component R. Summation over all components R yields the total organ equivalent dose, $H_T = \Sigma H_{R,T}$, which an organ or tissue T has incurred in that radiation field. For a continuum of radiation qualities such as in the space radiation field, the summation over R is to

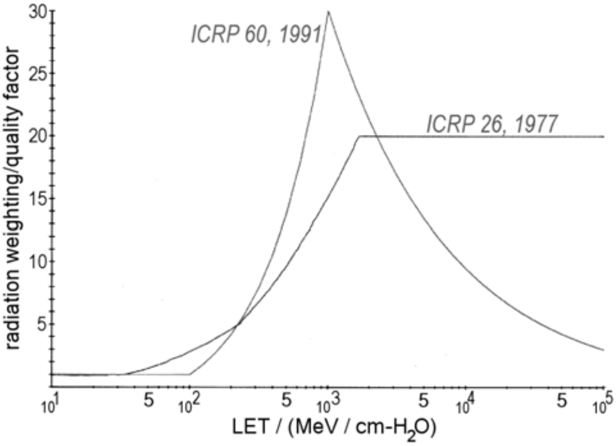


Figure 11.28. Radiation quality/weighting factor functions of LET, as defined by ICRP in 1977 and 1991.

be replaced by an integration over the LET range. If, in particular, at a given point in a tissue, $\Phi(L)$ denotes the integral planar fluence of charged particles with an LET $> L$, then the equation

$$H = k * \int Q(L) * |L * \partial\Phi(L)/\partial L| * dL \tag{11.1}$$

yields the equivalent dose deposited at the given point, with $Q(L)$ representing the quality factor dependence on LET (as in Figure 11.28), and k being an appropriate unit conversion factor. All physical dosimeters which attempt to 'measure' equivalent doses exploit this basic equation in that they somehow establish estimates for $\Phi(L)$.

It has to be stressed once more that radiation weighting or quality factors, w_R or Q, pertain – like any RBE – to a specific radiobiological effect only; that is, to cancer induction by low dose and low dose rate exposures. For other health effects or other exposures – for example, acute exposures to SPE protons – the equivalent dose calculated by equation (11.1) may not be appropriate. In 1990 the International Commission on Radiological Protection (ICRP) made a first attempt to address this problem in specifying RBE values for deterministic effects from exposures to densely ionizing α-particles from incorporated radionuclides in case of accidental (high-dose) exposures [R110]. Most recently, the ICRP reiterated the limitation regarding the applicability of quality factors: 'Those factors and the dose-equivalent quantities are restricted to the dose range of interest to radiation protection, i.e. to the general magnitude of the dose limits.' [R23]. In space we have to live with a finite probability that this general magnitude will be trespassed in SPEs so that deterministic effects might ensue. The US National Council on Radiation Protection and Measurements, NCRP, in drawing on the ICRP recommendations [R110], has suggested such weighting factors for space radiation applications in its recent report on that topic [R89]. However, it appears that a satisfactory solution has yet to be achieved, as in its most recent report the ICRP once more states [R23]: 'In special circumstances where one deals with higher doses that can cause deterministic effects, the relevant RBE values are applied to obtain a weighted dose. The question of RBE values for deterministic effects and how they should be used is also treated in the report, but it is an issue that will demand further investigations.' Therefore, the assessment of the risk for early deterministic effects from SPE exposure suffers from the additional uncertainty that the proper values for the RBE pertaining to a given effect may be only poorly known. Under terrestrial conditions it is normally found that RBE values for deterministic effects are smaller than those for stochastic effects [R111], so that equivalent doses calculated with the Q values for stochastic effects would overestimate the health risk, thereby yielding conservative limits.

11.4.2 Tissue weighting factors

As regards cancer: in order to account for the different propensity of human tissues to develop radiogenic tumours as well as for their different lethality, a further set of weighting factors – the tissue weighting factors, w_T – are applied to the equivalent doses, H_T, to which the organs of an irradiated individual have been exposed. The correspondingly weighted sum of tissue equivalent doses, $E = \Sigma w_T * H_T$, yields the so-called 'effective dose', which in radiation protection is considered the relevant quantity to assess health (cancer risks) from radiation exposure. Like the radiation weighting factors, w_R, the tissue weighting factors, w_T, are the result of educated guesswork by expert committees; and again, like w_R they are subject to change with progressing knowledge.

Only after the quantities H_T and E have been determined for a given space mission, does the assessment of the ensuing health risks become possible.

11.4.3 Acute irradiation, early (deterministic) effects

Figures 11.29 and 11.30 (see colour section) display, as a function of dose, the probability for several of the early effects and one late radiation effect relevant in man to occur after acute homogeneous whole-body exposure to sparsely ionizing radiation. Mortality data for curves in Figure 11.30 mainly stem from observations on victims of radiation accidents, and morbidity data represented by the dose response curves in Figure 11.29 also arise from observations on radiation therapy patients. Sparse as data (especially on victims) are, the human observations have to be supplemented by general relations derived from controlled animal experiments. The general shape of the dose response functions shown in Figure 11.29 and Figure 11.30 is that of Weibull distribution functions, which phenomenologically appear to provide the best and most flexible approximations to the empirical data. Also, conceptually the nature of these functions, as the distribution function of extreme values, supports their application to radiation sickness, which becomes manifest when the repair capacity of the irradiated tissue/organism becomes overstrained.

The conditions in Figure 11.29 (see colour section) – anorexia, fatigue, nausea, vomiting and diarrhoea – constitute the symptoms of the so-called prodromic syndrome as the early warning signs (depending on the dose – within hours) that potentially life-threatening doses may have been incurred. When homogeneous whole-body exposures have led to skin erythema, Figure 11.30 (see colour section) predicts that the irradiated person will most probably die from failure of the haematopoietic system – the most frequent cause of death after radiation accidents – unless substantial medical interventions, including anti-inflammatory and antibiotic treatments and ultimately bone marrow transplantation, are undertaken. After such interventions, radiation victims can often survive approximately twice the normally lethal exposures. Cataract formation (opacities in the ocular lens) represents another important deterministic radiation effect in man and also an example of a late deterministic effect, since due to the kinetics of lens cells it usually takes years for these opacities to become manifest. As long as in space the means for medical interventions after potentially precarious exposures are unavailable, the dose response curves for minimal medical treatment are appropriate.

The qualification 'acute' exposure in this context means that the duration of exposure is short compared to the characteristic timescales of cellular and tissue processes by which living systems can counteract the insult by radiation. If such repair capacities can be activated and have time to operate, the corresponding response functions in Figures 11.29 and 11.30 (see colour section) will be shifted to higher doses. Therefore, a proper assessment of the health risks from SPE exposures must also take into consideration the actual temporal profile of the dose rates during such an event.

Note the repeated mentioning of homogeneous whole-body exposures. It is a characteristic feature of space radiation that its components can produce extremely inhomogeneous distributions of dose. Then the equivalent dose to a given tissue is relevant to assess the corresponding health risk. The referral to sparsely ionizing radiation of course reflects the fact that – with the exception of selectively incorporated α emitters – our preponderant terrestrial experience pertains to such radiation fields. What the appropriate relative biological effectiveness of the space radiation components with respect to the early effects shown in Figures 11.29 and 11.30 (see colour section) might be is to a large extent a matter of conjecture at the present time (see Section 11.4.1 and [R23]).

11.4.4 Chronic irradiation, late (stochastic) effects

In space radiation protection, the dose to the haematopoietic tissue or blood-forming organs (BFO) is often taken as a substitute for the proper effective dose – the dose predictive of the associated cancer risk. This choice is based on the fact that leukaemia is the cancer with the shortest latency period of 5 years or less, and that the bone marrow belongs to the most sensitive tissues regarding cancer induction by radiation. Solid tissue cancers, in contrast, have longer latency times which can extend to 30 years. The dose response functions for these health effects are derived from observations of the survivors of the atomic bomb onslaughts.

Whereas for leukaemia a definite (convex) curvature of the dose response function is more or less established, for solid tumours a straight line through the origin is considered a reasonable approximation to the true dose response function (linear no-threshold (LNT) postulate). However, Figure 11.31 demonstrates that even here a straight line from the origin to the higher doses, where effects are well established, is a poor approximation to the low-dose empirical data of the sub-population from Nagasaki. Nonetheless, the risk coefficients determined from such data (the slope of the LNT line in Figure 11.31 modified by dose rate corrections) constitute the ultimate basis from which exposure limits for terrestrial radiation workers, as well as for exposure to space radiation, are derived.

11.5 RADIATION PROTECTION EXPOSURE LIMITS

Once the appropriate measures of exposures have been determined – effective dose as a measure of the late cancer mortality risk from chronic and acute exposures, and the tissue equivalent doses for critical organs from acute exposures as a measure for the assessment of early morbidity – the health risks ensuing from these exposures can be

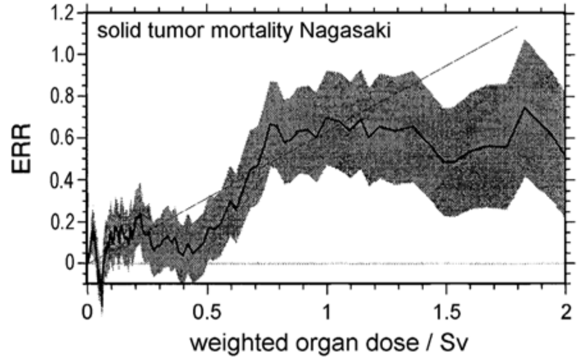


Figure 11.31. Excess relative risk of cancer mortality in atomic bomb survivors as the overriding data source for the estimation of risk from ionizing radiation. The solid line represents an approximation by the linear no-threshold (LNT) postulate.

assessed. The objective of radiation protection is to constrain these risks within levels deemed acceptable. From the knowledge of the proper dose response function, this acceptable risk level can be converted into a corresponding exposure limit. ‘Proper’, in this context, means that all relevant exposure conditions, as they pertain to the actual exposure situation in space, are properly incorporated in the dose response function applied. Self-evident as this might appear, in the practice of space radiation protection this proviso is usually only very approximately complied with, if at all. The radiobiological dose response relation is to a large extent determined by the many defence mechanisms against primary radiation damages which operate at the cellular level as well as above the cellular level; that is, on the tissue or immune system level. The established physiological changes brought about in man by microgravity, in particular in the humoral system, may well modify responses to radiation, especially late response after long-duration missions which would render invalid the derived dose limits. However, so far this aspect has hardly been addressed by suitable experimental work.

11.5.1 Chronic exposures, late cancer mortality

Since for long-term missions outside the terrestrial magnetosphere radiation exposures can reach levels in the order of 1 Sv, which are utterly unacceptable for terrestrial radiation workers, the error for risk prediction arising from the LNT approximation becomes less important. Based on the risk coefficients and the age–mortality profile for the incidence of spontaneous cancers, the total lifetime risk after exposure can be determined for a given radiation dose. By limiting this radiation-induced lifetime cancer mortality risk to 3%, the dose limit can in turn be established below which that 3% risk can be avoided. The 3% risk has in turn been adopted as acceptable by comparison with occupational mortality risks in terrestrial worker populations. Taking into account the already established risk of astronauts dying from technical failures during a mission, the increased risk of 3% to potentially die several years or perhaps decades later from a space radiation-induced cancer, appears reasonable. When the gender dependence of the risk coefficients is also taken into account, the following equation yields the age-dependent career exposure limits, $E_{\max}$, as they pertain to chronic space radiation exposures of US astronauts in LEO:

$$\begin{aligned} E_{\max}/mSv &\approx (\text{age}/a - 30) * 75 + 2000 \quad \text{for males} \\ E_{\max}/mSv &\approx (\text{age}/a - 38) * 75 + 2000 \quad \text{for females} \end{aligned} \tag{11.2}$$

Exposure limits set by other space agencies, such as the Canadian CSA, the Russian RSA, the Japanese JAXA and the European Space Agency, differ in some aspects from these NASA recommendations.

11.5.2 Acute exposures, early (deterministic) effects

Whereas the LNT postulate for cancer risk from chronic exposures implies that the risk can never by avoided totally, the response function for deterministic radiation effects from acute exposures allow for the assumption of a threshold below which the health effect can be avoided altogether. From such a dose response function as depicted in Figure 11.29 and Figure 11.30 (see colour section), approximate thresholds can be derived. After taking into account the reduction of deterministic radiation effects by protraction of exposure, the threshold values given in Table 11.1 have been recommended during various phases of the space age. Although radiation protection was an active concern in the pre-Apollo era, the radiation doses accumulated in those missions were too low to warrant the explicit regulation of dose limits. The first recommendations were therefore issued for the Apollo

Table 11.1. Radiation protection exposure limits in mSv, for prevention of early deterministic radiation effects in manned spaceflight.

Organ	Exposure timespan	Apollo§	NAS/NRC 1970 [R87]	NCRP 1989 [R88]	NCRP 2000 [R89]
BFO#	career	–	4,000	1,000–4,000$	400–4,000$
	annual	–	750	500	500
	30 d	2,000	250	250	250
Skin	career	–	12,000	6,000	6,000
	annual	–	2,250	3,000	3,000
	30 d	7,000	750	1,500	1,500
Ocular lens	career	–	2,000	4,000	4,000
	annual	–	380	2,000	2,000
	30 d	2,000	130	1,000	1,000
Extremities		9,800	–	–	–

§ maximum permissible single acute emergency exposure
$ depending on gender and age
# BFO, blood forming organs

missions. Obviously, the dominant concern during mission planning was to prevent exposures which by induction of early symptoms of radiation sickness might jeopardize the safe return to earth. Although the subsequent recommendations [R87] were still addressing primarily pioneering missions, the 30-day limits for the Apollo missions were already reduced by nearly an order of magnitude. Exposure limits in subsequent recommendations [R88] were designed under the premise that an astronaut should not unduly exceed the mortality risk of terrestrial workers. The recent change of the career limits in [R89] does not reflect a change of the risk deemed acceptable. Instead, it reflects the changes in the dose response function such as in Figure 11.31, due to the latest reanalysis of the data from atomic bomb survivors. In addition, a gender-specific larger risk for females materialized in these analyses.

11.6 IMPLICATIONS FOR MANNED SPACEFLIGHT

Radiation protection being an active concern from the very beginning of manned spaceflight, all manned missions were equipped with various radiation detectors that allowed the measurement of the total exposure accumulated during a mission, as well as the monitoring of dose rates during the mission. Initially only the energy dose or absorbed dose could be determined, but later analyses of the radiation fields allowed the estimation of approximate average quality factors by which absorbed doses might be converted into equivalent dose.

11.6.1 Approaches towards proper dosimetric techniques

Most of the radiation detector systems of the pre-Apollo era – in particular, active systems such as Geiger–Müller counters and ionization chambers – could not discriminate between the various sources of ionizing space radiation, and their reading only indicated absorbed dose. Even this measure of exposure suffered from systematic problems, since their sensitivity to the different space radiation components varied in often not fully documented ways. In order to provide a more comprehensive coverage of the different components of space radiation, during the Apollo programme the active systems were complemented by various passive detector systems [R104]. Diverse passive detector systems, responding to different components of space radiation by different physical interaction mechanisms, allow the identification of the contributions of these components to the total exposure.

Thermoluminescence detectors respond most efficiently to the sparsely ionizing component, primarily made up of energetic protons and then of secondary electrons, μ-mesons and photons. Separate analysis of this 'sparsely ionizing' category yields an average quality factor of 1.3 pertaining to the corresponding absorbed dose. Thermoluminescence detectors of different isotopic compositions allow the estimation of the absorbed dose produced by the neutron component, although here the response is restricted mainly to low-energy neutrons. Experimental data on the energy distribution of such neutrons is notoriously difficult to obtain. In the absence of sufficient data on their energy spectra, such neutrons are assigned an average

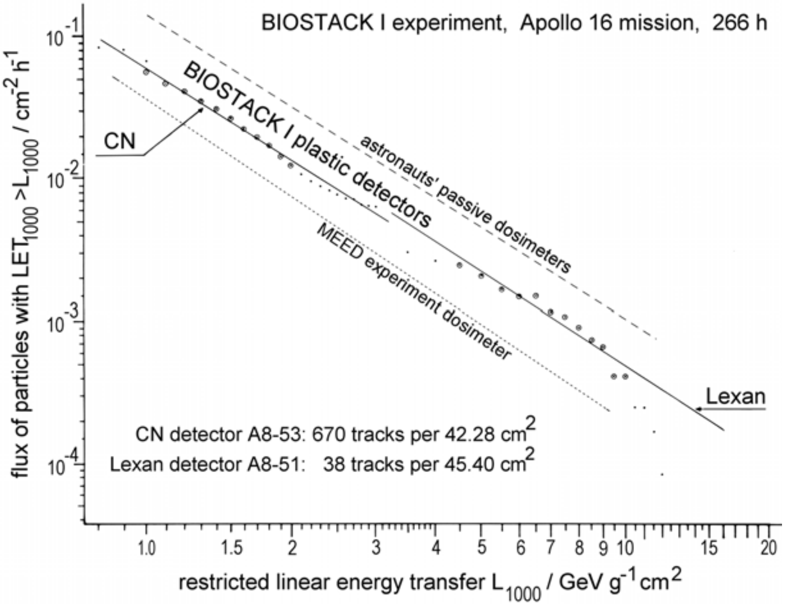


Figure 11.32. LET spectra of heavy ions in the Command Module of the Apollo 16 lunar mission. Astronauts' dosimeters were least heavily shielded, and MEED dosimeters were most heavily shielded.

quality factor of 20. Counting nuclear reaction stars in nuclear emulsions yields the volume density for stars with different prong counts from which the absorbed dose due to this component can be estimated. In matter dominated by light nuclei, such as tissue, the prongs arise mainly from secondary protons and α particles. Their energy spectra can be applied to estimate an average (nominal) quality factor of 5.75, pertaining to the absorbed dose produced by such nuclear-reaction stars. The absorbed and the equivalent dose due to heavy ions can be determined from LET spectra ($\Phi(L)$ in equation (11.1); for the calculation of absorbed dose, $Q(L)$ is set to one). These spectra can in turn be established from microscopic measurements on track etch cones generated by the passage of heavy ions through thin layers of plastic nuclear track detectors. The ratio of the absorbed and the equivalent dose yields an average quality factor applicable to GCR heavy ions with this respective LET spectrum. Figure 11.32 shows such integral LET spectra $\Phi(L)$ (or rather, their rates) of GCR heavy ions as they occurred in various experiment packages inside the Command Module of the Apollo 16 mission – one of the few manned missions outside the magnetosphere so far. In free space the spectra in this high LET range can reasonably be approximated by power functions where the different fluxes reflect the different shielding of the detectors. LET spectra in LEO are modified by the geomagnetic shielding, and have more complex shapes [R107].

Dosimetric techniques such as those just described [R105] constitute an expedient complement to the active detector systems, whose most important advantage, on the other hand, is that they can provide dosimetric information in real time, so that in case of severe space weather events, such as SPEs, protective or evasive measures can be taken.

11.6.2 Exposures during LEO missions

With the exception of the lunar missions, Apollo 8 and 10–17, man has so far not left the terrestrial magnetosphere, so that all other measured exposure data pertain to LEO missions. Furthermore, after the three manned Skylab missions and prior to the MIR and later ISS orbital stations, missions of the space transportation system (STS – the Space Shuttle) usually lasted two weeks or so at most. Table 11.2 summarizes essentially all manned US missions, with minima and maxima of total mission doses accumulated during the different phases of the manned spaceflight era. Obviously the exposure limits were never in danger of becoming infringed, although it deserves mentioning once more that in the case of the last two lunar missions, Apollo 16 and Apollo 17, this was due to fortunate circumstances (Figure 11.33). Annual dose rates during solar minimum, as they affect the crews of the MIR station or the ISS, however, can approach the limits set forth in Table 11.1. The largest total doses before permanent orbital stations became operative were received by the crew of the Skylab 4 mission [R103]. Table 11.3 gives the dosimetric details for this pioneering long-term LEO mission in order to demonstrate the quite satisfactory consistency of these measurements across the different crew-members as well as across the missions. The steady increase in dose rate by a factor of about 1.5 from Skylab 2 to Skylab 4 neatly reflects the noticeably steep decrease by about 200 MV of the heliocentric pseudo-potential during the terminal phase of solar cycle 20, as the highlighted monthly averages in Figure 11.17 reveal.

A compilation of the average dose rates during essentially all US manned space missions is given in Figure 11.33. The moderately shielded Apollo missions – the only missions outside the magnetosphere – were by no means exposed to the largest dose rates. Even the high-inclination orbits (40–60°), for which the geomagnetic

Table 11.2. Range of typical radiation exposures, in mSv, of astronauts measured during manned space missions.

	Mercury§ [R82]	Gemini§ [R83]	Apollo$ [R84]	Skylab§ [R85]	Shuttle§ [R86]	MIR# [1]	ISS# [1]
min	0.17	<0.15	6.8	24	0.17	1	0.5
	MA-8	Gem.-VIII	Apollo 8	Skylab 2	STS-2		Solar max.
max	0.42	12	33	116	24		1.5
	MA-9	Gem.-X	Apollo 12	Skylab 4	STS-61		Solar min.

§ absorbed dose converted to equivalent dose by average quality factor of 1.5
$ absorbed dose converted to equivalent dose by average quality factor of 4.5
# dose rate, mSV d^{-1}

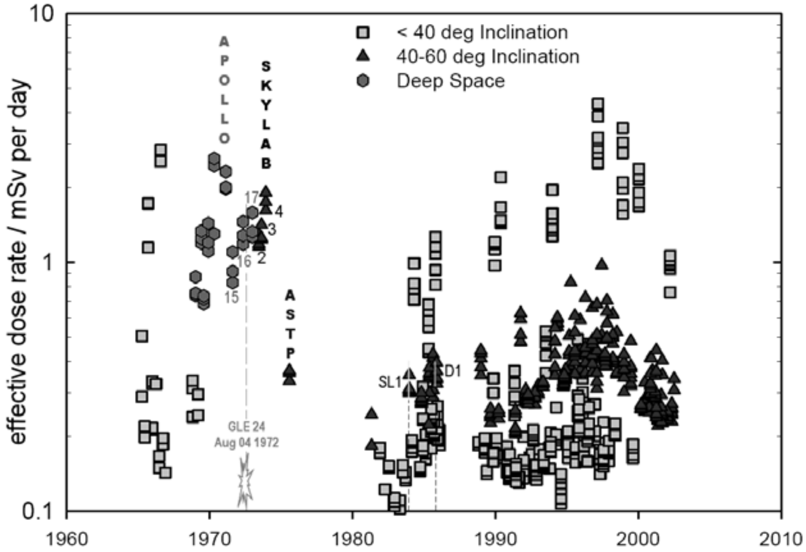


Figure 11.33. Compilation of dose rates experienced by US astronauts. Until the NASA–Mir and ISS missions, the only long-term missions were Skylab, and the only missions outside the terrestrial magnetosphere were the Apollo missions (Apollo–Soyuz Test Project).

shielding is poorer, did not experience the largest dose rates. The dominant factor in LEO orbits is the altitude which – apart from very low-inclination orbits – exposes the spacecraft to the protons of the inner Van Allen belt as the dominating source which is characterized by a very steep flux increase with altitude.

Results of the more comprehensive dosimetric approach, as described above, are presented in Figure 11.34 for the Spacelab 1 (SL1 mission, 28 November–8 December 1983) and for the first German Spacelab mission (D1 mission,

Table 11.3. Mission doses and dose rates measured in Skylab personnel dosimeters.

	Skin dose/mSV			Eye lens dose/mSv			BFO dose/mSv		
Skylab mission:	2	3	4	2	3	4	2	3	4
Crew member									
Commander	36.6	73.7	141.7	25.9	58.7	128.3	16.0	36.3	79.4
Pilot	31.5	78.9	161.0	29.0	67.4	108.8	17.9	41.7	67.3
Scientist Pilot	35.1	86.2	178.5	26.6	59.7	117.8	16.4	36.9	72.9
average	34.4	79.6	160.4	27.2	61.9	118.3	16.8	38.3	73.2
Rate/mSv d^{-1}	1.23	1.34	1.91	0.97	1.04	1.41	0.60	0.64	0.87
Missions, durations	Skylab 2: 19730525-19730622, 28 days Skylab 3: 19730728-19730925, 59.5 days Skylab 4: 19731116-19740208, 84 days								

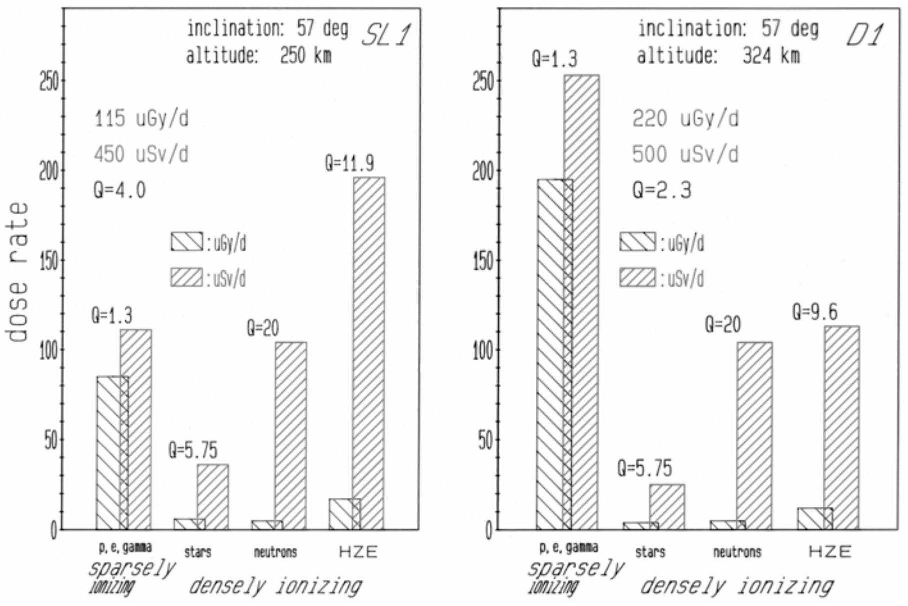


Figure 11.34. Contributions of space radiation components of different radiation quality to total doses, measured in the Spacelab missions SL1 and D1.

30 October–6 November 1985). Each panel shows, for each of the four categories described above, the absorbed dose rate either in μGy (left column) or μSv (right column) per day. The sum of the absorbed and the equivalent dose rate from all categories is also shown in each panel.

Having the same orbital inclinations, the major difference between the two missions is their altitude, which for LEO missions is the most important parameter for radiation exposure. Correspondingly, the absorbed dose rate from trapped protons rises by more than a factor of 2 for an increase in altitude by less than a factor of 1.3. Due to this disproportionate increase of the sparsely ionizing contribution, the average quality factor for the total exposure nearly halves from $\langle Q \rangle = 4.0$ to $\langle Q \rangle = 2.3$, so that despite the near doubling of the absorbed dose the equivalent dose rises by just 11%. A similar analysis of such passive dosimetric systems, exposed 17–25 March 1992 onboard MIR (400 km altitude, 51.5° inclination, see also [R106]), yielded a further reduction of the average quality factor down to $\langle Q \rangle = 2.1$. In addition to the once more increased altitude, a heliocentric potential 165 MV higher than that during the SL1 mission (see Figure 11.17) further subdued the relative contribution to absorbed dose from the GCR heavy ions as compared to that of the trapped protons.

The dose rates determined from personal dosimeters of US astronauts of the SL1 and D1 missions as shown in Figure 11.33, reflect the above-found moderate increase from the SL1 to the D1 missions by 11%. However, the absolute numbers in Figure 11.33 are systematically lower than those in Figure 11.34 – possibly due to the more sophisticated evaluation in the latter.

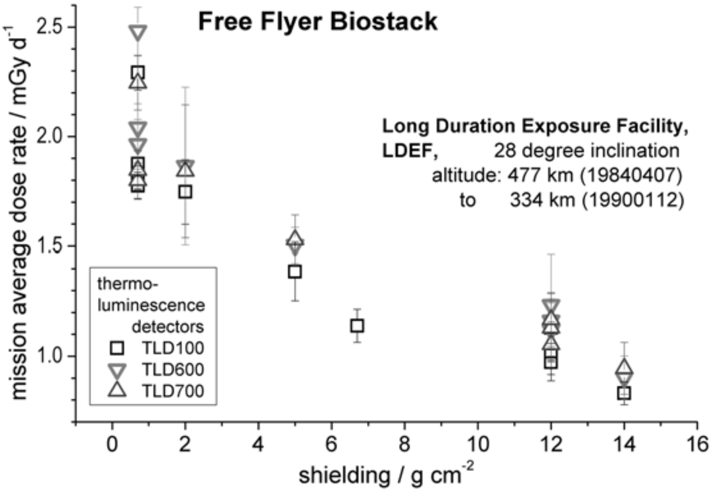


Figure 11.35. The influence of shielding on trapped proton and GCR ion dose rates, as measured in thermoluminescence detectors exposed during nearly 6 years of (mainly) solar minimum conditions between cycles 21 and 22.

Figure 11.35 displays the average dose rate measured with thermoluminescence detectors behind various thicknesses of cover layers exposed during slightly more than $5\frac{3}{4}$ years' exposure on the Long Duration Exposure Facility (LDEF). Due to its low inclination orbit of 28.5°, the particle energy spectrum to which LDEF was exposed can be separated into two regions. Below 1 GeV only trapped protons and above 1 GeV/nucleon only GCR ions irradiated the LDEF [R113, Figure 12]. Figure 11.35 reveals the preponderance of GCR ions behind thicker shield layers – say, $6\,\mathrm{g\,cm^{-2}}$ – in that a further doubling of the thickness reduces dose rates by just 20%. This so far unique mission covered, for the largest part, the phase of minimum solar activity between solar cycles 21 and 22, and thus yielded a solid experimental estimate of maximum radiation exposures to be expected in such orbits. Similar long-term results for the exterior radiation environment of the ISS are expected from the MATROSHKA facility – a human phantom profusely equipped with the whole array of active and passive radiation detectors presently available for space applications. Being installed outside the ISS, among its objectives is the measurement of dose profiles and depth dose distributions generated by SPE ions during such events, as it is required for risk assessment during extravehicular activities [R114].

The first time that a US space crew became exposed to a large energetic SPE occurred during the last 21.6 hours of the STS-28 mission of 8–13 August 1989. At that time the event was the second largest of the spaceflight era after the SPE of 4 August 1972. The 57° orbital inclination mission at an (unofficial) altitude of 300 km took place near the maximum of solar cycle 22 at the beginning of a phase of so far unparalleled SPE activity in September–October 1989. A tissue equivalent proportional counter (TEPC) dosimeter was monitoring radiation exposure during the

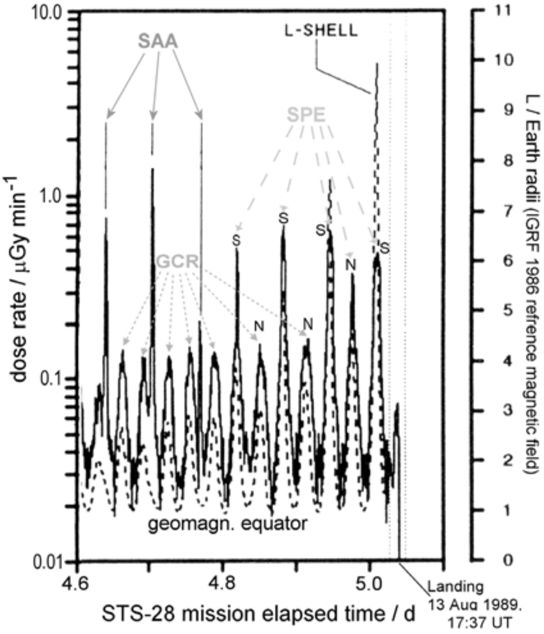


Figure 11.36. The first SPE dose rate profile measured during a US manned space mission. SAA points to passages through the South Atlantic Anomaly, GCR points to peaks of GCR dose rate near the poles, and SPE points to dose rate enhancement near the poles (S and N) by solar protons.

mission. The dose rates measured by this device during the approximately 11 final hours of the mission are shown in Figure 11.36 [R115]. Only 18 hours after the initial flare on 12 August 13:58 UT, did the TEPC record the first significant increase in dose rate when approaching the south polar aurora zone at mission elapsed time 4.816 days. No increase in dose rate was observed during the next two passages through the northern polar region, where the maximum L-value belonging to these orbits stayed below about 4 Earth radii. Only during the last passage near the northern polar zone this L-shell was crossed, and a corresponding increase in dose rate could be observed. In total, the absorbed dose that could be ascribed to this SPE was measured as 16 µGy as compared to 96 µGy accumulated during SAA passages, and 515 µGy accumulated from GCR exposure. Based on measurements made by other satellites in the initial phase of the SPE, estimates of the expected dose enhancement were calculated using a storm-disturbed geomagnetic transmission coefficient as, for example, shown in Figure 11.26. The results indicated that flight rule exposure limits would not be violated. However, overall the projected SPE dose turned out to be a factor of 2,000 larger than the measured TEPC dose. Beyond that comparison – which highlights the difficulties in forecasting SPE associated radiation exposures in LEO – the details of the dose rate profile nicely summarize the spatial and temporal features typical of such LEO exposures and space vehicles. During quiet solar times, about every other 8 orbits, the dose rates reach peak values created by trapped protons during two to three consecutive passages through the SAA, which are roughly an order of magnitude larger than the peak values engendered by GCR near the polar regions. Dose rates near the geomagnetic equator for such a mission are lower than the GCR peak dose rates by a factor of about 7. Enhancement by SPE ions is restricted to orbit segments near the poles, where the asymmetry of the geomagnetic field, as displayed in Figure 11.22, induces an additional longitude dependence as it is also shown in Figure 11.36 (two northern passages without any enhancement). So, in LEO only during a (usually small) fraction of the total event duration can SPE ions contribute to the mission dose.

Soon after this SPE, a phase of vehement solar activity set in and spawned most of the events which are now considered prototypical for a worst-case SPE (see Figure 11.11). The one with the hardest spectrum and the second largest associated GLE ever measured was the event of 29 September 1989. Luckily, during this event another active dosimetric device was monitoring this event in a LEO orbit, in addition to the usual solar particle flux surveillance by one of the dedicated Geostationary Operational Environmental Satellites – the GOES-7 satellite. On the (then Soviet) space station MIR, a silicon detector-based dosimeter, the Liulin device, was monitoring the dose rate inside the space station at an orbit of 57° inclination and an altitude of then about 400 km [R116]. About 12 hours into the event, the peak dose rate for orbit segments with $4.5 \leq L \leq 7.0$ – near the poles – had risen by a factor of nearly 500 above the pre-event level (Figure 11.37). The flux of protons with energies between 39 and 82 MeV, as measured by the GOES-7 satellite, rose by nearly an order of magnitude more; but for the dose rate inside MIR the higher energies are more important, and those fluxes rose much less steeply. The daily averaged dose rate rose from 0.31 mGy d^{-1} to more than a ten-fold 3.2 mGy d^{-1} during the event peak. The total dose, as recorded by the Liulin device in the period shown, was

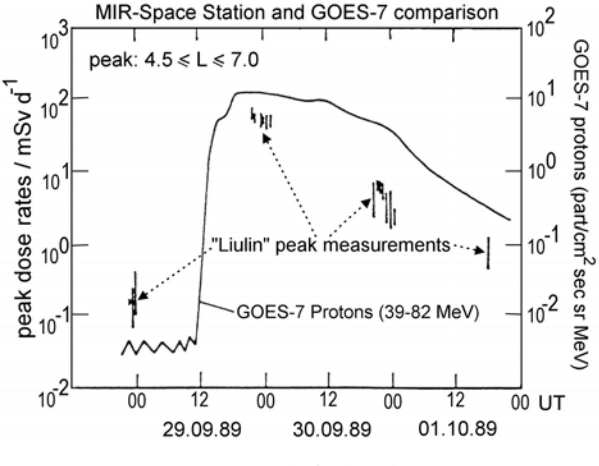


Figure 11.37. Dose rate measurements on Mir during the extreme SPE on 29 September 1989, compared with energetic proton flux measured on the GOES-7 satellite.

determined as 4 mGy, about 75% of which was ascribed to SPE protons and the rest to heavier ions which were also present in this event. Even if multiplied by an RBE of 2, this event dose would not pose any health risk to the crew. Figures 11.36 and 11.37 provide a graphic demonstration that this is owed to the shielding capacity of the geomagnetic field.

11.6.3 Exposures during interplanetary missions

With the exception of the Apollo missions, data for radiation exposures outside the terrestrial magnetosphere are mainly theoretical predictions that rely on an impressive amount of (by now) available physical data on particulate and energy spectra and on their spatial and temporal variation. One radiation component – which in LEO missions is the source most susceptible to stochastic space weather-related variability, and hence an obstacle to anticipatory assessments of radiation risks – the trapped radiation belts, are absent in interplanetary space and on celestial bodies without a sufficiently strong magnetosphere. However, this advantage is outweighed by the loss of geomagnetic shielding that the geomagnetic field provides against energetic SPE protons. So far, even the most menacing events of the space era, such as those listed in Figure 11.10 or Figure 11.11, have not led to radiation exposures with potentially significant risks for deterministic health effects.

To a large extent this is owed to the geomagnetic shielding, the effect of which is shown in Figure 11.24 and Figure 11.25. Given the extreme variability of SPE energy spectra as demonstrated in Figure 11.9 and in Figure 11.10, even for extreme events, a common approach to assess the radiation risk from SPE radiation has been to select either an actual event as a worst-case event against which adequate protection is to be designed, or to construe an artificial event which somehow is worse than all observed events. As mentioned above, previously the SPE of 4 August 1972 associated with the GLE 24, or the SPE of 23 February 1956 associated with GLE 5, or the SPE of 29 September 1989 associated with GLE 42, and also the SPE of 19 October 1989 associated with GLE 43, have been adopted as such worst-case events. Presently, the SPE associated with GLE 5 is considered the most appropriate estimate for a worst-case event. Its unknown energy spectrum is substituted by ten times the energy spectrum for GLE 42. Table 11.4 shows equivalent tissue doses calculated for this fictitious worst-case SPE in the skin, the lens of the eye, and in the bone marrow that would have been engendered in these critical 'benchmark' organs behind typical aluminium shielding thicknesses in interplanetary space and on the surfaces of the Moon and Mars. Doses on the lunar surface were in this case just estimated by dividing the free space value by 2, thereby allowing for the reduction of the solid angle of accessibility by the Moon itself. This approximation neglects the not marginal contribution of albedo neutrons coming from the soil, especially for an SPE with such a hard spectrum. On the other hand, in these calculations the potential presence of heavier ions was also allowed for, and their contribution to the equivalent tissue dose was calculated.

As far as Mars is concerned, it is evident that its $\sim 16\,\mathrm{g\,cm^{-2}}$ CO_2 atmosphere provides sufficient protection so that the induction of early symptoms of radiation

Table 11.4. Worst[§] case SPE radiation exposures in Sv during different mission phases for critical tissues under different mass shielding, given in parentheses in equivalent g/cm^2 aluminium; lens = ocular lens; BFO = blood-forming organs.

Mission phase	Concluding charges	Space suit (0.3)			Pressure vessel (1)			Equipment room (5)			Radiation shelter (10)		
		Skin	Lens	BFO	Skin	Lens	BFO	Skin	Lens	BFO	Skin	Lens	BFO
Free space	$Z = 1$	173.80	66.80	3.78	55.40	31.80	3.14	5.77	5.01	1.68	2.26	2.13	1.06
	$Z = 2$	114.90	13.30	0.40	8.20	3.30	0.35	0.66	0.49	0.24	0.35	0.29	0.19
	$3 \leq Z \leq 10$	5.30	0.90	0.01	0.50	0.20	0.01	0.02	0.02	<0.01	<0.01	<0.01	<0.01
	$11 \leq Z \leq 20$	0.90	0.20	0.01	0.20	0.10	0.01	0.01	0.01	<0.01	<0.01	<0.01	<0.01
	$21 \leq Z \leq 28$	0.20	0.10	0.01	0.10	0.10	0.01	0.02	0.01	<0.01	<0.01	<0.01	<0.01
	Total	**295.10**	**81.30**	**4.21**	**64.40**	**35.50**	**3.52**	**6.48**	**5.54**	**1.93**	**2.62**	**2.43**	**1.26**
Lunar surface	$Z = 1$	86.90	33.40	1.89	27.70	15.90	1.57	2.89	2.51	0.84	1.13	1.07	0.53
	$Z = 2$	57.45	6.65	0.20	4.10	1.65	0.18	0.33	0.25	0.12	0.18	0.15	0.10
	$3 \leq Z \leq 10$	2.65	0.45	<0.01	0.25	0.10	<0.01	0.01	0.01	<0.01	<0.01	<0.01	<0.01
	$11 \leq Z \leq 20$	0.45	0.10	<0.01	0.10	0.05	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	$21 \leq Z \leq 28$	0.10	0.05	<0.01	0.05	0.05	<0.01	0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	Total	**147.55**	**40.65**	**2.11**	**32.20**	**17.75**	**1.76**	**3.24**	**2.77**	**0.97**	**1.31**	**1.22**	**0.63**
Martian surface	**Total**	**0.45**	**0.44**	**0.32**	**0.44**	**0.42**	**0.31**	**0.38**	**0.37**	**0.28**	**0.33**	**0.32**	**0.25**

[§] 23 February 1956 event (GLE 5) as approximated by 10X flux of 29 September 1989 event (GLE 42).

sickness can be excluded under even the lightest 'shielding' of a spacesuit. During extravehicular activity in space or on the lunar surface, such exposure would have nearly immediate incapacitating consequences if no better shielding were to be accessible during the whole duration of that event (see Figures 11.29 and 11.30, colour section). Even behind $5\,\mathrm{g\,cm^{-2}}$ aluminium, the walls of an 'equipment room', the doses in interplanetary space reach levels where symptoms of early radiation sickness would be a probable consequence. Since $5\,\mathrm{g\,cm^{-2}}$ was estimated as representative for MIR [R116], such would have been the health detriment the cosmonauts would have experienced had the space station not been protected by the geomagnetic field. Early radiation sickness could be prevented in a 'storm shelter' behind $10\,\mathrm{g\,cm^{-2}}$ aluminium, but even there a significant increase in the probability for late cancer mortality (see Figure 11.31) would be the result of such an exposure. On the lunar surface the shielding by an equipment room would suffice to prevent early radiation sickness, although an increase in late cancer mortality would also ensue, even in the storm shelter. Some increase in late cancer mortality could be expected even from most of the exposures on the martian surface, although such an increased risk could be estimated only with large error margins.

Given the possibility that our Sun enters a phase of more violent activity [R65], the SPE associated with the 1859 Carrington flare might be a more realistic prototype for a worst-case SPE. The corresponding spectral approximations in Figure 11.11 were used to calculate [R20] – this time for only a proton component – the tissue doses for the critical organs behind the conventional mass shields as given in Table 11.5, with the exception of a spacesuit, since evidently a mission with a finite probability for such an exposure to occur will not lift off in the first place. The differences to the comparable predictions (for protons) in Table 11.4 are an indication of the uncertainty of such predictions. They arise mainly from the uncertainty of the true spectral shape at proton energies above, say, a few hundred MeV as they enter these calculations. The experimental data are very scarce and derogated by large uncertainties at these higher energies. The experimental uncertainty is augmented by the lack of a physical acceleration model that fits all events. Traditional parametrizations chose exponential functions in rigidity instead of energy which, however, reflected only some subset of events reasonably well. The recently advocated choice of Weibull functions of proton energy as well as of rigidity has been adopted for the calculations in Table 11.5. With the exception of the results for $1\,\mathrm{g\,cm^{-2}}$, these calculations agree with the assumption of the GLE 5 SPE being the worst-case prototype insofar as the spectrum of the 29 September 1989 event appears to be the most dangerous. Only for the lightest shielding and the organs with little self shielding (the skin and the eye lens) the spectrum of the 23 March 1991 event would deposit higher doses in these organs. As regards the choice of energy or rigidity, the last column in Table 11.5 shows that for the most energetic events – the SPEs of September–October 1989 – this choice is inconsequential, given the intrinsic uncertainties of these results.

Figure 11.38 (see colour section) compares the results in Table 11.5 with the NASA limits for deterministic acute effects in the critical organs. Apart from the earliest worst-case prototype – the August 1972 SPE – it is difficult to safely

Table 11.5. Tissue doses in Gy behind different aluminum mass shielding for the 1859 Carrington event, whose spectral shape is mimicked by energy/rigidity spectra of known worst-case events.

Al shield g cm^{-2}	Solar particle event	Skin	Eye	BFO	
				Energy spectrum	Rigidity spectrum
1	4 Aug 1972	34.26	23.83	1.41	1.96
	15 Aug 1989	43.62	24.71	1.29	0.78
	29 Sep 1989	35.39	23.37	**2.81**	**2.63**
	19 Oct 1989	39.67	24.04	2.12	2.20
	23 Mar 1991	**44.80**	**24.88**	1.09	0.35
2	4 Aug 1972	19.05	14.39	1.05	1.62
	15 Aug 1989	17.10	11.80	1.02	0.57
	29 Sep 1989	**18.01**	**14.00**	**2.44**	**2.23**
	19 Oct 1989	17.49	12.88	1.80	1.83
	23 Mar 1991	16.94	11.45	0.85	0.23
5	4 Aug 1972	5.56	4.61	0.47	0.99
	15 Aug 1989	4.14	3.50	0.59	0.27
	29 Sep 1989	**6.65**	**6.02**	**1.71**	**1.48**
	19 Oct 1989	5.46	4.83	1.22	1.16
	23 Mar 1991	3.78	3.13	0.46	0.08
10	4 Aug 1972	1.23	1.08	0.15	0.53
	15 Aug 1989	1.19	1.10	0.30	0.10
	29 Sep 1989	**2.82**	**2.73**	**1.09**	**0.86**
	19 Oct 1989	2.08	2.00	0.75	0.64
	23 Mar 1991	0.98	0.90	0.22	0.02

Stephens Jr., D. L., Townsend, L. W., Hoff, J. L., Interplanetary crew dose estimates for worst-case solar particle events based on historical data for the Carrington flare of 1859, *Acta Astronautica*, **56** 969–974, 2005.

avoid the violation of these limits, even behind the shielding of the 'storm shelter' of 10 g cm^{-2} thickness, and in particular if an RBE distinctly greater than 1 would be appropriate.

Chronic exposure to GCR heavy ions in interplanetary space and on planetary surfaces is never associated with a risk for early deterministic radiation sickness. On the other hand, due to the loss of geomagnetic shielding the risk of late cancer mortality is more difficult to contain. Essentially, the geomagnetic shielding effect has to be replaced by additional mass shielding. Presently, the establishment of a permanent lunar base and setting foot on the surface of Mars are at the forefront of missions planned for the next few decades.

For a set of most probable scenarios for such missions, Table 11.6 summarizes the radiation doses projected to be accrued during these missions from the journey through interplanetary space and from the sojourn upon the surfaces. Whereas GCR

Table 11.6. Mission doses from galactic cosmic rays for Reference Missions to the Moon and Mars.

Spacecraft shield thickness $\mathrm{g\,cm^{-2}}$	Solar activity	Shield material	BFO-equivalent dose rates			BFO-mission equivalent dose/mSv								
			in space $\mathrm{mSv\,a^{-1}}$	on Moon $\mathrm{mSv\,a^{-1}}$	on Mars[$] $\mathrm{mSv\,a^{-1}}$	Moon, 190 days			Mars, 450 days			Mars, 47 days		
						10 d trip	180 d stay	190 d total	420 d trip	30 d stay	450 d total	422 d trip	525 d stay	947 d total
1		Alu	711.7	355.9		19.5	175	**195**	818.4		**828**	822.3		**993**
(pressure	1977 min	PE	694.7	347.4	119	19.0	171	**190**	798.8	9.8	**809**	802.6	171	**974**
vessel)		Alu	271.7	135.9		7.4	67.0	**74.4**	312.4		**317**	313.9		**402**
	1970 max	PE	265.2	132.6	61	7.2	65.3	**72.6**	305.0	5.0	**310**	306.4	87.7	**394**
5		Alu	646.9	323.5		17.7	159.4	**177**	743.9		**754**	747.4		**918**
(equipment	1977 min	PE	584.3	292.2	119	16.0	144	**160**	671.9	9.8	**682**	675.1	171	**846**
room)		Alu	255.6	127.8		7.0	63.0	**70.0**	293.9		**299**	295.3		**383**
	1970 max	PE	229.2	114.6	61	6.3	56.6	**62.8**	263.6	5.0	**269**	264.8	87.7	**353**
10		Alu	589.0	294.5		16.1	145.1	**161**	677.3		**687**	680.5		**852**
(shelter)	1977 min	PE	499.0	249.5	119	13.7	123.0	**137**	573.8	9.8	**584**	576.5	171	**748**
		Alu	239.5	119.8		6.6	59.0	**65.6**	275.4		**280**	276.6		**364**
	1970 max	PE	198.7	99.4	61	5.4	49.0	**54.4**	228.4	5.0	**233**	229.6	87.7	**317**
20		Alu	517.6	258.8		14.2	127.5	**142**	595.2		**605**	598.0		**769**
	1977 min	PE	414.0	207.0	119	11.3	102.0	**113**	476.1	9.8	**486**	478.3	171	**649**
		Alu	217.7	108.9		6.0	53.7	**59.6**	250.3		**255**	251.5		**339**
	1970 max	PE	166.3	83.2	61	4.6	41.0	**45.5**	191.2	5.0	**196**	192.1	87.7	**280**

[$] minimal shielding only

doses accumulated during lunar missions clearly allow for several missions until the career limits become infringed (equation (11.2)), for manned missions to Mars this is reasonably achieved only for missions during maximum solar activity, where the magnetic field of the solar wind somewhat compensates the loss of the geomagnetic shielding. Once more it happens that the Sun dominates human affairs, even in space. Whereas an increase of mass shielding by a factor of 20 from 1–20 $\mathrm{g\,cm^{-2}}$ aluminium reduces the GCR dose for a 947-day mission to Mars during solar minimum activity by a factor of 2.93, nearly the same saving in dose, a factor of 2.5, can be achieved by simply changing to a mission during maximum solar activity, with an additional change from again only 1 $\mathrm{g\,cm^{-2}}$ aluminium to 1 $\mathrm{g\,cm^{-2}}$ polyethylene (PE). Of course, such an alternative is only a theoretical option, given the enhanced probability of SPE exposures during maximum solar activity. However, it emphasises the dominant role of the Sun. The change from aluminium to PE reduces the GCR dose, since in PE the dose contribution from secondary spallation and fragmentation products is reduced due to the much smaller cross-section of the hydrogen nucleus. The optimum material for GCR shielding would be pure hydrogen, while hydrogen-rich PE is a technically feasible approximation. Figure 11.39 (see colour section) compares the alternatives presented in Table 11.6, with gender and age specific career exposure limits from equation (11.2). It illustrates that for all scenarios and shields a mission during maximum solar activity would be the optimal choice for minimizing GCR exposures – neglecting, of course, the increasing risk of exposure to large SPEs. Secondly, it demonstrates, once more, the inefficiency of mass shielding in reducing GCR exposure. During maximum solar activity this inefficiency is also maximal, since in this phase the lower energies of the GCR ions are already shielded by the solar wind. Also, it visualizes the higher odds against female astronauts, who are prone to larger cancer risks due to added cancers of the sexual organs. Bearing in mind that the SPE doses, in addition to posing the threat of early radiation sickness, also add to the late cancer risk, a mission during minimum solar activity would probably be preferable. However, an optimum choice must also take into account that the risk of late cancer mortality may materialize only many years after the mission, whereas early radiation sickness might jeopardize a safe mission completion. Then, the loss of life-span potentially associated with extreme SPE exposures during the mission would be much larger than that from a late fatal cancer. Approaches to a rational choice for an appropriate mission selection criterion for long-term human missions in interplanetary space have been discussed in [R117], where the minimization of the Healthy Lifespan Lost due to a mission appears much superior to a selection according to assumptions regarding worst-case events, of which the Sun is the ultimate cause.

11.6.4 Observed health effects

For conventional epidemiological studies, the astronaut population is much too small to reveal adverse health effects (apart from fatal accidents, of course) from being an astronaut and the associated radiation exposures – in particular since for the vast majority these exposures have been kept well below levels which for terrestrial radiation workers are deemed risk-free. Nevertheless, for the American astronaut corps a close epidemiological surveillance has been established [R31, R32]. So far, no increased cancer frequency or other health effects could be observed in this population [R33, R34]. However, one health effect has been observed among astronauts which with some credibility can be ascribed to their exposure to space radiation. The induction of opacities in the lens of the eye (cataract) has long been recognized as a relatively sensitive indicator of radiation exposure [R35]. For this reason, some studies have investigated this [R28, R29], although in comparison to cancer it is a condition which nowadays can be easily corrected. In [R28], indeed, a statistically significant increase in the prevalence of cataracts was observed among those astronauts who accumulated higher doses (on average 45 mSv), than a less exposed group (on average 3.6 mSV). However, more important than the dose value of 8 mSv, which had been selected as the threshold separating these two groups, was the quality of the radiation exposure. The higher exposure was mainly incurred during missions with high orbital inclination, where a significantly larger fraction of the dose was produced by GCR heavy ions. Therefore, if confirmed in further studies, this finding is a notable reminder of the potentially unique and probably only insufficiently understood radiobiological mechanisms by which this component of space radiation interacts with biological systems.

For astronauts also, a radiobiological response to space radiation had been studied, which clearly allowed for an association with their respective exposure [R30]. In circulating lymphocytes of astronauts, an increased frequency of cells with aberrant caryotypes has been detected – cells with one or more chromosomes of irregular morphology. At the levels detected, such an increase cannot be associated with pathological consequences. Rather than foreboding disease, at such levels they can be exploited as a biologically weighting dosimetry system. Several studies are underway to establish the necessary techniques for long-term exploratory missions.

11.7 IMPLICATIONS FOR AIR CREWS

Atmospheric ionizing radiation (AIR) at air traffic cruising altitudes is made up of the fragments of cosmic radiation remaining after penetration through the residual atmosphere at these altitudes, and, apart from the cosmic radiation itself, comprises an unrivalled complex mixture of different radiation qualities [R123].

11.7.1 Exposures

The composition of AIR varies strongly with the thickness of the air layer which the cosmic radiation has penetrated at a given altitude. Figure 11.40 expresses this composition of the radiation field in terms of the contribution of the various components to the total effective dose rate as a function of altitude. In aviation, altitudes are conventionally expressed in units of 100 feet, abbreviated as flight level (FL). At

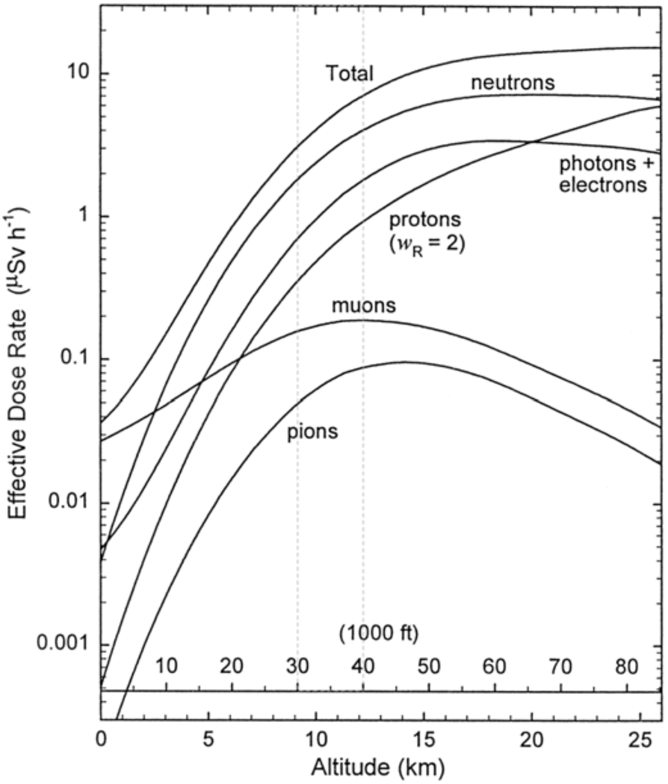


Figure 11.40. Dependence of composition and corresponding effective dose rate of AIR on altitude (LUIN-98F calculations, heliocentric potential for June 1997, near polar plateau with cut-off 0.8 GV).

commercial jet cruising altitudes between 30,000 and 40,000 feet (FL 300–400), the GCR protons as the primary cause of AIR contribute only about 10% to the effective dose, even if their average radiation weighting factor w_R is taken as high as 2. More than 50% of the effective dose is generated by energetic neutrons. Their energy distributions at two different altitudes, as shown in Figure 11.41, reveal that the shape of the spectrum near the two high-energy peaks at around 20 and 150 MeV is conspicuously insensitive to altitude changes – a feature which persists down to sea level. In Figure 11.41 the main spectral shape of the trapped proton belts is also foreshadowed, since it is the decay of these neutrons which constantly replenishes the inner trapped proton population. Next to densely ionizing energetic neutrons with a high quality factor, secondary electrons and photons, as sparsely ionizing radiation with a quality factor of 1, essentially contribute the remaining part to the effective dose from AIR. Below about 2.5 km altitude, sparsely ionizing muons start to represent the dominating contributor to AIR exposure. Above 25 km altitude the primary GCR protons become the dominant source, but the ambient dose rate once

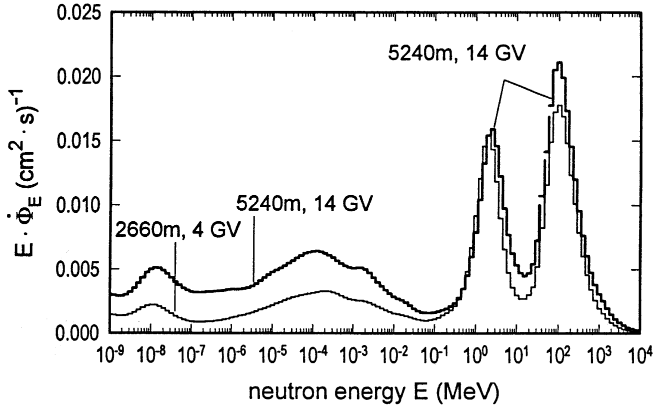


Figure 11.41. Energy distributions of atmospheric neutrons at different altitudes.

more starts to decline. The corresponding maximum bears the name of its discoverer, Pfotzer [R98, R99].

As an offspring of space radiation, AIR is modulated by the same factors as is the primary radiation in space.

The dependence of AIR exposure on (geomagnetic) latitude arises from the geomagnetic shielding that operates in LEO. Figure 11.42 presents the results from a multitude of experimental measurement campaigns, which in recent years mapped the altitude and latitude dependence of AIR in the northern hemisphere. Similar to the plateau region of GCR exposure in LEO, here a plateau of effective dose rate is shaped at latitudes above, say, 60°. Note the asymmetry in Figure 11.42 where the minimum dose rate does not occur at the geographical equator but rather near 15° N. This again reflects the asymmetry in the geomagnetic field, as displayed in Figure 11.22.

As in space, the long-term temporal variation in AIR dose rates reflects the cycles of solar activity. Figure 11.43 (see colour section) shows how this modulation of the primary GCR proton flux transforms into dose rate profiles. For two ranges of cruising altitude, the modulation by solar activity is shown. At higher altitudes at and above FL 500, which is used by corporate jets, the amplitude of this modulation is much more strongly developed than at commercial jet altitudes between FL 300 and 400. At latitudes below the polar plateau region where the curves in Figure 11.43 pertain, the amplitude variation becomes less and less.

Figure 11.44 (see colour section) comprises all regular dependences of AIR dose rates. For three years in solar cycle 23, the iso-dose rate lines clearly reveal the polar plateau region. Near the previous solar minimum phase in 1996, the maximum dose rate in the polar plateau region for commercial jet altitudes of about $10\,\mu\mathrm{Sv\,h^{-1}}$ drops to $9\,\mu\mathrm{Sv\,h^{-1}}$ in the middle of the cycle. The dose rate during the previous maximum of solar activity near 2001 attains a minimum value of about $7.3\,\mu\mathrm{Sv\,h^{-1}}$. At FL 300 this variation of the dose rate with the solar cycle is even less pronounced. For cruising altitudes of corporate jets, on the other hand, quite significant changes

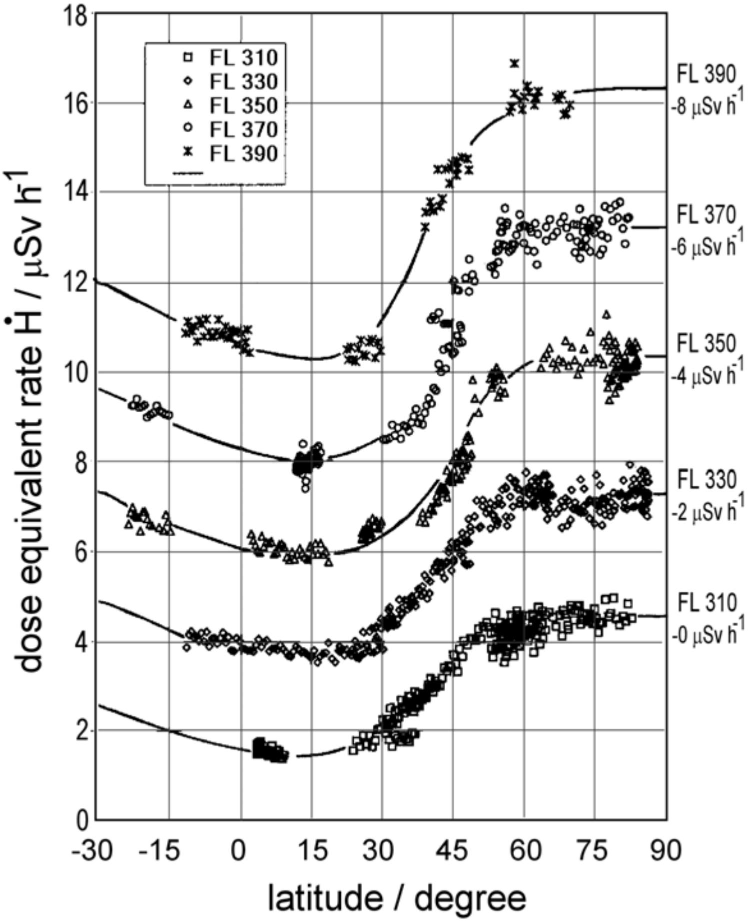


Figure 11.42. Comparison of measured and calculated effective dose rates covering commercial jet altitudes in the northern hemisphere.

in dose rate emerge. The maximum dose rate during minimum solar activity of 10 μSv h^{-1} is a rather stable value, since the minima of heliocentric potentials appear to vary rather narrowly (see Figure 11.17). In contrast, the maxima of solar activity show much larger variation, and thus the minima of dose rates attained during maximum solar activity will scatter more widely. Since the previous solar maximum was rather moderate, most dose-rate minima are likely to be smaller than the previous 7.3 μSv h^{-1}.

For air-crew exposure the impact of energetic SPEs is much smaller than in space. The reason is that at commercial jet altitudes nearly 200 g cm^{-2}, and at corporate jet altitudes 100 g/cm^{-2} of air layer, provide a shielding which in space is unattainable. For commercial jet altitudes, this translates into about 450 MeV that a solar proton must have in order to add to the AIR dose. Figure 11.10 shows that

Table 11.7. Estimated total event doses/mSv at typical cruising altitudes in the polar plateau region (60° N) for large solar particle events of solar cycle 22 and for a worst-case estimate represented by GLE 5.

Ground-level enhancement	Solar particle event	Flight altitude/FL (= 100 feet)		
		300	400	500
GLE 5#	23 February 1956	0.33	1.2	2.2
GLE 42	29 September 1989	0.033	0.12	0.22
GLE 43	19 October 1989	0.0054	0.019	0.039
GLE 44	22 October 1989	0.00062	0.0021	0.0047
GLE 45	24 October 1989	0.0012	0.038	0.080
GLE 42–45	29 Sep.–25 Oct. 1989	0.040	0.17	0.34

Estimated as 10 times the dose for GLE 42, which otherwise had a very similar event profile.

even among the large SPEs which have been used as worst-case events, only a few have significant proton fluxes above a few hundred MeV. Solar particle events which do not produce significant GLE events in neutron monitors can be ignored as far as the protection of air crews at commercial jet altitudes is concerned. Table 11.7 presents the results of transport calculations by the LUIN code [R101, W11], for the series of events in September–October 1989, which produced significant GLE events, 42–45. The event doses accrued by the worst of the documented events – the GLE 42 event – correspond to the doses incurred by one or two additional transatlantic flights at the corresponding altitude. Only an event with the intensities and energies of the GLE 5 event, as mimicked by a ten-fold boosted GLE 42 event, has the potential to violate the exposure limit of 1 mSv for pregnant crew. Of course, it must be realized that these calculations depend most critically on the proton energy spectrum adopted for these SPEs. Since the decisive energies in the GeV range have a rather poor empirical basis, these results must be viewed with some reservation. On the other hand, once a large SPE has passed Earth's orbit, it unfolds a protective effect, similar to that produced by the solar wind, in shielding off larger fractions of the GCR ions. The resulting Forbush decrease in GCR flux often lasts for some weeks, during which the 'quiet time' AIR dose rates are decreased, so that the net effect of an intense SPE may well be a reduction of the total annual exposure.

The limits pertaining to air crew since 2003 are the same as those for terrestrial radiation workers. Given the admissible working hours, it is inconceivable that these limits can be infringed by exposure to AIR. Even the lesser 'intervention' limit of 6 mSv is in practice only rarely reached. Under some regulations a medical check-up is required if occupational exposure is to be continued after that limit has been reached.

This raises the question of adverse health effects that might arise from exposures of air crew to AIR.

11.7.2 Observed health effects

The relevant health effect from exposures typical of radiation workers, including flight crews, is again induction of late malignant neoplasm. Presently, no experimentally validated model (let alone theory) covering the primary physical and the secondary molecular, cellular, tissue and immunological processes, leading from the initial energy absorption event to the expression of the final tumour, is capable of predicting the human cancer risk from exposures to low levels of ionizing radiation. Thus, empirical data from human exposures are the only reliable source for gauging the health effects of such low-level occupational exposures.

An early result from a corresponding epidemiological study was published in 1990 [R122]. This and subsequent studies in the 1990s never revealed a consistent increase of cancer mortality in the populations studied. When, in one study, a given cancer appeared to arise significantly more frequently, then in another study incidence or mortality from a different cancer appeared to be augmented. Since several or rather many cancer sites were usually included in the analysis, these 'significant' findings were most probably artefacts that arose from the problem of 'multiple testing'. A subsequent meta-study [R119] of the earlier ones corroborated this interpretation. Only in three cases did the cancer frequencies appear to be enhanced significantly at the 95% confidence level. For pilots these were incidents of prostate cancer and mortality from melanoma. For female cabin crew, incidence of breast cancer appeared to be significantly enhanced. Whereas melanoma rarely, if ever, had been associated with ionizing radiation, the more frequent exposure of pilots to solar UV appears to be a better explanation – you rarely see a pale pilot. Prostate cancer – the 'bus driver syndrome' – is a typical risk for sedentary occupations which again is typical in pilots. Enhanced breast cancer risk is strongly associated with pauciparity, and reduced or even suppressed child-bearing is again typical for female air crew. A large subsequent study comprising many European airlines further corroborated these findings [R120, R121], including the finding that all cancer mortality was always significantly and strongly reduced. So, in agreement with the results of virtually all epidemiological studies of radiation workers or of populations residing in areas of largely enhanced natural background radiation, exposures of air crew to AIR as an offspring of space radiation does not appear to entail a measurable risk to their health.

11.8 SPACE WEATHER IMPACTS ON THE BIOSPHERE

Influences of space weather conditions on terrestrial processes abound. Their impact on technical systems such as communication systems and large-scale systems of electric conductors such as powerlines or pipelines, have long been known (e.g., Chapters 9 and 10). However, observations of interactions with the terrestrial biosphere have multiplied in recent years. Although the focus of this chapter is the radiation health effects of the ionizing component of space radiation, we believe a condensed synopsis of some of the (evident as well as still conjectural) biological interactions of the space radiation environment and our ecosphere is in order.

Wise pigeon-racers monitor solar activity, and do not fly their homing-pigeons if geomagnetic disturbances can be expected. Probably less well known is that magnetoreception is also a widespread phenomenon among sea dwellers from lobsters [R67] to sea turtles [R90], to salmon and trout [R36, R37]. It has recently been established, from correlations of (fatal) whale strandings with solar activity indices [R38], that disturbances brought about by (in mechanical terms) weak forces exerted by some 100 nanotesla can virtually kill giant animals like whales. Even living organisms at the opposite size scale of living systems – bacteria – rely on a magnetotactic sense for navigation in an environment where navigation by other signals is not possible [R39]. Whether loss of orientation spells doom for such organisms depends, of course, on their capability to sit out a loss of feeding grounds for the duration of geomagnetic storms, or whether the distances they travel during the time of disturbed navigation can lead them far enough astray, such as in the case of stranded whales.

When navigation by man's primary tool of navigation – the eye – has to operate near the sensory threshold, even in man a modulation of this threshold by weak magnetic fields has been established which makes man, too, sensitive to changes in the direction of the Earth's magnetic field [R40, R41, R42, R43] under such conditions.

A doubling of the rate of heart attacks between 30 September and 6 October, in Tbilisi, Georgia, after the second largest SPE of the space era on 29 September 1989, initiated a series of studies in correlations of human performance – as expressed, for example, in the rate of car accidents – with solar and geomagnetic activity, since a direct radiation effect at a latitude of $\approx 42^\circ$ N and an altitude of $\approx 1{,}000$ m can be ruled out ([R44, R45] and references in [R48]). So far, no compelling case for an influence of space weather on 'meteorosensitive' persons could be produced from these studies. Even less convincing are the attempts to link the incidence rate of physical and psychic trauma to the 27-day solar rotation period [R46]. In contrast to such searches for statistical associations, others have investigated, in controlled experimental set-ups, the variation of physiological parameters such as (among others) blood pressure and of psycho-physiological indisposition with geomagnetic activity as expressed by the Ap index [R47]. For the end-points mentioned, they reported a definite dependence on Ap in the range between 1 and 5. So, in principle man, too, shares the sensitivity to magnetic fields, and might be influenced by their disturbances as engendered by space weather-related geomagnetic storms.

Whereas in the foregoing effects magnetoreceptive sensors evidentially can be identified as part of the reaction channel by which they are being brought about, in other cases of established space weather correlations with biological phenomena, such mechanisms still have to be elucidated. An example is the observed correlation of cyclic population outbreaks of forest moths with the sunspot number [R66]. Whether host-pest relations as modulated by UV irradiance, or climatic factors, or combinations of both, or still other mechanisms are responsible for this correlation, has yet to be determined.

Far beyond the importance of interactions with specific sensory systems of individual biological species is the influence that the Sun exerts on the terrestrial climate and thereby on the biosphere as a whole. The variation during the solar cycle of the optical energy input with the solar activity has long been studied, and particularly comprehensive data are available from continuous satellite surveillance [R70]. The consensus is that the observed amplitude of a 0.10–0.15% variation with the solar cycle is unable to exert climatic forcing. (The potential influence of the larger variability of the ultraviolet part of the solar irradiance is discussed in Chapter 8.) On the other hand, incontrovertible evidence of correlations of climatic indicators with solar activity [R71, R72, R80] raises the question of alternative mechanisms. As a mechanism possibly responsible for the observed associations, the modulation of cloud formation by the concentrations of ions as condensation nuclei has been proposed [R73, R81]. These ions are almost entirely created by GCR, and their flux modulation by the solar wind (reflected in Figures 11.15, 11.16 (see colour section), 11.18 (see colour section), and 11.43 (see colour section)) could well imprint the solar cycle on such an important climate factor as cloud cover. Presently, the physico-chemical details of all relevant atmospheric processes are too poorly understood to allow quantitative assessment of the amount by which this mechanism contributes to climatic developments. The presently favoured CO_2 ‘greenhouse’ mechanism (an unfortunate misnomer) as the dominant climatic driver, may well have to be modified by such space weather-related components. When carbon dioxide’s bearing for climate was tested on geological time-scales [R74], the conclusion was that either ‘the reconstructed past CO_2 levels are (partially) incorrect’, or ‘the role of pCO_2 as the main driving force of past global (long-term) climate changes is questionable, at least during two of the four main cool climate modes of the Phanerozoic.’ When, on the same timescale the hypothesis of a climate ‘forcing’ by GCR was tested [R75], the results ‘seem to endorse that long-term climate variability is indeed a measure of cosmic ray flux variability.’ Taken together, and despite several loose ends in a detailed mechanistic understanding, these findings imply that the Sun may, after all, dominate in driving climatic change – either directly through the well established optical irradiance, or indirectly by its space weather activities which by GCR can modulate cloud covers and thereby the absorption and reflection of the former.

Also on a large timescale, the possibility that our Sun might produce much more extreme SPEs than observed so far, and the probable number of acute radiation deaths ensuing directly from such an event, has been studied [R53, R54]. Figure 11.45 (see colour section) [R55] compares the ‘expected’ annual frequency as a function of the number of ensuing casualties with mortality risks from everyday life. For death counts below 10^3–10^4 per event, the mortality risks from man-made sources clearly dominate those from solar flares. According to these calculations, which drew on data in [R53], only for death counts above 10^5 per event, such a natural disaster, originating from extreme space weather events, was estimated to occur more ‘frequently’ (once every million years or so) than man-made disasters – with the exception of wars, of course. As the author realized: ‘The original basis of these predictions is rather doubtful, and considerable uncertainty remains regarding the actual frequency of large solar flare events.' Figure 11.12 reminds us of our current estimate of this uncertainty.

Beyond that, catastrophic events in space – such as nearby supernovae or giant gamma-ray bursts in nearby magnetars – remain a rich field for speculations about drivers of abrupt Phanerozoic developments, usually of mass extinctions. However, though possibly being effects of space radiation, such events can no more be considered to be space weather effects.

11.9 SUMMARY

Direct space weather impacts on man only became a concern when man set out to leave the protective layer of Earth's atmosphere. The earliest encounters with space radiation, nearly a century ago – such as the balloon flights of Victor Hess – did not foreshadow the real extent to which man would become entangled with space radiation. Even before Jurij Alexejetwitsch Gagarin ultimately left behind Earth's atmospheric shield for 108 minutes, on 12 April 1961, and thereby became exposed to primary space radiation, the onset of the jet era, with the first scheduled flight of a Boeing 707 on 25 January 1959, and later with the Douglas DC8, marked the actual beginning of enhanced exposure of man to cosmic radiation – if only to the secondary decay products left after passage through the overlaying air layer. Shortly thereafter the first attempts were made to assess the related radiation exposure of air crew [R76, R77, R78]. At the same time the first theoretical assessments of this exposure were made for the cruising altitudes where supersonic flight would have to operate [R79]. The real impetus for experimental and theoretical work on space radiation came from the Apollo programme when man would (or might) encounter all components of space radiation behind the rather moderate shielding provided by the Apollo modules. The satellite measurements of atomic and energetic composition of trapped, galactic and solar radiation made, to a large extent during the Apollo era, furnished the database. Theoretical work on geomagnetic material shielding now provides the means to assess radiation exposure of man for any mission scenario from commercial air flight to missions to Mars or a permanent lunar base. By and large, radiation exposure in spaceflight and *a fortiori* in civil air flight, turned out to impose manageable risks, with the one exception of solar particle events. The still unpredictable behaviour of our Sun, leading to potentially extreme solar particle events, remains a serious obstacle to the planning of long-term missions outside the shielding provided by the terrestrial magnetosphere.

As regards radiation in space: so far the only health effect that with some confidence can be correlated with exposure to GCR heavy ions, is an increased incidence of opacities in the eye-lenses of astronauts. This possibly is an indicator of unique radiobiological properties of heavy ions of the GCR. Exposure of air crew to atmospheric ionizing radiation has not been found to cause negative health effects, despite the fact that among radiation workers, air crew are among the more heavily exposed persons.

For man on Earth, as for the biosphere as a whole, the indirect impact of space radiation and of space weather via climatic repercussions or by more subtle interactions via magnetic interactions, may turn out to be more influential. Finally, it may be worth keeping in mind what the summary of a special report on the Sun–climate relationship concluded [R72]: 'The issue of a Sun–climate relationship is gaining more and more respect these days. Our understanding of the Sun and solar processes has increased dramatically during recent years. Moreover, it is understood that the Sun affects the Earth's environment in a much more complicated way than we have imagined. And we realise that it is impossible to describe the effect of the Sun on Earth by just a few parameters.'

The last two sentences highlight the probability that the modulation of galactic cosmic ray fluxes is another lever by which our Sun – the ultimate driver of space weather – exerts its rule over our biosphere, as well as the certainty that other levers yet await their discovery.

11.10 CONVERSIONS

Magnetic rigidity

Magnetic rigidity is a particle's momentum per unit charge. It is the relevant quantity for characterizing a cosmic ray's ability to penetrate Earth's magnetic field. In terms of the ion's atomic mass number A (the number of neutrons and protons in the nucleus), its charge Q (in units of the proton's charge) and its kinetic energy E (in units of GeV/nucleon), the rigidity of an ion is given by:

$$R \text{ (in GV)} = (A/Q)\,(E^2 + 2M_0E)^{1/2}$$

where $M_0 = 0.9315016\,\text{GeV}/c^2$ is the atomic mass unit.

(see *https://creme96.nrl.navy.mil/cm/rigidity.htm*)

Relativistic momentum and energy

W: total energy = kinetic energy $T +$ 'rest-mass-energy', M; $M = m_0 * c^2$
$p =$ momentum; $P = p * c$
$P^2 + M^2 = W^2 = (T + M)^2 = T^2 + 2TM + M^2$
$T = M[(1 + (P/M)^2)^{1/2} - 1]$
M in GeV, P in GeV, T in GeV, and p in GeV/c

11.11 REFERENCES

R1 Bhattacharjee, P., Sigl, G., Origin and propagation of extremely high energy cosmic rays, E-print astro-ph/9811011, Physics Reports.
R2 Blümer, J. for the Pierre Auger Collaboration, Cosmic rays at the highest energies and the Pierre Auger Observatory. *J. Phys. G: Nucl. Part. Phys.*, **29**, 867–879, 2003.

R3 Hathaway, D. H., Wilson, R. M., Reichmann, E. J., Group sunspot numbers: Sunspot cycle characteristics, *Solar Physics*, **211**, 357–370, 2002.

R4 Friis-Christensen, E., Lassen, K., Length of the solar cycle: an indicator of solar activity closely associated with climate, *Science*, **254**, 698–700, 1991.

R5 Lassen, K., Friis-Christensen, E., Variability of the solar cycle length during the past five centuries and the apparent association with terrestrial climate, *Journal of Solar and Terrestrial Physics*, **57**, 8, 835–845.

R6 Anderson, K. A., Energetic particles in the Earth's magnetic field, *Annual Review of Nuclear Science*, **16**, 291–344, 1966.

R7 Lathullière, C., Menvielle, M., Lilensten, J., Amari, T., Radicella, S. M., From the Sun's atmosphere to the Earth's atmosphere: an overview of scientific models available for space weather developments. *Annales Geophysicae*, **20**, 1081–1104, 2002.

R8 International Commission on Radiological Protection, 1990 Recommendations of the International Commission on Radiological Protection, ICRP Publication 60, Pergamon Press, Oxford, 1991.

R9 International Commission on Radiological Protection, Recommendations of the International Commission on Radiological Protection, ICRP Publication 26, Pergamon Press, Oxford, 1977.

R10 McIlwain, C. E., Coordinates for mapping the distribution of magnetically trapped particles, *J. Geophys. Res.*, **66**, 11, 3681–3691.

R12 Usoskin, I. G., Alanko, K., Mursula, K., Kovaltsov, G. A., Heliopsheric modulation strength during the neutron monitor era, *Solar Physics*, **207**, 389–399, 2002.

R13 Horneck, G., Baumstark-Kahn, C., Facius, R., Radiation biology, in *Fundamentals of Space Biology*, Clément, G., Slenzka, K. (eds), Kluwer Academic Publishers, Dordrecht (in press, 2006).

R14 Braun, H., Christl, M., Rahmstorf, S., Ganopolski, A., Mangini, A., Kubatzki, C., Roth, K., Kromer, B., Possible solar origin of the 1470-year glacial climate cycle demonstrated in a coupled model, *Nature*, **438**, 208–211, 2005.

R15 Chen, J., Physics of coronal mass ejections: a new paradigm of solar eruptions, *Space Science Reviews*, **95**, 165–190, 2001.

R16 Wilson, J. W., Cucinotta, F. A., Tai, H., Simonsen, L. C., Shinn, J. L., Thibeault, S. A., Kim, M. Y., Galactic and Solar Cosmic Ray Shielding in Deep Space, NASA Technical Paper 3682, National Aeronautics and Space Administration, Washington DC, 1997.

R17 McCracken, K. G., Dreschhoff, G. A. M., Zeller, E. J., Smart, D. F., Shea, M. A., Solar cosmic ray events for the period 1561–1994: 1. Identification in polar ice, *Journal of Geophysical Research*, **106**, A10, 21585–21598, 2001.

R18 McCracken, K. G., Dreschhoff, G. A. M., Smart, D. F., Shea, M. A., Solar cosmic ray events for the period 1561–1994: 2. The Gleissberg periodicity, *Journal of Geophysical Research*, **106**, A10, 21599–21610, 2001.

R19 Carrington, R. C., Description of a singular appearance seen in the sun on September 1, 1859, *Monthl. Not. Roy. Astronom. Soc.*, **20**, 13–15, 1859.

R20 Stephens Jr., D. L., Townsend, L. W., Hoff, J. L., Interplanetary crew dose estimates for worst case solar particle events based on historical data for the Carrington flare of 1859, *Acta Astronautica*, **56**, 969–974, 2005.

R21 Moussas, X., Polygiannakis, J. M., Preka-Papadema, P., Exarhos, G., Solar cycles: A tutorial, *Advances in Space Research*, **35**, 5, 725–738, 2005.

R22 Usoskin, I. G., Mursula, K., Long-Term Solar Cycle Evolution: Review of Recent Developments, *Solar Physics*, **218**, 1–2, 319–343, 2003.

R23 International Commission on Radiological Protection, Relative Biological Effectiveness (RBE), Quality Factor (Q), and Radiation Weighting Factor (WR), ICRP Publication 92, Elsevier, 2004.

R24 Xapsos, M. A., Barth, J. L., Stassinopoulos, E. G., Messenger, S. R., Walters, R. J., Summers, G. P., Burke, E. A., Characterizing solar proton energy spectra for radiation effects applications, *IEEE Trans. Nucl. Sci.*, **47**, 6, 2218–2223, 2000.

R25 Wilson, J. W., Townsend, L. W., Nealy, J. E., Chun, S. Y., Hong, B. S., Buck, W. W., Lamkin, S. L., Ganapol, B. D., Khan, F., Cucinotta, F. A., BRYNTRN: A Baryon Transport Model, Technical Paper 2887, NASA, Washington, DC, 1989.

R26 Wilson, J. W., Badavi, F. F., Cucinotta, F. A., Shinn, J. L., Badhwar, G. D., Silberberg, R., Tsao, C. H., Townsend, L. W., Tripathi, R. K., HZETRN: Description of a free-space ion and nucleon transport and shielding computer program, Technical Paper 3495, NASA, Washington, DC, 1995.

R27 Billings, M. P., Yucker, W. R., The Computerized Anatomical Man (CAM) Model. Report CR-134043, NASA, Washington DC, 1973.

R28 Cucinotta, F. A., Manuel, F. K., Jones, J., Iszard, G., Murrey, J., Djojonegro, B., Wear, M., Space radiation and cataracts in astronauts, *Radiat Res.*, **156**, 5, Pt. 1, 460–466, 2001; Erratum: *Radiat Res.*, **156**, 6, 811, 2001.

R29 Rastegar, N., Eckart, P., Mertz, M., Radiation-induced cataract in astronauts and cosmonauts, *Graefes Arch. Clin. Exp. Ophthalmol.*, **240**, 543–547, 2002.

R30 Obe, G., Johannes, I., Johannes, C., Hallman, K., Reitz, G., Facius, R., Chromosomal aberrations in blood lymphocytes of astronauts after long-term space flights, *Int. J. Radiat. Biol.*, **72**, 6, 727–734, 1997.

R31 Nicogossian, A. E., Pool, S. L., Leach, C. S., Moseley, E., Rambaut, P. C. Concepts for NASA longitudinal health studies, *Aviat. Space Environ. Med.*, **54**, 12, 2, S68–72, 1983.

R32 Hamm, P. B., Nicogossian, A. E., Pool, S. L., Wear, M. L., Billica, R. D., Design and current status of the longitudinal study of astronaut health, *Aviat. Space Environ. Med.*, **71**, 6, 564–70, 2000.

R33 Hamm, P. B., Billica, R. D., Johnson, G. S., Wear, M. L., Pool, S. L., Risk of cancer mortality among the Longitudinal Study of Astronaut Health (LSAH) participants, *Aviat. Space Environ. Med.*, **69**, 2, 142–4, 1998.

R34 Peterson, L. E., Pepper, L. J., Hamm, P. B., Gilbert, S. L., Longitudinal study of astronaut health: mortality in the years 1959–1991, *Radiat. Res.*, **133**, 2, 257–64, 1993.

R35 Merriam, G. R., Focht, E. F., A clinical study of radiation cataracts and the relationship to dose, *Am. J. Roentgenol.*, **77**, 5, 759–785, 1957.

R36 Kirschvink, J. L., Homing in on vertebrates, *Nature*, **390**, 339–340, 1997.

R37 Walker, M. M., Diebel, C. E., Haugh, C. V., Pankhurst, P. M., Montgomery, J. C., Structure and function of the vertebrate magnetic sense, *Nature*, **390**, 371–376, 1997.

R38 Vanselow, K. H., Ricklefs, K., Are solar activity and sperm whale Physeter macrocephalus strandings around the North Sea related?, *Journal of Sea Research*, **53**, 319–327, 2005.

R39 Scheffel, A., Gruska, M., Faivre, D., Linaroudis, A., Plitzko, J., Schüler, D., An acidic protein aligns magnetosomes along a filamentous structure in magnetotactic bacteria, *Nature* (2005), online doi: 10.1038/nature04382.

R40 Thoss, F., Bartsch, B., Tellschaft, D., Thoss, M., Periodic inversion of the vertical component of the Earth's magnetic field influences fluctuations of visual sensitivity in humans, *Bioelectromagnetics*, **20**, 459–461, 1999.

R41 Thoss, F., Bartsch, B., Fritzsche, B., Tellschaft, D., The magnetic field sensitivity of the human visual system shows resonance and compass characteristics, *J. Comp. Physiol. A.*, **186**, 1007–1010, 2000.

R42 Thoss, F., Bartsch, B., Tellschaft, D., Thoss, M., The light sensitivity of the human visual system depends on the direction of view, *J. Comp. Physiol. A.*, **188**, 235–237, 2002.

R43 Thoss, F., Bartsch, B., The human visual threshold depends on direction and strength of a weak magnetic field, *J. Comp. Physiol. A.*, **189**, 777–779, 2003.

R44 Alania, M. V., Gil, A., Modzelewska, R., On the statistical relationship of solar, geomagnetic and human activities, *Advances in Space Research*, **34**, 1602–1606, 2004.

R45 Alania, M. V., Gil, A., Wieliczuk, R., Statistical analyses of influence of solar and geomagnetic activities on car accident events, *Advances in Space Research*, **28**, 4, 673–678, 2001.

R46 Zakharov, I. G., Tyrnov, O. F., The effect of solar activity on ill and healthy people under conditions of neurous and emotional stresses, *Advances in Space Research*, **28**, 4, 685–690, 2001.

R47 Dimitrova, S., Relationship between human physiological parameters and geomagnetic variations of solar origin, *Advances in Space Research* (2005) in press; online: 16 June 2005.

R48 Durand-Manterola, H. J., Mendoza, B., Diaz-Sandoval, R., Electric current induced inside biological cells by geomagnetic and atmospheric phenomena, *Advances in Space Research*, **28**, 4, 679–684, 2001.

R49 Cliver, E. W., Hudson, H. S., CMEs: How do the puzzle pieces fit together?, *Journal of Atmospheric and Solar Terrestrial Physics*, **64**, 231–252, 2002.

R50 Todd, P., Unique biological aspects of space radiation hazards – An overview, *Advances in Space Research*, **3**, 8, 187–194, 1983.

R51 Facius, R., Schafer, M., Bücker, H., Unique radiobiological aspects of high-LET radiation, *Advances in Space Research*, **4**, 10, 175–185, 1984.

R52 Todd, P., Walker, J. T., The microlesion concept in HZE particle dosimetry, *Advances in Space Research*, **4**, 10, 187–197, 1984.

R53 O'Brien, K., Radiation doses from solar flares at ground level, *Health Physics*, **36**, 1, 63–65, 1979.

R54 Wdowczyk, J., Wolfendale, A. W., Cosmic rays and ancient catastrophes, *Nature*, **268**, 510–512, 1977.

R55 Mill, A. J., The effect of large solar flares, *Health Physics*, **39**, 6, 1001–1002, 1980.

R56 Freier, P., Lofgren, E. J., Ney, E. P., Oppenheimer, F., Evidence for heavy nuclei in the primary cosmic radiation, *Physical Review*, **75**, 2, 213–217, 1948.

R57 Freier, P., Lofgren, E. J., Ney, E. P., Oppenheimer, F., The heavy component of primary cosmic rays, *Physical Review*, **74**, 12, 1818–1827, 1948.

R58 Schaefer, H. J., Exposure hazards from cosmic radiation beyond the stratosphere and in free space, *J. Aviat. Med.*, **23**, 334, 1952.

R59 Tobias, C. A., Radiation hazards in high-altitude aviation, *J. Aviat. Med.*, **23**, 345–372, 1952.

R60 Eugster, J., Die kosmische Strahlung und ihre Bedeutung für den bemannten Raumflug, in *Handbuch der Astronautik* (Schütte, K, Kaiser, H. K., eds), pp. 154–175, Akademische Verlagsgesellschaft Athenaion, Dr. Albert Hachfeld, Konstanz/Bodensee, 1964.

R61 Chase, H. B., Cutaneous effects of primary cosmic radiation, *J. Aviat. Med.*, **25**, 388–391 and 427, 1954.

R62 Chase, H. B., Post, J. S., Damage and repair in mammalian tissues exposed to cosmic ray heavy nuclei, *J. Aviat. Med.*, **27**, 533–540, 1956.

R63 Lebish, I. J., Simons, D. G., Yagoda, H., Janssen, P., Haymaker, W., Observations on mice exposed to cosmic radiation in the stratosphere. A longevity and pathological study of 85 mice, *Military Med.*, **124**, 835–847, 1959.

R64 Yagoda, H., Behar, A., Davis, R. L., Kraner, K. L., Filz, R., Hewitt, J., Haymaker, W., Brain study of mice exposed to cosmic rays in the stratosphere and report of nuclear emulsion monitoring. In four balloon flights from Bemidij, Minnesota, July–August 1960, *Military Med.*, **128**, 655–672, 1963.

R65 Solanki, S. K., Usoskin, I. G., Kromer, B., Schüssler, M., Beer, J., Unusual activity of the Sun during recent decades compared to the previous 11,000 years, *Nature*, **431**, 1084–1087, 2004.

R66 Selas, V., Hogstad, O., Kobro, S., Rafoss, T., Can sunspot activity and ultraviolet-B radiation explain cyclic outbreaks of forest moth pest species?, *Proc. R. Soc. Lond. B.*, **271**, 1897–1901, 2004.

R67 Lohmann, K. L., Pentcheff, N. D., Nevitt, G. A., Stetten, G. D., Zimmer-Faust, R. K., Jarrard, H. E., Boles, L. C., Magnetic orientation of spiny lobsters in the oceans: Experiments with undersea coil systems, *J. Exp. Biol.*, **198**, 2041–2048, 1995.

R68 Langham, W. H. (ed.), *Radiobiological Factors in Manned Spaceflight*, Publication 1487, National Academy of Sciences/National Research Council, Washington, DC, 1967.

R69 Grahn, D. (ed.), HZE-particle effects in manned spaceflight, National Academy of Sciences, Washington, DC, 1973.

R70 Willson, R. C., Mordvinov, A. V., Secular total irradiance trend during solar cycles 21–23, *Geophys. Res. Lett.*, **30**, 5, 1199–1203, 2003.

R71 Lassen, K., Friis-Christensen, E., Variability of solar cycle length during the past five centuries and the apparent association with terrestrial climate, *Journal of Atmospheric and Terrestrial Physics*, **75**, 8, 835–845, 1995.

R72 Friis-Christensen, E., Solar variability and climate: A summary, *Space Science Reviews*, **94**, 411–421, 2000.

R73 Svensmark, H., Cosmic rays and earth's climate, *Space Science Reviews*, **93**, 175–185, 2000.

R74 Veizer, J., Godderis, Y., François, L. M., Evidence for decoupling of atmospheric CO_2 and global climate during the Phanerozoic eon, *Nature*, **408**, 698–701, 2000.

R75 Shaviv, N. J., The spiral structure of the Milky Way, cosmic rays, and ice age epochs on Earth, *New Astronomy*, **8**, 39–77, 2003.

R76 Allkofer, O. C., Bagge, E., Trümper, J., Untersuchungen über die Strahlengefährdung der Flugbesatzungen und des Bodenpersonals von Düsenflugzeugen (I) Messung der Kontamination der Kabinenluft durch Spaltprodukte des Stratosphären-Fallout, *Atomkernenergie*, **7**, 5, 173–178, 1962.

R77 Allkofer, O. C., Bagge, E., Trümper, J., Untersuchungen über die Strahlengefährdung der Flugbesatzungen und des Bodenpersonals von Düsenflugzeugen (II) Messung der Außenhautkontamination von Düsenmaschinen, *Atomkernenergie*, **7**, 6, 204–216, 1962.

R78 Allkofer, O. C., Bagge, E., Trümper, J., Untersuchungen über die Strahlengefährdung der Flugbesatzungen und des Bodenpersonals von Düsenflugzeugen (III) Messung der Strahlenbelastung durch die kosmische Strahlung, *Atomkernenergie*, **7**, 7/8, 249–257, 1962.

R79 Foelsche, T., Radiation exposure in supersonic transport. NASA Technical Note, NASA TN D-1383, National Aeronautics and Space Administration, Washington, DC, August 1962.

R80 Reid, G. C., Solar variability and the Earth's climate: Introduction and overview, *Space Science Reviews*, **94**, 1–11, 2000.

R81 Pallé, E., Butler, C. J., O'Brien, K., The possible connection between ionization in the atmosphere by cosmic rays and low level clouds, *Journal of Atmospheric and Solar–Terrestrial Physics*, **66**, 18, 1779–1790, 2004.

R82 Schaefer, H. J., Radiation monitoring on project Mercury, *Aerospace Medicine*, **40**, 12, Sect. II, 1509–1516, 1969.

R83 Richmond, R., A review of Gemini and Apollo astronaut dosimetry data, *Aerospace Medicine*, **40**, 12, Sect. II, 1517–1526, 1969.

R84 Richmond, R. G., *The space radiation environment and its measurement – from Mercury spacecraft to the Mars mission*, SAE Technical Paper Series No. 941614, SAE International, Warrendale, 1994.

R85 Jordan, T. M., *Radiation protection for manned space activities*, NASA-CR-173202, Washington, 1983.

R86 Wilson, J. W., Overview of radiation environments and human exposures, *Health Physics*, **79**, 5, 470–494, 2000.

R87 National Academy of Sciences, National Research Council, *Radiation protection guides and constraints for space mission and vehicle design involving nuclear systems*, National Academy Press., Washington DC, 1970.

R88 National Council on Radiation Protection and Measurements, *Guidance on radiation received in space activities*, NCRP Report No. 98, Bethesda, 1989.

R89 National Council on Radiation Protection and Measurements, *Radiation protection guidance for activities in low Earth orbit*, NCRP Report No. 132, Bethesda, 2000.

R90 Irwin, W. P., Horner, A. J., Lohmann, K. J., Magnetic field distortions produced by protective cages around sea turtle nests: unintended consequences for orientation and navigation?, *Biological Conservation*, **118**, 117–120, 2004.

R91 Johnson, M. H., Kierein, J., Combined Release and Radiation Effects Satellite (CRRES): Spacecraft and Mission, *Journal of Spacecraft and Rockets*, **29**, 4, 556–563, 1992.

R92 Gussenhoven, M. S., Mullen, E. G., Sperry, M., Kerns, K. J., Blake, J. B., The effect of the March 1991 storm on accumulated dose for selected satellite orbits – CRRES dose models, *IEEE Transactions on Nuclear Science*, **39**, 6, 1, 1765–1772, 1992.

R93 Gussenhoven, M. S., Mullen, E. G., Brautigam, D. H., Improved understanding of the Earth's radiation belts from the CRRES satellite, *IEEE Transactions on Nuclear Science*, **43**, 2, 353–368, 1996.

R94 Daly, E. J., Leeuwen, F. V., Evans, H. D. R., Perryman, M. A. C., Radiation-belt and transient solar-magnetospheric effects on Hipparcos radiation background, *IEEE Transactions on Nuclear Science*, **41**, 6, 2376–2382, 1994.

R95 Reedy, R. C., Constraints on solar particle events from comparisons of recent events and million-year averages, in *Solar Drivers of Interplanetary and Terrestrial Disturbances* (Balasubramaniam, K. S., Kiel, S. L., Smartt, R. N., eds), pp. 429–436, ASP Conf. Ser. Vol.95, Astronomical Society of the Pacific, 1996.

R96 Bieber, J. W., Evenson, P., Dröge, W., Pyle, R., Ruffolo, D., Rujiwarodom, M., Tooprakai, P., Khumlumlert, T., Spaceship Earth observations of the Easter 2001 solar particle event, *Astrophysical Journal*, **601**, L103–L106, 2004.

R97 Truscott, P., *Ion-Nuclear Models for the Analysis of Radiation Shielding and Effects (IONMARSE)*, Contract Final Report, CR041585, QinetiQ Ltd, Farnborough, 25 June 2004.

R98 Regener, E., Pfotzer, G., Vertical intensity of cosmic rays by threefold coincidences in the stratosphere, *Nature*, **136**, 718–719, 1935.

R99 Pfotzer, G., Coincidence counter measurements of the intensity depth curve, *Zeits. f. Physik*, **102**, 23–58, 1936.

R100 Agostinelli, S., Allison, J., Amako, K., Apostolaki, J., Araujo, H., Arce, P., Asai, M., Axen, D., Banerjee, S., Barrand, G., *et al.*, GEANT4 – a simulation toolkit, *Nuclear Instruments and Methods*, A506, 250–303, 2003.

R101 O'Brien, K., Luin, *A code for the calculation of cosmic-ray propagation in the atmosphere (update of HASL-275)*. Report EML-338, Environmental Measurements Laboratory, New York, 1978.

R102 Smart, D. F., Shea, M. A., The space-developed dynamic vertical cutoff rigidity model and its applicability to aircraft radiation dose, *Advances in Space Research*, **32**, 1, 103–108, 2003.

R103 Bailey, J. V., Hoffman, R. A., English-II, R. A., Radiological Protection and Medical Dosimetry for the Skylab, in *Biomedical Results from Skylab* (Johnston, R. S., Dietlein, L. F., eds), Chapter 9, NASA SP-377, Washington, DC, 1977.

R104 Bailey, V. J., Radiation Protection and Instrumentation, in *Biomedical Results of Apollo* (Johnston, R. S., Dietlein, L. F., Berry, C. A., eds), Section II, Chapter 3, NASA SP-368, Washington, DC, 1975.

R105 Reitz, G., Space radiation dosimetry, *Acta Astronautica*, **32**, 715–720, 1994.

R106 Badhwar, G. D., Atwell, W., Reitz, G., Beaujean, R., Heinrich, W., Radiations measuremets on the Mir Orbital Station, *Radiation Measurements*, **35**, 5, 393–422, 2002.

R107 Heinrich, W., Wiegel, B., Ohrndorf, T., Bücker, H., Reitz, G., Schott, J-U., LET spectra of cosmic-ray nuclei for near earth orbits, *Radiation Research*, **118**, 63–82, 1989.

R108 Reitz, G., Beaujean, R., Kopp, J., Leicher, M., Strauch, K., and Heilmann, C., Results of dosimetric measurements in space missions, *Radiation Protection Dosimetry*, **70**, 413–418, 1998.

R109 Reitz, G., Facius, R., Sandler, H., Radiation protection in space, *Acta Astronautica*, **35**, 4–5, 313–338, 1995.

R110 International Commission on Radiological Protection, *RBE for Deterministic Effects*, ICRP Publication 58, Elsevier, 1990.

R111 Barendsen, G. W., RBE for non-stochastic effects, *Advances in Space Research*, **12**, 2–3, 385–392, 1992.

R112 Watts, Jr., J. W., Parnell, T. A., Heckman, H. H., Approximate angular distribution and spectra for geomagnetically trapped protons in low-earth orbit, in *High Energy Radiation Backgrounds in Space*, pp. 75–85, Proc. AIP Conf. 186, American Institute of Physics, 1989.

R113 Armstrong, T. W., Colborn, B. L., Benton, E. V., Model calculations of the radiation dose and LET spectra on LDEF and comparisons with flight data, *Radiation Measurements*, **26**, 6, 751–764, 1996.

R114 Reitz, G., Berger, T., *The MATROSHKA Facility – Dose Determination during an EVA*, Radiation Protection Dosimetry, 2005, in press.

R115 Golightly, M. J., Hardy, A. C., Atwell, W., Hardy, K., Description, analysis and impact of major solar activity during recent U.S. Shuttle missions, *Advances in Space Research*, **12**, 2–3, 335–338, 1992.

R116 Dachev, Ts. P., Matviichuk, Yu. N., Bankov, N. G., Semkova, J. V., Koleva, R. T., Ivanov, Ya. J., Tomov, B. T., Petrov, V. M., Shurshakov, V. A., Bengin, V. V., Machmutov, S., Panova, N. A., Kostereva, T. A., Temny, V. V., Ponomarev, Yu. N., Tykva, R., Mir radiation dosimetry results during the solar proton events in September–October 1989, *Advances in Space Research*, **12**, 2–3, 321–324, 1992.

R117 Horneck, G., Facius, R., Reichert, M., Rettberg, P. *et al.*, *HUMEX, a Study on the Survivability and Adaptation of Humans to Long-Duration Exploratory Missions*, ESA SP-1264, Noordwijk, 2003.

R118 O'Brien, K., Friedberg, W., Sauer, H. H., Smart, D. F., Atmospheric cosmic rays and solar energetic particles at aircraft altitudes, *Environment International*, **22**, Suppl. 1, S9–S44, 1996.

R119 Ballard, T., Lagorio, S., deAngelis, G., Verdecchia, A., Cancer incidence and mortality among flight personnel: a meta-analysis, *Aviation Space and Environmental Medicine*, **71**, 3, 216–224, 2000.

R120 Blettner, M., Zeeb, H., Auvinen, A., Ballard, T. J., Caldora, M., Eliasch, H., Gundestrup, M., Haldorsen, T., Hammar, N., Hammer, G. P., Irvine, D., Langner, I., Paridou, A., Pukkala, E., Rafnsson, V., Storm, H., Tulinius, H., Tveten, U., Tzonou, A., Mortality from Cancer and other causes among male airline cockpit crew in Europe, *Int. J. Cancer*, **106**, 6, 946–952, 2003.

R121 Zeeb, H., Blettner, M., Langner, I., Hammer, G. P., Ballard, T. J., Santaquilani, M., Gundestrup, M., Storm, H., Haldorsen, T., Tveten, U., Hammar, N., Linnersjö, A., Velonakis, E., Tzonou, A., Auvinen, A., Pukkala, E., Rafnsson, V., Hrafnkelsson, J., Mortality from cancer and other causes among airline cabin attendants in Europe: A collaborative cohort study in eight countries, *American Journal of Epidemiology*, **158**, 1, 35–46, 2003.

R122 Band, P. R., Spinelli, J. J., Ng, V. T. Y., Moody, J., Gallagher, R. P., Mortality and cancer incidence in a cohort of commercial airline pilots, *Aviation Space and Environmental Medicine*, **61**, 4, 299–302, 1990.

R123 Reitz, G., Radiation environment in the stratosphere, *Radiation Protection Dosimetry*, **48**, 1, 5–20, 1993.

Web resources

W1 http://www.phy6.org/Education/wmap.html (*The Exploration of the Earth's Magnetosphere*)
W2 http://see.msfc.nasa.gov/ire/model_trap.htm (*AP8 Trapped Proton Model*)
W3 http://see.msfc.nasa.gov/ire/model_elec.htm (*AE8 Trapped Electron Model*)
W4 http://web.dmi.dk/fsweb/solarterrestrial/sunclimate/SCL.txt (Solar Cycle Lengths)
W5 http://science.msfc.nasa.gov/ssl/pad/solar/sun_wind.htm (Solar wind)
W6 http://www-istp.gsfc.nasa.gov/Education/wbirthrb.html (Transient radiation belts)
W7 http://aadc-maps.aad.gov.au/aadc/gle/index.cfm (AADC GLE lists)
W8 http://wwwasd.web.cern.ch/wwwasd/geant4/geant4.html (GEANT4)
W9 http://www.fluka.org/ (FLUKA)
W10 http://www0.gsf.de/epcard2/index.phtml (EPCARD)
W11 http://www.faa.gov/education_research/research/med_humanfacs/aeromedical/radiobiology/cari6/index.cfm (CARI)

12

Effects on spacecraft hardware and operations

Alain Hilgers, Alexi Glover and Eamonn Daly

12.1 INTRODUCTION

The most obvious environmental characteristics of space are low gravity, very low ambient pressure, and strong thermal gradients, which all result in special engineering approaches being necessary in building and operating space systems. However, the background environmental characteristics consisting of residual gases, various plasma regimes, energetic particle radiation, electric and magnetic fields, and solid particle impacts, can all result in important engineering problems. In this chapter, these various space environment components are presented, along with their effects on space systems. A description of the causes and consequences of the variability of the space environment affecting spacecraft operation and hardware is given, followed by a review of the relevant data source available for assessment and monitoring. Finally, the possible evolution of the relevant data acquisition, distribution and modelling infrastructures is outlined.

12.1.1 High-energy charged particles

The high-energy charged particle environment has three main sources:

- High-energy electrons, protons and other ions trapped in the Earth's (or in some other planet's) magnetic field.
- High-energy protons and heavier ions resulting from energetic 'events' close to the Sun and elsewhere in the heliosphere.
- Cosmic rays, originating outside the Solar System.

In the context of effects on space systems, 'high energy' is usually thought of as being above energies of 0.1–1 MeV. Since the main effects of these particles are related to radiation damage, the environment is sometimes referred to as the 'penetrating radiation environment'.

Table 12.1. The effects of the space environment on space systems.

Environment	Main effect on space systems
Cosmic rays (~100 MeV to ~1 GeV) High-energy solar particles (~100 keV to ~300 MeV (sometimes GeV))	Ionizing and non-ionizing dose effects (components and material degradation, living cell damage) Single-event effects
Radiation belt particles (electrons: ~100 keV to ~10 MeV; protons: ~1 MeV to ~1 GeV)	Ionizing and non-ionizing dose effects (components and material degradation, living cell damage) Single-event effects Deep dielectric charging
Ionospheric plasma (~0.1 eV) Plasmasphere (~1 eV) Solar wind plasma (electron: ~1 eV; proton: ~0.1 to 1 keV)	Particle flux effect (electrical current noise, erosion, sputtering) Small electrostatic potential
Auroral filament plasma (~100–10 keV) Plasmasheet plasma (~1 to 10 keV)	Particle flux effect (electrical current noise, erosion, sputtering) Strong electrostatic potential
Visible and IR photons	Thermal
UV, X and γ photons	Ionizing effects
Radio waves	EM noise
Geomagnetic field	Magnetic torque on current loop Induced electric field in conductors
Interplanetary magnetic field	Torque Induced electric field
Neutral atoms	Spacecraft drag Chemical reactions on surfaces, and associated degradation
Microparticles	Puncture Transient dusty plasma cloud

Figure 12.1 (see colour section) shows the ranges of electrons and protons in aluminium as functions of energy. It can be seen that to penetrate typical spacecraft walls of ~1–5 mm thickness, electrons have to have more than ~500 keV and protons more than ~5 MeV. In addition, photovoltaic solar cells have glass covers of the order of 100 μm thick. Particles having energies able to penetrate these thicknesses are plentiful in the radiation belts, and in the heliosphere following solar particle events. The effects of these radiation environments are summarized in Table 12.1.

For Earth-orbiting spacecraft the main concern with respect to radiation has been component damage from radiation belt particles. Figure 12.2 (see colour section) shows the doses expected behind a sphere of aluminium shielding for various satellite orbits, as functions of the sphere radius. This shows that hundreds of kilorads are easily encountered, particularly in the commercially interesting geostationary and GPS-type orbits. Most commercial components fail when exposed to doses in the kilorad range.

While dose effects continue to be very important – especially as system builders are keen to fly advanced and often commercial-off-the-shelf (COTS) components – the problems caused by interactions of single ions (and neutrons) are growing. These ‘single event effects’ (SEE) occur when the ionization track in a sensitive region of a component results in a status flip or signal ‘transient’, or in destructive current paths. The charge deposition can also arise from the products of nuclear spallation reactions between protons or neutrons and nuclei of the materials of the component. Therefore, the proton radiation belts, solar particle events and cosmic rays are all important, as are their variations. During the design phase, careful assessments of SEE rates make use of environment models, together with component sensitivity characterization tests. Error rates in memories can often be considered as little more than a nuisance, and errors can usually be corrected. However, errors ocurring in programme memory or within processors can be more serious. While care is taken, there is always a residual risk of this type of error, and when they occur they can cause effects such as attitude control failure, system crashes, and so on. Such effects can also arise when an SEE occurs in analogue circuits where a transient pulse may appear on an output and be latched or sensed by a comparator, resulting in an unwanted status being flagged. Consequent effects at system level have to be considered, and this can result in questions being posed as to how many SEEs may appear over given times, or how often the rate would exceed a particular threshold. For evaluation of such questions during solar particle events, for example, examination of historical data over long periods has to be performed, rather than assessments of peak rates at the peak of a worst-case event.

Radiation background in sensors is a particular class of SEE, and the sensor systems or subsequent processing have to be designed to remove the effects of background. During enhancements of the radiation belts during geomagnetic storms or during solar particle events, the background level may render the sensor data useless. If the sensor is on a platform system (in particular star trackers as part of the attitude control system), the high level of background may interrupt normal functioning. Cases have been reported of attitude loss due to background in star trackers.

Sensors can be very sensitive to radiation damage, even from relatively low-energy particles. For example, protons of only a few hundred keV can penetrate to the sensitive parts of CCDs, and the open nature of X-ray optics makes X-ray missions particularly sensitive. Both XMM and Chandra have had to take special operational measures to protect the CCDs while in orbit, by blocking or shifting CCD locations. Many astrophysics missions planned for the future explicitly take space weather into account, and may fly radiation monitors or plan operations and data-processing to account for space weather events such as radiation belt encounters and solar particle events. Future developments of advanced systems give rise to concern mainly with regard to single-event effects. The methods used in the past have been adequate for the design of circuits on components, but in the future, the very small 'feature sizes' of electronics means that currently employed test and analysis methods may no longer be adequate or suitable.

A further important class of radiation effect is the charging inside dielectric materials or of ungrounded metalic parts inside spacecraft. High energy electrons from the radiation belts cause this charging, and discharges which can result cause electromagnetic interference with systems.

Also, manned space programmes are extremely concerned with radiation effects on astronauts (see Chapter 11). Clearly, energetic solar particle events can represent a serious health hazard. The large solar particle event of August 1972 occurred between the flights of Apollo 16 and Apollo 17. If it had occurred during a flight, the consequences could have been fatal (suited astronaut skin dose >4 krem; dose in the Lunar Module >1 krem; dose in the Command Module >300 rem). While the Earth's geomagnetic field provides some protection, for spacecraft in low-altitude and low-inclination orbits, from solar energetic particles and galactic cosmic rays, the International Space Station's (ISS) inclination of 53° means that some of the orbit is more exposed. In addition, at high latitudes the orbit encounters the low-altitude extension of the outer, electron, radiation belt. Compared to many past Shuttle flights, this environment is more hazardous, and requires greater care, for example, in performing extravehicular activities. In the future, missions returning to the Moon and going to Mars will require even greater attention to radiation hazards.

12.1.2 Plasmas

The plasma environment is found above an altitude of a few hundred kilometres. At low altitudes it results mainly from the ionization of the upper atmosphere by solar UV radiation, and constitutes there the ionosphere, which is important for ground-to-ground signal propagation, as well as propagation from and to satellites. Solar and geomagnetic activity lead to variations in the ionospheric density, which in turn may impact systems which involve radio-wave propagation through the ionosphere, such as communication or navigation systems (see the relevant chapters in this book).

At higher altitude, the plasma environment is strongly organized by the Earth's magnetic field and constitutes, in this region of geomagnetic influence, the magnetosphere (Figure 12.3). The whole magnetosphere is embedded in the plasma expanding from the solar atmosphere: the solar wind. The magnetospheric plasma is composed of a mixture of ions (essentially hydrogen, helium and oxygen) and electrons. The typical energy and density of the plasma in space vary considerably from one region to another (Figure 12.4). From the engineering point of view, one can consider three main plasma regions in the magnetosphere: The cold ionospheric plasma region, which includes the ionosphere – the closest to Earth – with low energy (typically 0.1 eV) and high density (up to $10^6\,\mathrm{cm}^{-3}$); the plasmasphere, which is mainly the equatorial expansion of the ionosphere with low energy (1 eV) and

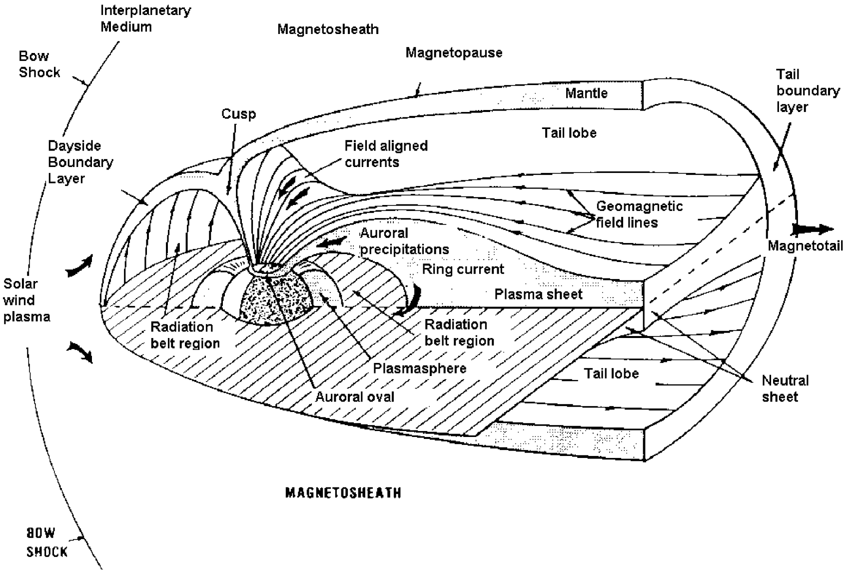


Figure 12.3. Sketch of the Earth's magnetosphere embedded in the solar wind.

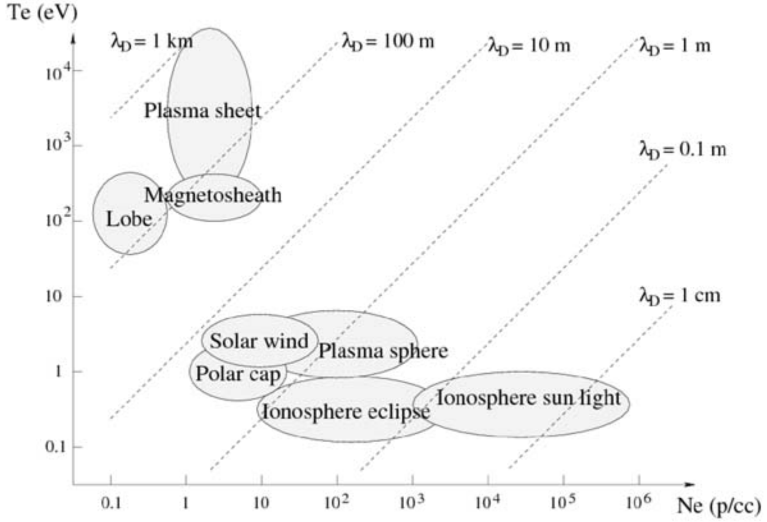


Figure 12.4. Typical range of density and temperature value in various plasma regions of the Earth magnetosphere.

moderate plasma density (10–1,000 cm^{-3}); and the high-temperature plasma regions (aurorae, plasmasheet, substorm and storm precipitations) where the plasma has been energized up to energies typically in the range of 100–10 keV, and with usually low plasma density (of the order of 1 cm^{-3}). The solar wind is where the plasma is cool (of the order of 1 eV) and dense (of the order of 10 cm^{-3}), but with high drift velocity (300–2,000 km s^{-1}). It is worth noting that there is also a large region of the magnetosphere that includes the polar caps and the major part of the tail, which is characterized by extremely low density (typically 0.1 cm^{-3}) and rather high energy (10–100 eV).

Bodies immersed in space plasma are submitted to fluxes of charged particles on surfaces, and related electrical phenomena (including induced electrical current, electrostatic potential, and arc or coronal discharges). The theoretical potential that an isolated surface would reach in a plasma is typically negative and of the order of the ambient electron temperature in high-density regions or in eclipse, while it is typically positive and of the order of the energy of emitted photo-electrons in low density regions in sunlight. Therefore, high-level negative charging phenomena of surfaces (up to a few kV negative) and most of the powerful discharge phenomena are observed in the hot plasma region, especially in eclipse. In the cold ionospheric plasma region and in the solar wind, the induced electrostatic potential on surfaces is generally small (from -1 V to a few tens of V positive) but not necessarily negligible (for example, scientific applications have strong requirements regarding electrostatic cleanliness). In the lobe and the polar cap, where the density is very low, the spacecraft potential structure may reach several tens of V positive.

Most of space engineering practices related to orbital plasma environment issues consist in limiting the expected charging level between surfaces within empirical threshold values above which powerful electrostatic discharges may occur (see ECSS-E-20-6). Normally the threshold electric field is of the order of 10^7 V/m between a dielectric element and a metallic element. In practice, failures occur on microscopic-scale surfaces such as edges, and spikes where electric fields are the strongest but are difficult to measure or to estimate. Therefore, empirical criteria based on more macroscopic potential measurements are generally used. Because the microscopic fields depend on the structure of the material, the macroscopic potential at which discharge is observed can depend on sample geometry, the type of discharge, the environment, material properties and surface state.

For instance, a critical value for the surface voltage of a dielectric with respect to a more negative metal (inverted potential gradient) is considered to be 500 V. For a dielectric adjacent to a more positive metal (normal potential gradient), the critical voltage is around 1 kV. Electric fields of 10^6–10^7 V/m normally lead to punch-through, but the discharge does not propagate along the surface (lead to flash-over) without a substantial transverse electric field, and so surface potentials of greater than 1 kV are an additional characteristic for these discharges. Discharges initiated by the dense plasma environment in low Earth orbit can occur, with potential differences above around 100 V. These discharges are illustrated in Figure 12.5.

Radiation and plasma effects are not always easily separated. Ions from the plasma with energy as low as 50 keV can penetrate coating materials and deposit all of their energy, resulting in accumulated damage. Electrons with energy of the order of 1 keV can create SEU on CCDs directly or indirectly accessible (through multi-scattering) from space (Hilgers *et al.*, 1999). Furthermore, the radiation envir-

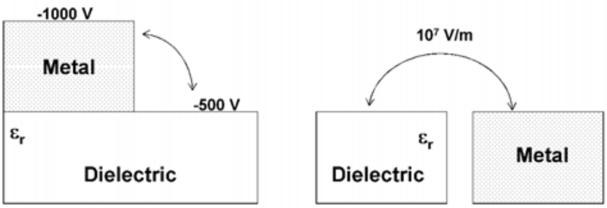


Figure 12.5. Critical potential and field on multi-material elements immersed in plasma.

onments modify the conductivity of materials, and therefore influence surface charging behaviour.

12.1.3 Electromagnetic environment

The electromagnetic environment includes electromagnetic waves from the radio to the gamma range, the interplanetary magnetic and geomagnetic field, and interplanetary and Earth electric field. The electromagnetic wave background has mainly a solar origin, but a significant component is provided by the Earth. There is also a non-negligible component generated by human activity.

The main effect of electromagnetic waves is thermal, and is mainly related to photons in the infrared and visible range mainly of solar origin. High-energy photons (UV, X and gamma) have ionizing effects, thereby generating secondary particles or change of electrical properties of materials. The high-energy photon background has mainly a solar origin, but a minor component is generated by precipitating particles in the auroral regions.

The geomagnetic field direction is a reference used by magneto-torquers for spacecraft attitude control. Its variability is of direct concern. There is an electric field generated in conductive parts moving across the geomagnetic field. This effect is especially the basis of a system based on long conductive cable, either for generating electrical power or drag.

The interplanetary magnetic field is typically much weaker than the geomagnetic field, and is much more variable. It is therefore not used as a reference for spacecraft attitude control, and usually has no impact on spacecraft systems.

12.1.4 Atomic environment

The atomic population is mainly of atmospheric origin, and is therefore more significant at low altitude. The major concern is the drag effect on spacecraft, which affects trajectory and re-entry. Another concern is the chemical action of atomic oxygen on spacecraft surfaces. Small fluxes of energetic neutral atoms produced by charge exchange in regions of high-energy particles are also observed. The main interest of energetic neutral atoms in this context is therefore mainly as a tracer of these high-energy charged-particle regions.

12.1.5 Micro-particle environment

The micro-particle environment is dominated by micrometeoroids of interplanetary origin, interstellar dust and orbital micro-debris. The latter ones are of artificial origin. The main concern is the direct mechanical effect of an impact, which may modify or puncture spacecraft surfaces. During the impact process a dense cloud of gas with a significant ionized component is also generated, which may also lead to electrical interactions.

12.1.6 Environment and effects: summary

The corpuscular and electromagnetic environment described above is very diverse, and has a potentially strong impact on spacecraft and space systems. A summary of the effects and of the environmental cause is given in Table 12.1. It must be noted that the effects of the space environment are not only adverse. Some systems are designed to make use of the space environment by extracting information (scientific instruments, magneto-torquer) or energy (electrodynamic tether).

12.2 DYNAMICS AND VARIABILITY OF THE SPACE ENVIRONMENT

It is very important, in the various phases of a space mission from spacecraft development to operation, to have access to a good quantitative description of the relevant space environment. The most complete description for each component is usually the flux density as a function of energy, time and location. The range of most of the parameters, however, may vary by orders of magnitude as a function of time and location, and the use of simplified averaged values is not always possible. The dynamics and time variability has then to be taken into account.

The space environment components described above are not independent of each other. The dynamics of the solar–terrestrial environment is driven by fluxes of emitted particles from the Sun, photons, electrons, protons, and a magnetized solar wind which interacts with the Earth's atmosphere and the geomagnetic field via numerous processes taking place in the solar photosphere and corona, the solar wind, the magnetosphere, the ionosphere and the thermosphere. In addition, some particles (such as cosmic dust and galactic cosmic rays) have an interstellar or interplanetary origin, and interact with the solar wind before reaching the near-Earth environment. This section primarily aims at describing the causes and consequences of the variability of the space environment.

12.2.1 Space environment and solar–terrestrial dynamics

Photons from the Sun travel the distance to Earth (about 150 million km) within a few minutes before impacting the atmosphere. The background ionosphere is primarily generated by photoionization of the Earth's atmosphere by solar photons in the UV and X range. Enhanced ionization by high-energy particles also occurs, but since they are deflected by the Earth's magnetic field, this phenomenon is mainly significant in the polar regions.

The solar-wind plasma blown from the Sun fills a considerable region of space far beyond the known planets, and constitutes the heliosphere in the interstellar environment. The heliosphere is crossed by a number of interplanetary and interstellar particles, including dust and cosmic rays. Cosmic rays and high-energy solar particles are scattered by solar wind inhomogeneities, tending to have an isotropic distribution. There are also interaction processes of the solar wind, leading to the generation of highly energetic particles.

Cosmic rays, while impacting the atmosphere, generate secondary high-energy neutrons which eventually decay into high-energy protons that populate the radiation belts. The solar cosmic rays – which have lower energy but higher flux than the galactic cosmic rays – can still produce ionization deep within the Earth's atmosphere.

Solar wind material travels from the solar surface to Earth's orbit within several days. The interaction of the solar wind with the Earth's magnetic field leads to a complicated system of plasma convection and electrical current and the resulting magnetospheric field. The energy loaded into the magnetosphere is related to the local solar wind density, speed, and the interplanetary magnetic field. This energy is stored in the geomagnetic tail in plasma and magnetic flux, and is released intermittently in events known as substorms during which most of the plasma is released away from Earth, although a significant part is accelerated toward Earth, leading to particle precipitations. The substorm development involves drastic magnetospheric changes. Enhanced ionization and heating of the atmosphere occur due to auroral and radiation belts precipitations in the atmosphere during geomagnetic substorms and storms. Auroral particle precipitations also generate UV and X photons. A small fraction of the dissipated energy is emitted in the interplanetary medium via radio waves.

Among the space environment components discussed in Section 12.1, the microparticle environment is least connected to the other components. Nevertheless, the low-altitude population is subject to thermospheric drag. Also, light debris is subject to photon radiation pressure, and there are possible geomagnetic influence on the debris trajectory once electrically charged by photoionization and interaction with charged particles.

12.2.2 Variability of the space environment

The space environment described above is far from being static. There are both periodic and sporadic variations, with amplitudes varying from a few percent to several orders of magnitude. Solar activity and geomagnetic activity are the terms commonly used to define the variable components of, respectively, the solar output and the Earth's magnetic environment. Other variations of the space environment that are related to, or that occur in parallel with, the solar and geomagnetic activities, are also discussed below.

12.2.2.1 Solar variability

The Sun varies over both long and short timescales. While the total irradiance from the Sun varies by a fraction of a percent during the course of a solar cycle, important localized variations of the photon flux are observed in the radio, visible, UV and X-ray range. Short-term variations, lasting from minutes to hours, are related to eruptive phenomena. Medium-term variations of the order of days or months are modulated by the 27-day rotational period of the Sun, and are often associated with the appearance of active regions and coronal holes on the solar disk. Long-term variability is related to the polarity reversal of the Sun's magnetic field, leading to the 11-year solar cycle.

Eruptive phenomena

There are two main types of solar eruptive events which have direct consequences on the space environment: Coronal Mass Ejections and solar flares.

Coronal Mass Ejections (CME)

CMEs are eruptions of magnetized plasma from the Sun, seen as bright features expanding outwards from the corona in coronagraph images. CMEs are ejected from the Sun with speeds of up to 3,000 $\mathrm{km s^{-1}}$. A typical mass estimate for these strustures is 10^{15} g. The CME further propagates in the interplanetary medium to become an interplanetary coronal mass ejection (ICME), observed at the Earth's L1 Lagrangian point, where it constitutes a large-scale disturbance of the solar wind properties. Acceleration of particles up to several times 100 MeV may take place ahead of the fastest CME. These particles also propagate along magnetically connected interplanetary field lines.

Solar flares

Solar flares are sudden and rapid releases of magnetic energy in the solar atmosphere. During a flare, energy is explosively released in various forms, including particle acceleration, plasma heating and acceleration, and enhanced radiation fields. The amount of energy released is of the order of 10^{25} J. Solar flares occur in active regions, characterized by strong and complex magnetic fields. The high-energy particles are emitted into the interplanetary medium along magnetically connected interplanetary field lines.

Slow variations

Initially there are two remarkable long-term variations of the averaged solar activity: the solar rotation period – 27 days – and the solar cycle – 11 years. The average time from cycle minimum to maximum is 4.3 years, and from maximum to minimum, 6.6 years. The magnetic cycle of the Sun has an average duration of 22 years. The maximum activity of the solar cycle corresponds to a magnetic polarity reversal. The number of sunspots and numerous solar features are well correlated with the solar activity cycle. The background electromagnetic energy emitted in the radio range and in the UV to X range has a slowly varying intensity, and is closely linked to the number of active regions present on the disk.

12.2.2.2 Interplanetary medium variability

High/low-speed solar wind

The solar wind is not a uniform medium. There are high-speed and low-speed solar wind streams. The high-speed solar wind streams originate from coronal holes, which may be steady over periods of months and rotate with the Sun. As a result, properties of the solar wind, when reaching the Earth's magnetosphere, often display a 27-day periodicity – in particular during periods of low solar activity.

Irregularities and shocks

In addition, important solar wind irregularities and shocks result from the interaction of streams with different speeds, and from the propagation of CMEs in the interplanetary medium. These irregularities may accelerate and scatter high-energy particles. As a consequence, a decrease of about 20% of the flux of cosmic-ray particles is observed during solar maximum. A decrease of the order of 10% may also be observed on smaller timescales during periods when interplanetary ICMEs are observed to cross the Earth's orbit. Conversely, a significant increase of radiation may be observed as the ICME passes the Earth, if it is fast enough to accelerate particles.

12.2.2.3 Magnetospheric variability

Sporadic geomagnetic variations

As indicated above, the geomagnetic environment is constantly interacting with the solar wind. There is, therefore, geomagnetic activity directly induced by the solar wind variability. Furthermore, there are sporadic and violent reconfigurations of the magnetosphere that lead to particle acceleration and injection into the inner magnetosphere, including the radiation belts and the auroral region. This leads to auroral emissions in the visible, UV and X range, as well as some radio emissions in the decametric and kilometric wavelength range. The associated change of the current system in the magnetosphere and the ionosphere induces perturbations of the geomagnetic field that are even detectable on ground.

Sporadic radiation belt variations

During storms there are injections of high-energy particles toward the lower altitudes. During this process, particles gain further energy and populate the ring current and the radiation belts. Several hours after injection, a significant increase of electrons flux in the MeV range is observed at geostationary orbit. The exact mechanisms of this energization process is still a research topic. Exceptionally strong injection events that seem directly related to solar wind geo-effective structures, may lead to the sudden creation of secondary trapped proton belts in a region normally nearly empty of high-energy particles (the slot), which may remain there for several months.

Diurnal and annual variations

Diurnal and annual variations of the space environment are due to the rotation and revolution of the Earth. One effect is to modify the quantity of incident UV and X flux, and therefore the ionospheric plasma source. For instance, the low-altitude ionosphere nearly disappears during the night. Another effect is the change of the geomagnetic field due to the inclination of the Earth's magnetic dipole with respect to the Earth's rotation axis. Plasma and radiation regions that are tied to the geomagnetic field topology move accordingly.

Debris at low altitude are influenced by the thermosphere, and are therefore subject to solar cycle variations (a few percent less during solar maximum).

12.2.2.4 Variability: summary

In Table 12.2 a summary of the variable components of the space environment is given, together with the main cause of the variability, typical frequency occurrence, and timescale. These numbers are mainly indicative, and the uncertainty is high due to difficulties related to the sparsity of the observations and the sporadicity of some of the phenomena involved.

12.3 SPACE ENVIRONMENT MONITORING FOR SPACECRAFT

It has been shown in the previous Sections that the space environment is highly variable, and that this variability is primarily linked to solar activity and is also partly connected to interplanetary and interstellar phenomena.

Regarding space environment effects, spacecraft and space application engineers may have one or more of the following requirements:

(1) Extracting accurate specifications of the environment from a good statistical basis.
(2) Forecasting the long-term space environment conditions for spacecraft design.
(3) Forecasting the short-to-medium term space environment conditions for operation planning.
(4) Evaluating in real time whether the space environment conditions are compatible with planned operations or are requiring specific actions.
(5) *A posteriori* determination of the condition of the space environment that prevailed at dates requiring in-depth analysis or data quality label (for example, date of spacecraft anomaly or date of scientific data collection).

During the last few decades, a large amount of data have been gathered on the solar–terrestrial dynamics and the related space environment, and significant progress has been made in understanding the physical processes involved. Regarding the first two requirements from the above list, there are relatively good models of the averaged

Table 12.2. Variable space environment components, main causes, frequency and timescales.*

Environment	Variability	Frequency	Timescale
Galactic cosmic rays	Scatter by solar wind	Sporadic structures with 11-year modulation	Hours 11 years
Solar proton events	Generated by flares		Hour
	Generated by CMEs	$\sim$6 per year	Day
Radiation belt variation	Direct injection into inner belt	$\sim$2 per solar cycle	Hour
	Storm injection in outer belt	$\sim$1–7 per month	Hour
	Thermospheric enhanced absorption	$\sim$6 per day	Hour
Ionospheric density enhancement	Flare-related	Sporadic	Hour
	Storm-related	$\sim$1–7 per month	Hour
Plasmasphere expansion	Shrinks during substorm Expands during quiet time	$\sim$6 per day	Hour
Solar wind (density, speed and magnetic field)	Solar rotation	27-day modulation	
	Coronal hole dynamics	Modulated by solar cycle	Month or Year
	CME	$\sim$10 to $\sim$100 per month	Hour
	Shock-related flare	Sporadic	Hour
Auroral plasma	Substorm	$\sim$6 per day	Hour
Plasmasheet plasm Earthward injection	Substorm	$\sim$6 per day	Hour
UV, X photons, Radio waves	Background	Modulated by solar cycle	Day
	Flares	Sporadic	Hour
Radio waves	Generated by CME	$\sim$10 to $\sim$100 per month	Hour Day
	Substorm	$\sim$6 per day	Hour
Geomagnetic field variation	Solar wind pressure increase	Sporadic	Hour
Geomagnetic field variation	Storm	$\sim$1–7 per month	Hour
	Substorm	$\sim$6 per day	Hour
Thermospheric heating	Mainly solar EUV flux	Modulated by solar cycle	Hour
	Storm	$\sim$1–7 per month	Hour
Micrometeoroids	Meteor streams	2–3 per year	Days
Debris	Human space activity	N/A	Trend

[*] Only major solar proton events are considered with flux $>1\,\mathrm{cm}^{-2}\,\mathrm{sr}^{-1}\,\mathrm{s}^{-1}$ for protons $>10\,\mathrm{MeV}$. The frequency given is a long-term average. There are years without any major event observed.

space environment conditions and their slowly varying components (see a compilation of models and data for space engineering application in ECSS-E-10-04). However, there are still significant gaps in the model due to the undersampling of some regions and states of the radiation belts, and of the sparse observation of the relatively rare solar particle events (Rosenqvist and Hilgers, 2003). Engineering specifications could also be improved if reliable predictions of the amplitude of the solar cycle and related phenomena were available.

Meeting the last three requirements in the above list may prove more difficult, owing to their dependence on sporadic and transient components of the space environment. These components may lead to variations of space environment parameters over several orders of magnitudes and, at the present time, very few predictive quantitative models exist. Given our current understanding, these requirements cannot be fulfilled without near real-time measurements of the relevant space environment parameters.

There already exists a long record of observations of the solar and geomagnetic activity originally initiated for scientific purposes. Thanks to technological advances, these data can now be made available in near-real-time. Existing data are briefly described in the following Section, according to the observational techniques. For the purposes of this chapter, only data already recorded over a significantly long period of time, and foreseen to be continued for the next few years, are listed here.

12.3.1 Ground-based measurements

Ground-based measurements are the most ancient ones, and are therefore relatively well understood and widely used. Several datasets cover many solar cycles, and are therefore suitable for long-term trend estimates. The difficulties with ground-based measurements are the need for intercalibration – the effect of local disturbances that need correction or may completely prevent observations (such as atmospheric weather conditions affecting solar observation) – and the often incomplete coverage in longitude and latitude (critical for magnetospheric and ionospheric phenomena). Furthermore, although these services have been maintained for a long time, their funding depends on local policy and may disappear in the near future.

12.3.1.1 Solar observatories

White light/sunspot number

These observatories provide daily observations of the Sun, with nearly continuous coverage thanks to their large number and widely spread distribution over the globe. Currently, about sixty observatories contribute to the provision of the sunspot number index R – a measure of the area of solar surface covered by sunspots. The most frequently used long-term indicator of solar activity is the 12-month running mean value of sunspot numbers, R12. A provisional sunspot number is available within 24 hours from the Solar Influences Data Centre (SIDC) (*http://sidc.oma.be*).

Recent white-light disk images are available from, for example, Catania Astrophysical Observatory, Italy.

Hα/flare observations

The Hα line (wavelength = 656.3 nm) corresponds to the first atomic transition in the hydrogen Balmer series. This absorption line of neutral hydrogen falls in the red part of the visible spectrum, and on the Sun is partly formed in the solar chromosphere (~4300–30,000 K). As such, Hα images provide information about the chromosphere, including the fine structures of active regions and filaments. Hα observations also provide information about the location and nature of transient activity, including solar flares. (Several laboratories have on-line data: Hiraiso Observatory, Japan, current full disk Hα; Catania Astrophysical Observatory, Italy, current white-light and Hα full disk images (*http://web.ct.astro.it/sun/*).)

Radio observatories

Flare activity can also be detected in the radio range. A list of facilities intended to be operated routinely to monitor the solar radio activity at discrete frequencies is available on line (see *http://www.astro.phys.ethz.ch/rapp/cesra/sites_nf.html*).

Background radio observation/F10.7

The solar flux density from the entire solar disk at a frequency of 2,800 MHz (corresponding to 10.7-cm wavelength) has been recorded routinely by a radio telescope near Ottawa since February 1947, leading to the *10.7-cm solar flux index*. The atmosphere and ionosphere is transparent at this wavelength, but for the long-term consistency of the index the observed values have to be adjusted for the changing Sun–Earth distance (adjusted values) and for uncertainties in antenna gain (absolute values). Fluxes are given in units of $10^{-22}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$ (solar flux unit, SFU). Empirical formulae have been established for the relationship between F10.7 and the solar sunspot number. It is a useful indicator of day-to-day change of EUV flux which has an impact on the ionosphere and thermosphere. (Latest measurements are available at *http://www.drao.nrc.ca/icarus/www/current_flux.shtml.*)

Solar magnetograph

Magnetographs provide an indication of the magnetic field at the solar photosphere through use of the Zeeman effect. These images of the solar disk are termed magnetograms. Study of the evolution of the magnetic field structure within an active region, or surrounding a prominence, can provide insight into risk of eruption. Recent magnetograms, updated every 3 hours, are available from Learmonth Solar Observatory, Western Australia. Many research groups are working towards models which are able to predict eruptive phenomena based on magnetograms. (One example of an experimental model for predicting eruptive phenomena from magnetograms can be found at *http://www.dasop.obspm.fr/fromage/.*)

12.3.1.2 Neutron monitors

There are several tens of neutron monitors world-wide which monitor galactic cosmic rays and can detect solar particle events. Several observatories provide data in near-real-time. (For instance, the Moscow neutron monitor from Izmiran provides data on-line at *http://helios.izmiran.troitsk.ru/cosray/main.htm* for the last 30 days, corresponding to 10–20 GeV primary cosmic rays, and is updated every hour. More data sources are listed on their web site at *http://helios.izmiran.rssi.ru/cosray/sort.pl?showtable*.)

12.3.1.3 Ground magnetometers and geomagnetic activity indices

Magnetospheric activity, and especially storm and substorm-related charged particle injection and energization, produces enhanced currents in the ionosphere and the magnetosphere that can be monitored by ground-based magnetic observatories recording the horizontal magnetic field components. By combining measurements from well-chosen observatories, various indices can be produced to monitor substorm and storm development. Many of the magnetometer data are now available in real-time, especially via the INTERMAGNET network, while provisional and 'quick look' indices are also available. (INTERMAGNET data are available via web access nodes; for example, *http://obsmag.ipgp.jussieu.fr/intermagnet/*.)

A index

An A index is a linear measure of the disturbance level in the Earth's magnetic field defined over a period of 1 day. It can be defined for any location on Earth, and also for the entire globe. Levels of A index are often described as follows: A index less than 8 – 'quiet'; from 8 up to and including 15 – 'unsettled'; from 16 up to and including 24 – 'active'; from 25 up to and including 35 – 'minor storm'; from 36 upwards – 'major storm'. A very useful planetary index is the Ap index derived from a set of standard stations around the world.

The AE and related indices describe the disturbance level recorded by auroral zone magnetometers. Horizontal magnetic component recordings from a set of 10–14 globe-encircling stations are plotted to the same time and amplitude scales relative to their quiet-time levels, and then graphically superposed. The upper and lower envelopes of this superposition define the AU (amplitude upper) and the AL (amplitude lower) indices, respectively. The difference between the two envelopes determines the AE (Auroral Electrojet) index: AE = AU–AL. AO is defined as the average value of AU and AL. While the definitive AE index is available only after months of data processing and quality checking for scientific purposes, an indicative 'quick look AE index' is available on line (at the WDC-C2 homepage) with a 12-hour delay. Currently, the AE index has 1-minute resolution, but data exist, back to 1957, with 1-hour resolution (see *http://swdcdb.kugi.kyoto-u.ac.jp/aedir/ae1/quick.html*).

K index

A three-hourly index of geomagnetic activity relative to an assumed quiet-day model for the recording site. K index values range from 0 (very quiet) up to 9 (very disturbed). The global or planetary Kp index is obtained as the mean value of the disturbance levels observed at a dozen selected subauroral stations. Kp has been widely used in ionospheric and atmospheric studies since it was introduced in 1938, and is generally recognized as a good indicator for the effects of energetic charged particles arriving in Earth's upper atmosphere after periods of intense solar activity. The final Kp is available with about a 1-year time lag. There is also a preliminary Kp available with about a 1-day time lag from NOAA/SEC (*http://www.sel.noaa.gov/Data/index.html*).

Dst index

The Dst index monitors the variation of the globally symmetrical ring current, which encircles the Earth close to the magnetic equator in the Van Allen belt of the magnetosphere. It is derived from four equatorial magnetic stations. During large magnetic storms, the signature of the ring current can be seen, in ground magnetic field recordings worldwide, as so-called main phase depression. Dst responds differently than AE to changes of solar wind parameters during intense storms.

The unit of Dst is nT or γ (where $1\ \gamma = 1\,\text{nT}$). The typical range of Dst is from $+50\,\text{nT}$ to $-150\,\text{nT}$ (during the ring current energization), although these values can be largely exceeded. Therefore, the provisional Dst index, which is currently based on five observatories, is made available on a monthly basis with approximately 2 months delay, while the final Dst index is published annually with several months delay. However, a near-real-time Dst is available online from the WDC-C2 in Kyoto, Japan.

Polar cap index

The simplest index is the polar cap index. It is a near-real-time index derived from two antipodal stations. A provisional version is, in principle, available with about 24 hours delay, from DMI (*http://web.dmi.dk/fsweb/projects/wdcc1/master.html*).

12.3.1.4 Ionosondes and ionospheric radars

The bottom ionosphere can be monitored with the help of ionosondes and radars. Ionosondes especially provide the location and peak density value of the F layer, and radars provide maps of the ionospheric convection. Some ionosonde data are available in real-time. The SuperDARN network of high-frequency radars is being prepared to provide radar data that can be downloaded in real-time anywhere in the world (*http://superdarn.jhuapl.edu/rt/map/index.html*).

However, real-time ionospheric data which are very useful for transionospheric or ground-to-ground radio-wave propagation are of little interest to spacecraft environment modelling at the moment, since the magnetospheric and ionospheric models are not mature enough to trace high-energy particle dynamics from these observations. Significant modelling effort is required in this field.

12.3.1.5 TEC mapping

The total electron content (TEC) is the integrated number of electrons along a wave path measured in electrons per cm^{-2}. The TEC is a relevant parameter in determining delay and directional changes of a wave in the ionosphere. Several groups have developed techniques to derive maps of TEC from GPS dual-frequency measurements data (see *http://www.rcru.rl.ac.uk/live.htm*, Radio Communications Research Unit (RCRU) at Rutherford Appleton Laboratory, over Europe; and *http://iono.jpl.nasa.gov/latest.html*, over the US).

12.3.2 Space-based measurements

A certain number of space-based measurements are available in near-real-time. Some are performed by operational spacecraft and are part of a routine observational programme that includes provision for continuous spacecraft replacement. Others are the result of unique (usually science) missions and will, *a priori*, not be available after the nominal operational period.

12.3.2.1 GOES/SEM and energetic particle monitors

The NOAA Geostationary Operational Environmental Satellites (GOES) constitute a unique source of very valuable near-real-time solar and magnetospheric data recorded over decades at geostationary orbit. These spacecraft – which are the major elements of the overall US National Weather Service – carry, in addition to their main meteorological payload, a Space Environment Monitor (SEM) that has instruments for X-ray, energetic particle and magnetic field measurements.

The SEM consists of the four following instruments:

- The Energetic Particle Sensor (EPS), which detects electrons above 0.6 MeV, α particles and proton activity within specific lower-energy ranges (1–100 MeV in seven channels).
- The High Energy Proton and Alpha Detector (HEPAD), which detects protons and α particles in higher-energy ranges.
- The X-Ray Sensor, which measures solar X-ray flux in two bands (0.5–4 and 1–8 Ångströms).
- Magnetometers.

There are always two satellites present at any one time, and the longitude spacing between the satellites ranges from 30–60°, or 2–4 hours local time. The latest GOES data (1997 to the present) are available over the Internet from SEC. The data exist both in 1-minute and 5-minute resolution, and are also available in separate files for each day, about 1 month back in time. The latest two hours of data exist in a separate file, and the latest 5-minute data are generally 5–10 minutes old.

The solar X-ray flux is described as follows:

Level	Flux (watts/sq m)	Description
<A	Less than 10^{-8}	Very Low Background
A	Between 10^{-8} and 10^{-7}	Low Background
B	Between 10^{-7} and 10^{-6}	Moderate Background
C	Between 10^{-6} and 10^{-5}	High Background/Low Flare
M	Between 10^{-5} and 10^{-4}	Moderate Flare
X	Above 10^{-4}	High Flare

For a given decade, a letter is used to specify the flux. Hence a value M3.2 indicates that the flux is 3.2 $10^{-5}\,\mathrm{W/m^{-2}}$, and X1.2 would indicate $1.2 \times 10^{-4}\,\mathrm{Wm^{-2}}$.

GOES spacecraft now also provide X-ray imaging to allow more accurate detection of eruptive phenomena and impacts (*http://www.sec.noaa.gov/sxi/*).

Several other spacecraft contain instruments for energetic particle radiation monitoring. The ESA missions Chandra and Integral, for example, contain radiation monitors as a support for the science operations and data processing. Although the monitors are hosted on a case-by-case basis, the data can prove useful for environment modelling. Monitors of a similar nature also fly on Rosetta, Proba-1, SAC-C and GSTB. (Some of these data are available at *http://srem.web.psi.ch/html/srem_ home.shtml* and *http://plasma.oma.be/mrm/home.html.* XMM radiation monitor data are available in near real time at *http://xmm.vilspa.esa.es/external/xmm_obs_info/radmon/index.php*.)

12.3.2.2 LANL geosynchronous energetic particle data

The LANL geosynchronous energetic particle data derive from ten different US DoD satellites flown over the period 1976 to the present. Typically, data are available from three or four satellites simultaneously. The satellites operate at a geostationary orbit. Data are acquired in real time at Los Alamos, and are then processed, formatted and put on line every night. Therefore, digital data are typically available within 24 hours. The recent energetic particle data from the LANL satellites comes from the SOPA instrument in use on satellites launched from the beginning of 1989. It measures electrons from 50 keV to 26 MeV in sixteen channels, and protons from 50 keV to > 50 MeV in fifteen channels. It also measures heavier ions. (More information can be found at: *http://leadbelly.lanl.gov/lanl_ep_data/*.)

12.3.2.3 SOHO

The SOHO spacecraft is a joint ESA–NASA scientific mission. SOHO was launched in December 1995 and was placed into a halo orbit at the first Lagrangian point L1 (1.5 million km from Earth). The main objectives of the mission centre around study of the solar interior, heating of the solar corona, and acceleration of the solar wind. The SOHO payload consists of twelve instruments, dealing with helioseismology, solar atmosphere remote sensing, and solar wind *in situ* measurements. Many instruments onboard SOHO provide unique capabilities for imaging the solar disk in the EUV range and of the solor corona with onboard visual coronographs. Continuous observation by the SOHO coronographs have generated an unprecedented dataset on CMEs (*http://cdaw.gsfc.nasa.gov/CME_list*). Interesting new techniques have been also elaborated to infer UV flux or the presence of sunspots on the far side of the Sun (the anti-Earthward side). One of these techniques, based on the Lyman-α photons (121.6 nm) backscattered by the neutral hydrogen atoms present in the interplanetary medium and measured by the SWAN instrument, is being used for forecasting proxy of solar EUV flux. SWAN performs three full-sky observations per week (usually on Tuesdays, Thursdays and Saturdays). Images of the solar flux spatial distribution are displayed on-line 2–4 days after the observation.

Thanks to the US Deep Space Network, most SOHO data are available, with only a few hours' delay, via the SOHO website (*http://sohowww.estec.esa.nl/*).

12.3.2.4 ACE

Like SOHO, the NASA spacecraft ACE is located on a halo orbit at the first Lagrangian point L1, and is therefore permanently between the Sun and the Earth's magnetosphere at about 1.5 million km from the Earth. It is also a scientific spacecraft, however, and the data-link has been enhanced due to an NOAA contribution to provide continuously and nearly real-time data on solar wind protons, electrons, and interplanetary magnetic fields. It also provides data on solar flare particles and solar particle events. Data from the latest month with 1–5 minute resolution are available from the SEC web site. The latest data are generally $\sim$5 minutes old. ACE is a scientific spacecraft, and is not part of a continuous monitoring programme. Consequently, it is uncertain whether such data will be available following the end of the mission (see *http://www.sel.noaa.gov/ace/*).

12.3.2.5 NOAA POES SEM

The NOAA Polar-orbiting Operational Environmental Satellites (POES, formerly TIROS) are in Sun-synchronous orbit at 850 km altitude and 98° inclination. In the near future they will be part of an international constellation together with the Eumetsat METOP spacecraft. Like the GOES spacecraft, they carry space environment monitors besides their main meteorological payload. The Space Environment Monitor (SEM) contains two sets of instruments that monitor the energetic charged-particle environment near Earth. An upgraded SEM, called SEM-2, began operations with the launch of NOAA-15, and is currently the prime source of observations. The two instruments monitor the medium- and high-energy particle environment in the respective range 50 eV to 20 keV (the typical energy range in auroral precipitations), and 30 keV and 69 MeV (the typical energy range in radiation belts) respectively. These instruments have been used to derive two indices directly related to auroral activity and radiation belt activity.

Auroral activity index

Auroral measurements from more than 100,000 satellite passes have been used to construct statistical patterns of auroral power flux for ten levels of auroral activity corresponding to different total power dissipation. Thanks to this database, the total power deposited in an entire polar region by auroral particles can be estimated from the power flux observations obtained during a single pass of the satellite over a polar region (which takes about 25 minutes). The power input estimate is converted to an auroral activity index that ranges from 1 to 10 (see *http://www.sel.noaa.gov/pmap/*).

Radiation belts indices

Data from the POES high-energy particle sensors are processed to compute the integrated difference of individual sensor responses observed on a given day from the median responses of all sensors over the previous year's observations. This derivation provides an estimate of whether the current particle environment is more or less intense than usual. Such radiation belt indices are given for the whole belts, the inner belt, the slot and the outer belts.

The >30 keV electron detectors generally show elevated fluxes at geographic latitudes between 55° and 75° during periods of high auroral activity. They do not respond so strongly to auroral activity, but will observe unusually high fluxes in the outer radiation belts (geographic latitudes 45–60°) during periods of relativistic electron events, as also observed by geostationary orbiting satellites.

The proton detector telescopes (especially the 30–80 keV and the 80–250 keV energy channels), will measure very elevated fluxes over a region extending from the outer Van Allen radiation belt to the auroral zone (geographic latitudes from 45° to 75°) during global magnetic storms. The >16-MeV and >35-MeV proton omnidirectional detectors will display very high responses over the Earth's polar regions (geographic latitudes higher than 65°) during solar energetic particle events (see *http://www.sec.noaa.gov/tiger/*).

12.3.3 Near-real-time monitoring data: summary

As far as effects on spacecraft only are concerned, the ideal key parameters along the spacecraft orbit are indicated in Table 12.3. The corresponding measurements are indicated, together with their caveat.

12.3.4 Forecast, precursors and models

As already mentioned at the beginning of Section 3, there are potential applications in space engineering related to the ability to forecast the space environment conditions. It has been seen in the previous sections that most of the short-term and sporadic variation of the space environment is related to specific phenomena: solar flares, CMEs, SPEs, substorms and storms. There are known precursors that can be used to assess the probability of risk of certain level of flux of corpuscular environment to be exceeded. Examples of precursors are given in Table 12.4.

Use of precursors, however, is only indicative, since in general there is no one-to-one correlation or the information is missing at the time of occurrence of the expected event. In some cases it is possible to make use of predictive models. There are two types of model: first principle models and empirical models.

First principle models rely on known physical principles. There are models for nearly all regions of the solar–terrestrial environment. However, only a few of them

Table 12.3. Relevant near-real time data source for space environment effects.

Component	Available measurement	Caveat	Delay before availability
Galactic cosmic rays: 100 MeV to 1 GeV	Neutron monitor	Proxy	1 hour
Flare protons: 10 MeV to ~300 MeV	GOES	Require model for low altitude correction	5 min
	Neutron monitors	Proxy Only above ~400 MeV	1 hour
SPE particles: 100 keV to ~100 MeV	GOES ACE	Require model at low altitude	5 min
X, EUV, UV photons	GOES SOHO		5 min
Magnetopause boundary location	GOES	Only 2 points at GEO	5 min
	ACE	Input to model	5 min
Plasma: 0.1 to 100 eV	TEC	Integrated value	1 hour
	NPOES	Only at 830 km altitude	1 hour
Plasmasheet and auroral Electron energy spectrum (100 eV to 20 keV)	LANL	Only at GEO	1 day
	NPOES	Only at 830 km	1 hour
	Kp	Proxy	1 day
	A	Proxy	1 day
Radiation belts proton energy spectrum 1 MeV to 1 GeV	GOES	Only at GEO Reduced energy range	5 min
	NPOES	Only at 830 km	1 hour
Radiation belts electrons energy spectrum 100 keV to 10 MeV	GOES	Only at GEO	5 min
	LANL	Only at GEO	24 hours
	NPOES	Only at 830 km	1 hour
	Kp	Proxy	1 day
	A	Proxy	1 day
Thermospheric flux	F10.7	Input to model	Once a day
	Kp		1 day

are run in an operational context or accessible on-line. Empirical models are also much faster and somehow more reliable, as limitations are clearly identifiable. A particular class of model is based on neural network techniques. There have been several discussions on the relevance of such models. It must be noted that they are merely non-linear filters, and may potentially apply to a broader range of phenomena than the more simple linear models.

12.3.5 Services

Besides the original data-providers, various organizations have developed support service capabilities for customers (paying or not). The type of service ranges from

Table 12.4. Examples of precursors of sporadic phenomena.

Phenomenon	Tracer	Precursor
Flare	X, UV, Vis, MeV protons	Sunspot and less intense magnetic field regions Magnetic structure
CME	Vis image	Sunspot
ICME	Radio Interplanetary scintillation	CME
SPE	MeV protons	fast CME
Substorm	AE, Kp NPOES	
Storm	Kp GOES LANL	IMF Bz <0, CME
Ionospheric and thermospheric change	TEC Radars	EUV from solar backside IMF Bz <0

Table 12.5. Main international service providers for space environment data.

Name	Type of service	Web site
ISES	International Space Environment Service	*http://ises-spaceweather.org/*
NOAA/SEC	ISES node	*http://www.sec.noaa.gov/*
IPS	ISES node	*http://www.ips.oz.au/*
NICT	ISES node	*http://hiraiso.nict.go.jp/*
IMO	Meteor shower information	*http://www.imo.net/calendar/cal00.html*
ISGI	Access and description of all IAGA indices	*http://www.cetp.ipsl.fr/~isgi/homepag1.htm*
WDC-C2	World Data Centre C2 for Geomagnetism Definitive Dst and AE, including quick-look values	*http://swdcdb.kugi.kyoto-u.ac.jp/*
SWENET	Space Weather European Network of Services	*http://esa-spaceweather.net/*

provision of space environment data to model output and tailored services for specific customers. The largest service provider is by far the NOAA/SEC, where links to most of the real-time data available world-wide can be found due to collaborative agreements with its International Space Environment Service (ISES) partners. Other large international service providers among ISES members are IPS (Australia) and NICT (Japan). In Europe the development of services is growing rapidly, and several organizations have become members of ISES. Many of these services have joined the Space Weather European Network, SWENET. A non-exhaustive list of the main international service providers of general interest is given in Table 12.5.

12.4 THE FUTURE

It has been shown in the previous section that a large amount of continuous measurements of the space environment are available, with a short delay, on-line. Many of these data are very useful for assessing the space environment conditions in near-real-time or for quickly retrieving the conditions that prevailed at a time of special interest; for instance, whenever a spacecraft anomaly occurs or for the quality check of scientific data. Furthermore, the quality and size of some of these datasets have allowed the contruction of new models that take into account the full variability of the environment at various timescales (from hours to decades). However, the current situation has some major limitations, and improvements should be foreseen regarding the data coverage, the continuity and the reliability of critical data provision and the accuracy of models for forecasting or interpolating. These issues are discussed further in this section.

12.4.1 Data coverage

Although many space environment datasets are available world-wide, there is a striking lack of data covering some regions of space which are critical regarding their effects on spacecraft. For instance, the spatial scale of the gradient of the properties of low- or medium-energy particles environment is often very small compared to a spacecraft orbit length. Therefore, observations performed, for example, on another spacecraft, even on a similar orbit, are too ambiguous. In such cases, if no accurate models exist to extrapolate the environment from observations elsewhere in space or earlier in time, the deployment of onboard monitors for the local, or orbit specific, details is often the only solution. Examples are as follows:

- *Electron environment in keV energy range for high-altitude orbit*. Regions with keV electrons are rather narrow in the space environment. For instance, the extension of the auroral arcs in a direction perpendicular to the magnetic field is not more than a few hundred km. Also, keV electron injection at geostationary orbit is rather localized. Information on the location of aurorae could be retrieved from ground-based all-sky cameras under clear sky conditions, from space-based auroral imagers, or from onboard hitch-hikers in this energy range. In the two former cases, an accurate magnetic field model is required to trace the auroral arc location up to spacecraft altitude. At geostationary orbit, hitch-hikers are needed.
- *Electron environment in 1 eV energy range for high-altitude orbit*. The plasmasphere – the region of cold plasma – is very dynamic at high altitudes, and it can expand beyond geosynchronous orbit. It is useful information for spacecraft charging effects, as when a spacecraft is in the plasmasphere one does not expect high-level surface charging. Models of the plasmasphere and/or hitch-hikers could be used in this energy range. Hitch-hikers would work best on electrostatically clean spacecraft.

- *Proton and electron environment in the range between 100 keV and 1 MeV at very high altitudes.* Injection of >100-keV electrons and protons at high altitudes is also a rather localized phenomena, although they rapidly drift around the Earth, where they can be detected by other spacecraft at geostationary orbit. Bursts of particles in this energy range and above are also generated in the solar wind, and may be a concern for spacecraft with orbits at very high altitude and with very sensitive devices such as the X-ray observatories Chandra and XMM/Newton. Such particles are not detectable with the GOES or NPOES spacecraft. In the absence of accurate models for such phenomena, the best option is the use of hitch-hikers in this energy range.
- *Thermospheric flux.* Although direct measurement of the thermospheric flux is locally possible with the help of incoherent scatter radars or interferometers, they allow only local measurements, and under certain conditions. Better knowledge of this environment could be provided with the use of systematic tracking of spacecraft and accelerometers.

12.4.2 Long-term continuity of data provision

It has been shown in the previous section that already a large amount of continuous measurements of the space environment are available on-line with a short delay. Many of these data are very useful for assessing the space environment effects on space hardware in near real-time or for quickly retrieving the conditions that prevailed at a time of special interest (e.g., whenever a spacecraft anomaly occurred). Furthermore, the quality and size of some of these data sets have allowed the contsruction of new models that take into account the full variability of the environment at various timescales (from hours to decades). However, the current situation has some major limitations and improvements should be foreseen if operational services were to be developed. These issues are discussed further in this section.

12.4.3 Reliability of data provision

For data that may be used as part of operational procedure the reliability of the datalink is crucial. However, there is very little redundancy in the current space data collection and distribution architecture. The main source of data is currently the web site maintained at the Space Environment Centre of NOAA. Thanks to the GOES and NPOES programmes there is a guarantee of continuity of the measurements performed by these spacecraft. For instance, the successor of the GOES spacecraft currently in operation is already in place. Also, some data provided by scientific spacecraft have no requirements on real-time availability or of continuity, and may therefore be suddenly interrupted without notice. Data from ground observatories are not always distributed in real-time or made publicly available because of their often old equipment, although slow improvements performed over recent years can be noticed.

Improvements in data provision reliability could be obtained by more redundancy of the data provision architecture from onboard acquisition to ground network distribution. Companies or industries needing data for operational procedures may find it useful to have tailored data distribution service or guaranted timely data provision. Such enhanced service is possible on a commercial basis only.

12.4.4 Model accuracy

Current models are often too inaccurate for operational purposes. Physics-based models require more testing, and the processes taken into account are often not sufficient to describe the full dynamics at work in the solar–terrestrial environment. Models are often developed for specific regions corresponding either to specific scientific expertise or because particular approximations hold. Therefore, the interoperability of the models is often problematic. In most cases, empirical models perform better than physics-based models, and they are already in use for forecasting indices or for thermospheric and ionospheric models. However, their response is normally optimized for the most common environments, because they are the models with maximum statistical weight. There is therefore a strong risk that they would not accurately predict the stronger events. There is clearly a need for improving the modelling activities by favouring coordination and interoperability in the development of models and by increasing the range of data on which they can be validated. Also, because of their higher reliability and computing speed, models based on non-linear filters are often chosen for operational purpose; but their validation must ultimately rely on well-accepted first principle-based models.

12.5 CONCLUSION

There are many space weather-related effects on spacecraft hardware and operations involving environments as diverse as the thermosphere, the ionosphere, the magnetosphere, and solar and cosmic-ray particles. Many of them have adverse effects, and problems are increasing because technological advances are susceptible to disruption from space weather and its disturbances (solar storms, radiation belt increases, geomagnetic perturbations, atmospheric changes, and so on). In addition, the expanding population of astronauts and future plans related to manned missions to the Moon or Mars increase the importance of the effects of cosmic radiation on humans. Advances in solar–terrestrial physics and information technology together make possible the provision of services to quantify, detect and predict hazardous conditions which would have been inconceivable a decade ago.

As seen above, routine space-based observations of the space environment have been performed for more than 30 years by US operational space programmes. The NOAA/SEC service, partly sponsored by the Department of Defense, is providing data and services free of charge to a very large community of users world-wide. Nowadays the study of space weather activity in the US is growing rapidly and being reorganized in the light of future strategic and commercial opportunities in the form of an interagency programme. Opportunities for similar initiatives are also being discussed in Japan and in China. International coordination of research and scientific missions which will seek to further develop the scientific understanding of space weather is currently underway within the framework of the International Living with a Star Programme.

For many years, Europe has been at the forefront of the development of the understanding of the physics of the Sun, the production of energetic solar events, their propagation through the interplanetary medium, and the consequent responses of the magnetosphere, ionosphere and atmosphere to the events. The SOHO, Ulysses and Cluster projects have been key space-based observatories in exploring this highly complex Sun–Earth connection. The excellence they represent should, in the future, extend to Solar Orbiter and SWARM. European technical groups have been actively supporting European programmes in coping with the hazards caused by space weather on satellites and on ground systems. However, the data acquisition segment has some significant gaps and weaknesses, including the lack of reliability and continuity of data and the fact that most modelling activities are disconnected from the engineering world. The recent ESA initiatives (Daly and Hilgers, 2001, and Glover *et al.*, 2003) – including the setting up of the Space Weather Working Team, and the network of services, SWENET, together with the COST 724 and COST 296 actions from European Science Foundation – are anticipated to be instrumental in improving the data provision infrastructure and in fostering the links between the scientific and engineering communities.

12.6 ACKNOWLEDGEMENTS

A large amount of useful information has been extracted from the Internet – especially from the NOAA Space Environment Center web server of National Oceanic and Atmospheric Administration (NOAA), US Department of Commerce, and data from the studies performed by ALCATEL Space industry and Rutherford Appleton laboratories funded under ESA ESTEC contracts 14069/99/NL-SB and 14070/99/NL-SB.

12.7 REFERENCES

Daly, E. J. and A. Hilgers, Space Weather: European Space Agency Perspectives, in *Geophysical Monograph 125*, American Geophysical Union, Washington DC, USA, pp. 53–58, 2001.

European Cooperation for Space Standards, ‘Space Environment’, ECSS-E-10-04, January 2000, (*http://www.ecss.nl/*).

Glover, A., A. Hilgers, E. Daly, and R. Marsden, Tomorrow’s space weather forecast?, *ESA Bulletin 114*, ISSN 0376-4265, pp. 28–37, May 2003.

Hilgers, A., P. Gondoin, P. Nieminen, and H. Evans, Prediction of plasma sheet electrons effects on X-ray mirror missions, pp. 293–296, ESA-WPP-155, ESTEC, Noordwijk, The Netherlands, March 1999.

Rosenqvist, L. and A. Hilgers, Sensitivity of a statistical solar proton fluence model to the size of the event data set, *Geophys. Res. Letters*, **30**, No 16, 1865, 2003.

Web resources

ACE	http://www.sel.noaa.gov/ace/
BGS	http://www.geomag.bgs.ac.uk/gifs/on_line_gifs.html#UK
Catania Observatory	http://web.ct.astro.it/sun/
CESRA	http://www.astro.phys.ethz.ch/rapp/cesra/sites_nf.html
CLRC-RCRU	http://www.rcru.rl.ac.uk/live.htm
CME list	http://cdaw.gsfc.nasa.gov/CME_list/
COST 724	http://cost724.obs.ujf-grenoble.fr/
COST 296	http://www.cost296.rl.ac.uk/
NICT-Hiraiso	http://hiraiso.nict.go.jp/
CLS	http://www.cls.fr/previsol/
DASOP	http://www.dasop.obspm.fr/previ/
DMI	http://web.dmi.dk/
DMI PCN	http://web.dmi.dk/fsweb/projects/wdcc1/master.html
ECSS	http://www.ecss.nl/
ESA Space Weather Initiatives	http://esa-spaceweather.net
ESA SREM	http://srem.web.psi.ch/html/srem_home.shtml and http://plasma.oma.be/mrm/
ESA XMM Radiation Monitor	http://xmm.vilspa.esa.es/external/xmm_obs_info/radmon/index.php
GOES SXI data	http://www.sec.noaa.gov/sxi/
FROMAGE	http://www.dasop.obspm.fr/fromage/
GFZ	http://www.gfz-potsdam.de/pb2/pb23/Geomag/niemegk/kp_index/
Glossary	http://www.sec.noaa.gov/info/glossary.html
Hiraiso observatory	http://sunbase.nict.go.jp/solar/denpa/realtime.html
IMO	http://www.imo.net/calendar/cal00.html
Intermagnet	http://obsmag.ipgp.jussieu.fr/intermagnet/
IPS	http://www.ips.oz.au/
IRF-Lund	http://www.irfl.lu.se/
ISES	http://ises-spaceweather.org/
ISGI	http://www.cetp.ipsl.fr/~isgi/homepag1.htm
IZMIRAN	http://helios.izmiran.troitsk.ru/cosray/main.htm
LANL	http://leadbelly.lanl.gov/lanl_ep_data/
OULU	http://cosmicrays.oulu.fi/
Penticton	http://www.drao.nrc.ca/icarus/www/current_flux.shtml
NOAA/SEC	http://www.sec.noaa.gov/
NOAA/SEC all data	http://www.sel.noaa.gov/Data/index.html
NOAA/SEC auroral activity index	http://www.sel.noaa.gov/pmap/
NOAA/SEC Today space weather	http://www.sec.noaa.gov/today.html

SAAPS	http://www.irfl.lu.se/saaps/
SIDC	http://sidc.oma.be/index.php3
SOHO	http://sohowww.estec.esa.nl/
Superdarn	http://superdarn.jhuapl.edu/rt/map/index.html
SWENET	http://esa-spaceweather.net/swenet/
WDC-C2	http://swdcdb.kugi.kyoto-u.ac.jp/

13

Effects on satellite navigation

Bertram Arbesser-Rastburg and Norbert Jakowski

Electromagnetic radio waves transmitted from satellites of Global Navigation Satellite Systems (GNSS) such as GPS, GLONASS, or Galileo interact with the ionospheric plasma on their path to the receiver, changing their speed and direction of travel. At the commonly used L-band frequencies (between 1.2 and 1.6 GHz), the ionosphere causes signal delays that correspond to range errors of up to 100 m. In a first-order approximation the range error is proportional to the integral of the electron density along the ray path (total electron content, TEC) and can be mitigated by a simple linear combination of the navigation signals at two different frequencies. However, horizontal gradients in the ionospheric ionization and ionospheric irregularities producing phase fluctuations and radio scintillations may significantly degrade navigation signals up to loss of lock. Strong TEC gradients, travelling ionospheric disturbances (TIDs), and phase fluctuations are often correlated with space weather phenomena such as geomagnetic/ionospheric storms.

Depending on the user requirements, the current GNSS technology takes into account space weather effects by developing proper mitigation techniques. Thus, space weather effects can be reduced but cannot be completely removed. Uncertainties remain in particular in the case of strong and highly dynamic ionospheric plasma density irregularities. Therefore, to fulfil growing safety requirements in GNSS applications, the impact of the ionospheric plasma on GNSS applications has to be studied in more detail. Fortunately, the GNSS technique itself provides a unique opportunity for high-resolution monitoring of the ionosphere on global scale.

13.1 INTRODUCTION

Principally, space weather has a measurable impact on radio signals up to frequencies of about 10 GHz. In GNSSs the interaction of trans-ionospheric radio waves with the plasma causes a propagation delay indicating a travel distance larger than the real one. Due to the inertia of the free electrons this propagation error decreases with increasing radio frequency ($\sim 1/f^2$), called ionospheric dispersion. Although direct damaging of navigation satellites by energetic solar wind particles cannot be excluded, the dispersive interaction of the navigation signal with the ionospheric plasma is the main reason for the vulnerability of space-based radio systems working at L-band frequencies.

The currently operating American Global Positioning System (GPS) and the Russian Global Navigation Satellite System (GLONASS) were originally designed for defence purposes. However, a rapidly growing number of civilian users are taking advantage of the freely available signals in space for a wide range of applications in navigation, surveying, tracking, etc., opening a broad market for developers of receivers and applications. Terrestrial overlay systems are being put in place to augment the navigation signals so that they can be used with higher precision or to provide integrity to life-critical operations such as landing a commercial aircraft. Europe, in partnership with a number of countries around the globe, is currently building the civilian navigation satellite system Galileo, which will offer extended services.

Developed by ESA in collaboration with the European Union and co-funded by the two organisations on a 50–50 basis, Galileo is a complete civil system, designed to provide accurate, secure, and certified positioning services to users all over the world. Amongst the many potential applications are guidance and navigation services for road, rail, air, and maritime traffic, location-based services in the transport sector, clock synchronisation for mobile networks, precise surveying and geodesy, as well as a whole range of scientific applications. The Galileo Space Segment is formed by a constellation of 30 satellites in a Walker configuration of 27 satellites in three orbital planes, plus one spare per orbital plane. All the satellites have circular orbits with nominal semi-major axis of 29,993.7 km and nominal inclination of 56°. The orbital planes are equally (120° apart) spaced.

13.2 SATELLITE-BASED NAVIGATION TECHNIQUE

Satellite navigation relies on precise measurement of the time it takes for a signal to propagate from a spacecraft to the antenna of the navigation receiver. If there is no refractive medium on the path, the location can be determined by calculating the distance to a number of visible satellites by multiplying the time observed by the vacuum speed of light. With the position and clock offsets of the satellites precisely known, the user of the navigation receiver can easily establish his own position in a given coordinate system [1].

In reality there are some additional propagation delays, which have to be taken into account. These delays are caused by the interaction of the electromagnetic wave field with the charged particles of the ionospheric plasma, the neutral but polarized particles in the troposphere, and obstacles. The resulting observation equation can be written for the code and the carrier phase Φ and ϕ, respectively as:

$$\Phi = \rho + c(dt - dT) + d_I + d_T + \varepsilon_P \tag{13.1}$$

$$\phi = \rho + c(dt - dT) - d_I + d_T + N\lambda + \varepsilon_L \tag{13.2}$$

where ρ is the true geometrical range between GPS satellite and receiver, c is the vacuum speed of light, dt and dT are the satellite and receiver clock errors, d_I is the ionospheric delay along the ray path s, d_T is the delay caused by the troposphere, λ is the wavelength of the corresponding carrier frequency, N is the phase ambiguity number (integer), and ε_P and ε_L are the code and carrier phase residual errors, respectively. The space weather sensitive ionospheric propagation term d_I is a function of the refraction index and related effects such as diffraction and scattering.

In a first-order approximation of the refraction index, the ionospheric delay is proportional to the integral of the electron density along the ray path according to:

$$d_I = \frac{K}{f^2} \int_{\mathrm{S}}^{\mathrm{R}} n_e \, ds \tag{13.3}$$

with $K = 40.3\,\mathrm{m}^3\,\mathrm{s}^{-2}$.

Here the integral of the local electron density n_e along the ray path between satellite S and receiver R is the TEC. The frequency dependency characterizes the ionosphere as a dispersive medium as already mentioned above.

Due to this dispersion the differential code phase $\Delta\Phi = \Phi_2 - \Phi_1$ measured at the two frequencies f_1 and f_2 reveals a first-order estimation of the TEC:

$$\mathrm{TEC} = \frac{\Delta\Phi \cdot f_1^2 f_2^2}{K \cdot (f_1^2 - f_2^2)} + \mathrm{TEC_{cal}} \tag{13.4}$$

Besides random errors the calibration term $\mathrm{TEC_{cal}}$ includes specific satellite and receiver code phase delays which do not cancel out in $\Delta\Phi$ according to (13.1).

It becomes clear that dual-frequency measurements allow the mitigation of the ionospheric term d_I at the navigation frequency L_1 in (13.1) as will be outlined in Section 13.5. Furthermore, the determination of TEC enables the monitoring of ionospheric key parameters by GNSS techniques as discussed in more detail in Section 13.3. Typical vertical TEC values range from a few TEC units ($1\,\mathrm{TECU} = 10^{16}$ electrons m^{-2}) for night-time conditions at high latitudes to well above 200 TECU at equatorial latitudes and during ionospheric storm conditions.

At the L_1 GPS frequency of 1.575 GHz, 1 TECU is equivalent to a path length increase of 0.162 m indicating maximum ionospheric vertical delays of up to about 30 m under heavily perturbed conditions. The delay may even increase by a factor of more than 2 at low elevation angles (i.e. when the path length through the ionosphere increases considerably). Equation (13.3) does not include the higher order terms of the refractive index. As illustrated in Figure 13.1, these effects are usually less than 0.1% of the first-order range error and therefore practically ignored in navigation. However, precise positioning and navigation services have to consider these terms [2].

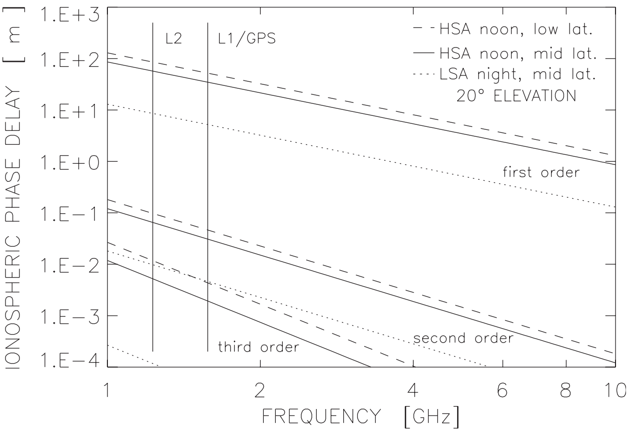


Figure 13.1. Frequency dependence of ionospheric range errors for different ionospheric conditions at 20° elevation (HSA, high solar activity; LSA, low solar activity).

To achieve high-precision positioning with Galileo, propagation errors have to be mitigated as accurately as possible. Therefore, mitigation of the ionosphere-induced propagation errors is a challenging task in high-precision satellite positioning and navigation (near-real-time positioning with an accuracy of better than 10 cm). In the last decade several methods and techniques have been developed to achieve this goal with Galileo.

Galileo can benefit from four well-separated frequencies in the L-band, two of them are compatible with GPS.

The first-order ionospheric range error (IRE) typically varies from 1–100 m at the Galileo frequencies. Using an ionospheric model, single-frequency users can reduce these effects by approximately 50%. To achieve a higher accuracy a more sophisticated ionospheric correction model is required.

Due to the dispersive ionosphere the availability of a third (or more) carrier frequencies essentially improves the performance of a GNSS. In the case of a third frequency, the simplest approach is the linear combination of three different frequency pairs for computing ionosphere-free solutions.

13.3 USE OF GNSS TECHNIQUES FOR SPACE WEATHER MONITORING

13.3.1 Ground-based monitoring

As outlined in the previous section, dual-frequency GNSS measurements enable the estimation of the TEC along the observed ray paths providing valuable information on the ionospheric state (e.g., [3]).

To obtain absolute TEC values, the ionospheric travel time delay or range error

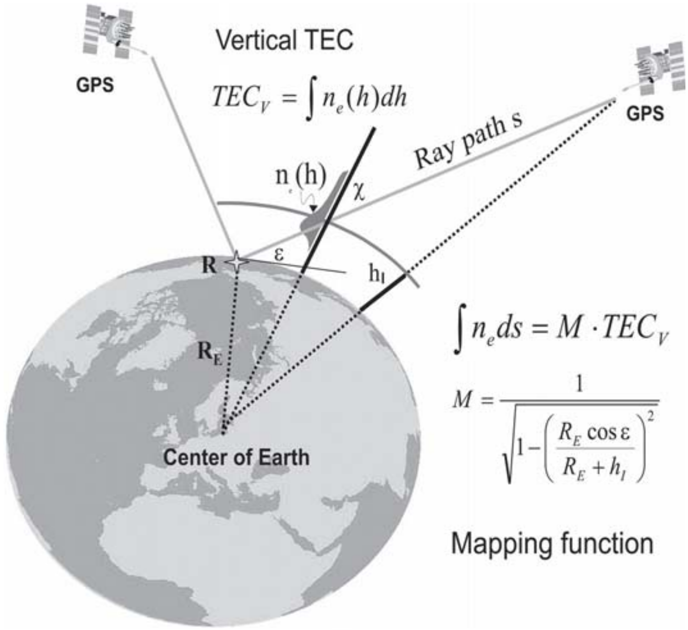


Figure 13.2. Conversion of slant TEC measurements into vertical TEC and vice versa. To obtain vertical TEC for reference, the ionosphere is assumed to be compressed in a thin single layer at height h_I. The location of the piercing point of the radio link with the ionospheric layer is considered to be the 'point' of measurement.

has to be calculated from differential code phases given in (13.1). Since the link related travel times are biased by the instrumental delays at the satellite and at the receiver, the derived TEC data must be calibrated (e.g., by a Kálmán filter technique [4]. Because carrier phase measurements are much less noisier than code measurements, the TEC retrieval can be significantly improved by levelling the phase into the code measurements simply applying least squares methods (e.g., [5]).

Assuming a single-layer ionospheric shell at a fixed height of about 350-400 km (single-layer approximation, see Figure 13.2), the slant TEC values are mapped to the vertical at the piercing points by the geometrical mapping function. Vice versa, this function is also used to compute the slant ionospheric propagation error if grid-based vertical TEC information is provided.

For ionospheric imaging the availability of a sufficient number of measurements is required. Different approaches are reported in the literature by different research groups (e.g., [5], [6], [7], [8]).

TEC maps are generated in DLR Neustrelitz routinely since 1995. After the measured and calibrated slant TEC data are mapped onto the vertical by applying a single-layer approximation at $h_I = 400$ km as shown in Figure 13.2, the resulting vertical TEC data are assimilated into a regional TEC model, NTCM [9]. This technique, even in the case of a low number of measurements, still delivers reasonable ionospheric corrections for enhancing the accuracy and integrity of positioning.

To give an example, Figure 13.3 (see colour section) illustrates the TEC level and the data coverage over Europe during a few hours on 10 January 1997, a day which is characterized by large space weather effects (cf. [10]).

The worldwide existing large database, containing data representative for a broad spectrum of solar and geomagnetic conditions, is an optimal background for the development and validation of ionospheric models and for studying the ionospheric impact on GNSS applications. The GNSS technique has the advantage to work robustly even under conditions where other measurement techniques, such as ground-based vertical incidence sounding are not reliable. Therefore, there are permanent efforts for improving ground-based GNSS monitoring and reconstruction techniques (e.g., [12]).

13.3.2 Space-based monitoring

Whereas ground-based measurements of propagation effects, such as travel time delays and phase changes, have been well established since the mid-1990s, space-based GNSS measurements on board low-Earth orbiting (LEO) satellites are rather new. The proof-of-concept GPS/MET experiment on Microlab-1, flown within the years 1995–1997, has demonstrated the huge potential of the limb-sounding technique on LEO satellites for atmosphere/ionosphere sounding (e.g., [13]). Subsequent satellite missions such as Oerstedt, CHAMP, and SAC-C have confirmed the high potential of GNSS radio occultation measurements for sounding the ionosphere.

The radio occultation technique provides a rather simple and inexpensive tool for a global profiling of the entire vertical electron density structure from satellite orbit heights down to the bottom of the ionosphere, not achieved so far by any other technique (Figure 13.4, right panel, see colour section). The reception of multi-satellite navigation signals, affected on their travel through the ionosphere, provides integral key information on the ionospheric state. Modern inversion and data assimilation methods allow a reliable reconstruction of the electron density structure if the amount of data is sufficient. Extensive information provided by current and future satellite missions with a GNSS receiver on board enables permanent monitoring of the Earth's co-rotating plasma environment in near-real-time (Figure 13.4, left panel). The plasmasphere, from about 1,000 km upward, is the innermost region of the Earth's magnetosphere and co-rotates with Earth.

The obtained global data sets contribute to a comprehensive understanding of solar–terrestrial relationships, and are valuable for developing and improving global ionospheric models. The space-based GPS technique is well-suited for providing operational space weather information on a routine basis. DLR has established a permanent space weather service under *http://www.kn.nz.dlr.de/swaci*, which provides both ground- as well as space-based GPS measurements and corresponding ionospheric data products. Consequently, accuracy and reliability of space-based com/nav radio systems will take benefit from this monitoring.

Generally speaking, combined ground- and spaced-based GNSS measurements can essentially contribute to improve our understanding of ionospheric perturbations which are closely related to space weather phenomena from solar, magnetospheric, and thermospheric origins (e.g., [16]).

13.4 SPACE WEATHER IMPACT ON THE SIGNAL PROPAGATION MEDIUM

Space weather, which describes the complexity of space environment conditions, can adversely affect the signal propagation of radio systems at frequencies below 5 GHz. Since the IRE in GNSS applications is proportional to the TEC of the ionosphere, the range error is principally subjected to all changes of the ionospheric ionization (i.e., strongly dependent on space weather conditions).

13.4.1 Solar control of ionospheric ionization

Similar to the terrestrial weather at the Earth's surface, the ionosphere is also mainly controlled by the intensity of solar radiation and the geometry of its incidence. Ionization is essentially produced at wavelengths shorter than 130 nm which is in the extreme-ultraviolet (EUV) range. Because the so-called solar activity is characterized by an enhanced energy transmission especially at shorter wavelengths, the ionospheric range error d_I is also a sensitive function of the solar activity shown in Figure 13.5. It is interesting to note that the ionospheric response to solar irradiation changes is delayed by about 1–2 days [17]). Even solar flare effects are visible in the TEC (e.g., [18]).

This close coupling of TEC to the solar EUV radiation can directly be seen during a solar eclipse event when the solar radiation is switched off and on according to the occultation function (e.g., [19]).

It is obvious that the photo-ionization of the Earth's atmosphere due to solar radiation depends on the incidence angle of the irradiation. Thus, the diurnal, seasonal (Figure 13.5), and latitudinal variation (Figure 13.6) of the total ionization can easily be explained. However, some anomalies such as the equatorial or seasonal anomaly indicate more complex solar–terrestrial relationships which are far from being fully understood nowadays (Figure 13.6).

Whereas the seasonal dependency shows the importance of the atmospheric composition in the thermosphere (e.g., [20]), the equatorial anomaly demonstrates the action of the electric fields (e.g., [21]). Thus, it becomes clear that space weather events accompanied by enhanced energy input due to rapid solar radiation bursts or by enhanced particle precipitation can cause a broad variety of plasma perturbations at quite different scales in time and space.

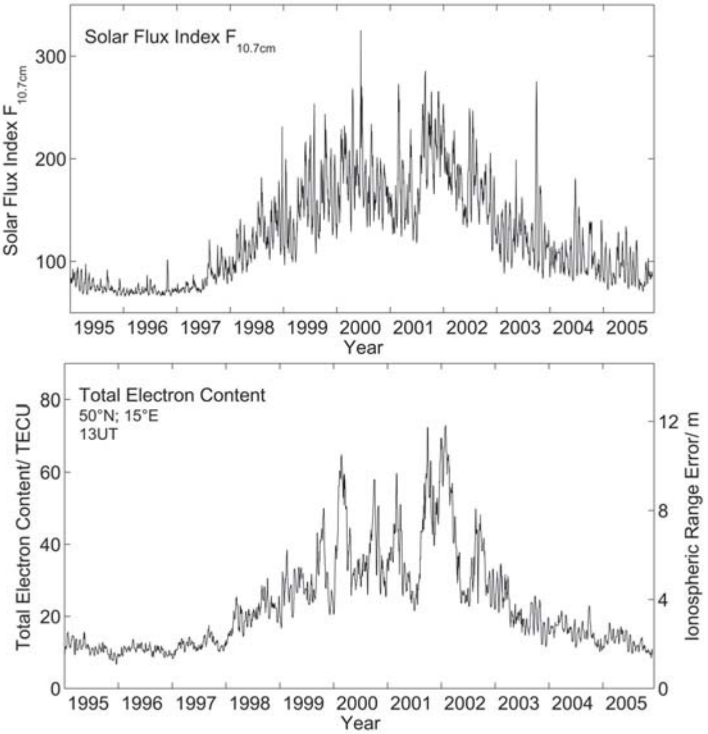


Figure 13.5. Compared are 7-day averaged TEC values at 50°N, 15°E at 2:00 pm local time (bottom panel) with the solar radio flux index F10.7 (upper panel), showing a close correlation with the solar flux but also significant seasonal effects.

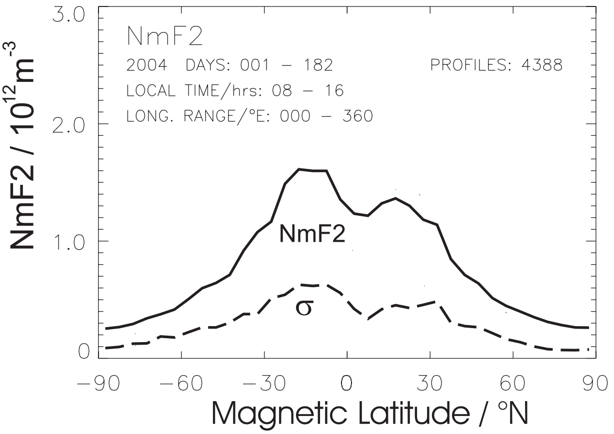


Figure 13.6. Latitudinal variation of the peak electron density as derived from CHAMP data in the second half of 2004 at day-time conditions (08:00-16:00 local time) indicating a strong latitudinal dependency of the ionization and the equatorial anomaly as well. The dashed line shows the standard deviation.

13.4.2 Ionospheric storms

Space weather-induced disturbances in the ionosphere are characterized by strong variations in the vertical and horizontal electron density distribution. As discussed in the previous section, ground-based GNSS measurements can effectively be used to estimate the horizontal distribution of TEC on regional and global scales.

The big space weather event on 6 April 2000 was monitored by quite different techniques at different spheres from the Sun down to the Earth's surface, demonstrating the broad spectrum of space weather effects observable in different spheres of the Earth's environment.

The highly variable temporal and spatial structures in the ionospheric TEC maps are probably due to particle precipitation through the cusp region and strong plasma convection across the pole due to perturbation electric fields. The strong impact on the magnetospheric/ionospheric systems in the evening hours of 6 April is closely related to the southward direction of the interplanetary magnetic field ($B_z < 0$) that has been measured from about 17:00 until 24:00 UT ([22], [23]).

As Figure 13.7 clearly shows, the polar ionization is strongly irregular between 17:00 and 24:00 UT. The amplitude of TEC can reach extremely high values of up to $60 \times 10^{16}\,\mathrm{m}^{-2}$ (60 TECU) that has a severe impact on accuracy and reliability of GNSS navigation. Strong phase fluctuations were observed at mid-latitude European GPS stations in particular around 23:00 UT [24], indicating the existence of significant horizontal structures.

Although the time development of ionospheric storms differs considerably, depending on a large variety of solar–terrestrial conditions, they also show some typical features [25].

Perturbation induced changes of TEC are more pronounced in so-called differential maps (percentage deviation $\Delta\mathrm{TEC} = (\mathrm{TEC} - \mathrm{TEC}_{\mathrm{med}})/\mathrm{TEC}_{\mathrm{med}} \times 100$) which are quite useful for studying general features of ionospheric storms. Figure 13.8 (see colour section) represents the average storm pattern of TEC along the 15°E meridian from 32.5°N up to 70°N as a function of a specially fixed storm timescale over 5 days. The storm time ST = 00:00 is defined by the temporal geomagnetic Dst index gradient at the storm onset phase where $\Delta Dst/\Delta t$ exceeds $10\,\mathrm{nT\,h}^{-1}$ (see Figure 13.8, upper panel). The pre-storm day was restricted to be quiet within a Dst interval of $\pm 15\,\mathrm{nT}$. These selection criteria select storms which are characterized by a sharp onset. Thus, the arithmetic average of superposition of percentage TEC storm patterns of 10 severe storms ($\mathrm{Dst} < -40\,\mathrm{nT}$ during the main phase) occurring from 1995 until the end of 2004 is presented in Figure 13.8.

The observed simultaneous increase of differential TEC at different latitudes near 00:00 ST (Figure 13.8, lower panel) indicates the action of an eastward directed electric field which penetrates from high to at least mid-latitudes at the storm onset. The delayed response of the storm pattern toward lower latitudes observable a few hours later is a clear indication of the action of storm induced equator-ward blowing neutral winds. Both effects were also seen in the case study from 10 January 1997 [10]. These main driving forces originate from the

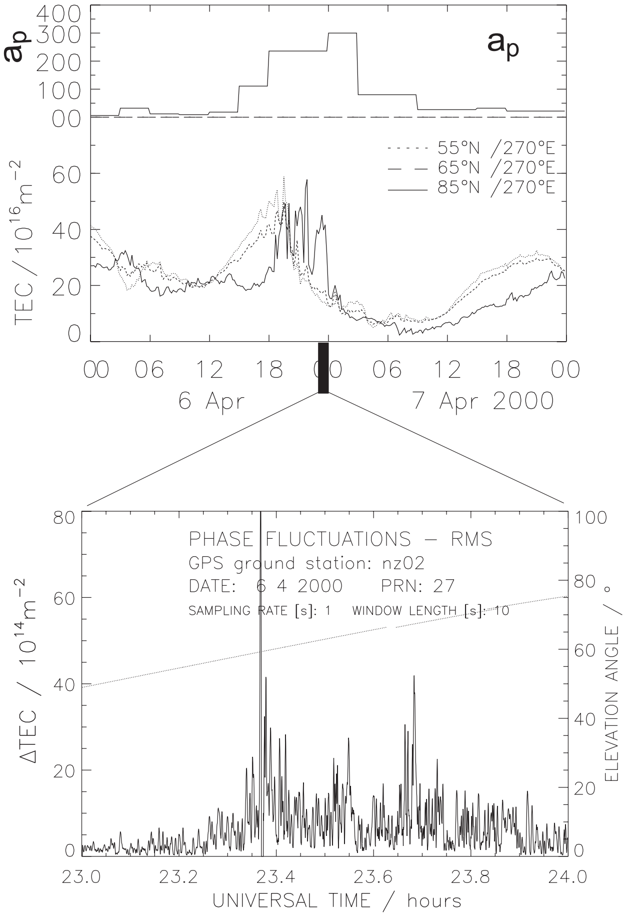


Figure 13.7. Space weather event on 6 April 2000 showing strongly enhanced geomagnetic activity (ap up to 300) which is correlated with strong fluctuations of TEC in particular at high latitudes accompanied by a high variability of the differential GPS carrier phases of GPS satellite PRN 27 measured at mid-latitudes in Neustrelitz (nz02).

magnetosphere and from the enhanced heating of the polar thermosphere due to the auroral electrojet and/or due to particle precipitation at high latitudes.

13.4.3 Small-scale irregularities in the ionosphere

Satellite navigation in the L-band is susceptible to interruptions due to the effects of ionospheric irregularities producing rapid fluctuations of the signal amplitude called scintillations. Principally, through the mechanisms of forward scattering and diffraction, small-scale irregular structures in the ionization density cause scintillation phenomena in which the steady signal at the receiver is replaced by one which is fluctuating in amplitude, phase, and apparent direction of arrival. The most commonly used parameter characterizing the intensity fluctuations is the scintillation index S_4, defined by equation:

$$S_4 = \left(\frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2} \right)^{1/2} \tag{13.5}$$

where I is the intensity of the signal and $\langle\,\rangle$ denotes averaging [26].

The scintillation index S_4 is related to the peak-to-peak fluctuations of the intensity. The exact relationship depends on the distribution of the intensity which is best described by the Nakagami distribution for a wide range of S_4 values. As S_4 approaches 1.0, the distribution approaches the Rayleigh distribution. Occasionally, S_4 may exceed 1, reaching values as high as 1.5. This is due to wave focusing caused by the irregularities.

For values less than 0.6, S_4 shows a consistent f^{-u} relationship, with the spectral index u being 3, for observations at GHz frequencies. As the scintillation becomes stronger, with S_4 exceeding 0.6, the spectral index decreases. This is due to the saturation of scintillation for Rayleigh fading under the strong influence of multiple scattering [27].

Empirically, the following table provides a convenient conversion between S_4 and the approximate peak-to-peak fluctuations P_{fluc} (dB).

S_4	P_{fluc} (dB)
0.1	1.5
0.2	3.5
0.2	6
0.4	8.5
0.5	11
0.6	14
0.7	17
0.8	20
0.9	24
1.0	27.5

This relationship can be approximated by:

$$P_{\text{fluc}} = 27.5 \cdot S_4^{1.26} \tag{13.6}$$

When scintillations occur, also the received carrier phase suffers rapid and strong changes. To measure these changes the phase standard deviation is commonly used. This parameter is crucial when estimating the loss of lock probability of GNSS receivers.

Geographically, as Figure 13.9 shows, the two areas most affected by scintillations are equatorial (+/−30° magnetic latitude) and auroral (around the poles) regions. Equatorial scintillations typically occur after local sunset and last for

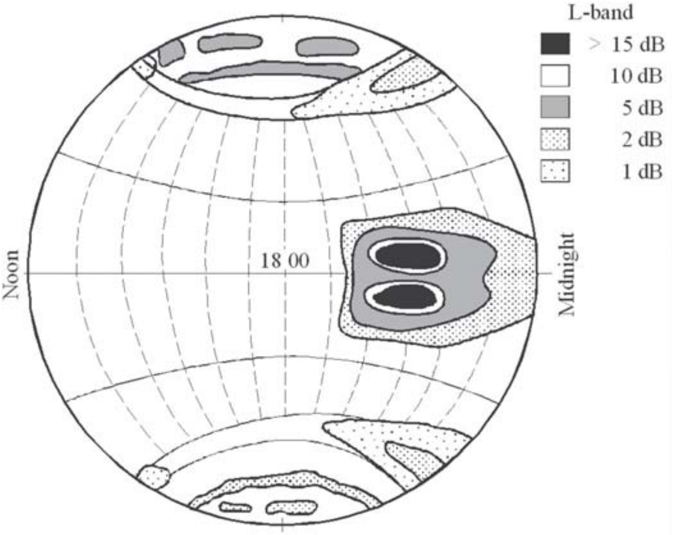


Figure 13.9. Map depicting the scintillations fading as a function of local lime and location for solar maximum conditions (Source: ITU-R Rec. 532-8, originally supplied by U.S. Delegation, [28]).

several hours – see also Figure 13.10 (see colour section) [29]. Due to the unique physical conditions at low latitudes, the equatorial ionosphere produces the most severe scintillation levels and highest TEC levels (cf. Figure 13.3, see colour section). Nevertheless, also at other regions such as at subauroral latitudes, scintillations may be observed as a direct consequence of geomagnetic/ionospheric storms [30].

The improvement of scintillation models like the Global Ionospheric Scintillation model (GISM) [32] or other models like Secan's [33, 34] will be helpful forecasting ionospheric scintillations. Forecasting of scintillations is in particular required by Safety of Life applications. Corresponding efforts led to the development of the Communication and Navigation Outage Forecast System in the U.S.A. which will essentially contribute also to explore underlying physical processes [35].

Enhanced scintillation activity in solar active years as well as in the course of ionospheric storms can cause significant link outages (e.g., cycle slips and even loss-of-lock, which means that the respective satellite has to be re-acquired (e.g., [36], [37]).

13.5 SPACE WEATHER ISSUES IN SPECIFIC NAVIGATION AND POSITIONING TECHNIQUES

Since the trans-ionospheric propagation errors are a major source of positioning errors in satellite-based navigation, the users of satellite navigation systems have to apply appropriate mitigation techniques such as: corrections based on dual-frequency techniques, model-assisted corrections, and local and/or global augmentation systems. With the help of transmitted GNSS corrections, positioning with an accuracy of a few centimeters is achievable under quiet space weather conditions.

Depending on the requirements at the user level (accuracy, observation time), different positioning techniques are applied. These can be divided into:

- Point positioning using a single receiver for determining the position.
- Differential GNSS technique using the code measurements in a network of reference stations.
- Real-time kinematic (RTK) using the low-noise carrier phases in a dense network of reference stations.

In the subsequent chapters these different techniques and their sensitivity with respect to space weather effects are discussed.

13.5.1 Point positioning

If the determination of the latitude, longitude, and altitude of a point with respect to a given coordinate system is performed with a single receiver, single-point positioning has to be taken into account. After the U.S. government stopped the artificial degradation of clock and ephemeris data (Selective Availability, SA) at midnight on 1 May 2000, the single-point accuracy increased immediately from about 100 m to about 13 m in the horizontal plane and 22 m in the vertical direction which is sufficient for a number of applications.

The majority of navigation receivers for civilian use are currently single-frequency receivers. As opposed to the dual-frequency receiver, where the ionospheric delay can be removed internally, the single-frequency receiver (or multi-frequency receivers which due to interference or other causes had to fall back to single-frequency operation), when used in stand-alone mode, requires external information about the state of the ionosphere to carry out a suitable correction. The GPS system transmits the 8 coefficients to be used with the ionospheric correction algorithm proposed by Klobuchar [38, 39].

In the Galileo system, only three coefficients plus five disturbance flags will be transmitted. The three coefficients a_0, a_1, and a_2 initializing the NeQuick Ionospheric model ([40], [41]) describe a dip-latitude dependent 'effective ionization parameter' Az, which substitutes the solar flux input parameter in the standard NeQuick model:

$$Az = a_0 + a_1 \cdot \mu + a_2 \cdot \mu^2 \tag{13.7}$$

where μ is the modified dip-latitude as defined by Rawer [42]:

$$\tan \mu = \frac{I}{\sqrt{\cos \phi}} \tag{13.8}$$

with $I =$ true magnetic dip and $\phi =$ geographic latitude.

The space weather relationship is clearly manifested in the solar flux and geomagnetic field dependencies. The expected performance of this algorithm, initialized with a broadcast message generated over a period of 24 hours is an error of less than 30% of the actual slant TEC value or 20 TEC units, whichever is the larger.

Dual-frequency receivers in stand-alone mode can largely remove the ionospheric delay by virtue of the dispersive nature of the ionosphere. With two frequencies, the TEC can be retrieved (cf. (13.3)) and thereby an 'ionosphere-free' solution can be created. To mitigate the ionospheric first-order error the new observable Φ_3 is formed by a linear combination of measured phases $\Phi_1(f_1)$ and $\Phi_2(f_2)$ for code and phase measurements at frequencies f_1 and f_2:

$$\Phi_3 = (1 + \alpha_3)\Phi_1 - \alpha_3\Phi_2 \tag{13.9}$$

where

$$\alpha_3 = \frac{f_2^2}{f_1^2 - f_2^2} \tag{13.10}$$

By ignoring higher order propagation error terms, Φ_3 becomes 'ionosphere-free'. It is evident that higher order terms depending on $1/f^3$ or $1/f^4$, as indicated in Figure 13.1, are not mitigated by this linear approach (e.g., [2, 43]). The disadvantage of this technique is that the combination is about three times noisier than the original Φ_1 and Φ_2 measurements.

13.5.2 Satellite-based augmentation systems

In order to use a satellite navigation receiver in safety-of-life critical application, an integrity concept has to be deployed which allows detecting any irregularity or error in the position exceeding the required protection level. For GPS a Satellite-based Augmentation System (SBAS) has been set up by the U.S.A. (called the Wide Area Augmentation System, WAAS, [44]) and is also under development in other parts of the world. In Europe it is called 'European Geostationary Navigation Overlay System (EGNOS)' [45], in Japan 'MT-Sat Augmentation System (MSAS)' [46], and in India 'GPS Aided Geo Augmented Navigation (GAGAN)' [47]. The concept of the SBAS is based on a network of reference stations, which are continuously observing the signals from navigation satellites, and one or more geostationary satellites to transmit the derived corrections to the users. It allows improvement of the position accuracy (differential corrections) and measurement integrity, which is required for Safety-of-Life (SoL) critical applications. The ionospheric correction is transmitted in the form of grid values in a 5 × 5 degree matrix. At the user level the vertical TEC at the ionospheric pierce points is calculated by interpolation between grid points. In addition to the vertical delay, also grid-point vertical error estimates are being transmitted. These error values are calculated in the SBAS processing center end they are designed in such a way that the user receiver (typically on board a civilian aircraft) can safely determine whether the actual navigation solution is within the specified error bounds or not. If the error calculated by the user receiver exceeds a pre-determined threshold, the operator is alerted that the system is not meeting the requirements. The probability of this procedure failing (probability of missed alarm) is specified as 10^{-7}. This means that the error bounds are set up in an extremely conservative way so that system integrity has priority over any other parameter such as system availability or continuity.

SBAS systems have the advantage of allowing safe aircraft navigation and landing over a large area without specific installations needed at every airport. The initial phase of operation of the WAAS system has shown good performance, even in the presence of ionospheric storms. It is however expected, that in order to achieve a comparable good performance in equatorial regions, where gradients and rapid fluctuations of TEC are much more pronounced, the system may need to incorporate more information in the broadcast message to the user.

The European system has been tested by using synthetic ionospheric storms which have been derived from actual observations by ionosondes and GPS receivers. Five different scenarios have been created, based on real storms from March 1989, November 1991, August 1998, September 1998, and October 2003. The 3-D distribution of electron density was created by deriving grid values of foF2 and M(3000) from the observed data and applying the NeQuick model (e.g., [40, 41]) to this grid. These scenarios, which represent snapshots of the ionosphere at 1-second intervals, were fed into the EGNOS End-to-End Simulator (EETES) [48]. Five different storms were generated and all of them were handled by the EGNOS simulator without any problems.

Besides the problems of ionospheric storms, other critical questions relating to the ionosphere, such as the threat posed by travelling ionospheric disturbances (TIDs) and ionospheric radio scintillations have been carefully analysed for the EGNOS system.

13.5.3 Local augmentation systems

Local augmentation systems are using differential corrections from nearby reference stations. Using this technique, all common errors such as clock errors and orbit errors cancel out when differencing the observation equations ((13.1) and (13.2)) at two frequencies. Residual errors due to differences in the ionospheric and troposphere delay cannot be removed. This chapter addresses the error that can be expected due to changes in the ionospheric delay.

For single-frequency systems which are using differential corrections, the residual ionospheric delay error is directly related to the difference in slant ionospheric delay between the satellites of common visibility and the base station on the one hand and the user station (rover) on the other hand. This difference will depend on the ionospheric TEC gradient and the baseline length (the distance between base station and rover). Under non-storm conditions in mid-latitudes, a single-frequency GPS receiver using differential corrections with a baseline below 50 km will typically achieve a horizontal position accuracy of better than 5 m (95%). However, in equatorial regions during daytime the maximum baseline length allowable to remain within this error bound can be significantly lower.

For aeronautical use, a Local Area Augmentation System (LAAS) has been devised. This system uses a local reference receiver at the airport to support the landing of aircraft. Compared with SBAS, LAAS has less information about the state of the ionosphere since it does not have a widely spread network of reference stations and is therefore more risky if a large storm induced TEC gradient passes through the site [49].

High-precision RTK positioning networks apply data from multiple reference stations belonging to the network. Accurate positioning at centimetre-level is achieved by solving the carrier phase ambiguities for each satellite visible in the network. Typical distances between the reference stations are in the order of 20–50 km. The unambiguous solution of the positioning algorithm requires measurements of the local errors (ionospheric and tropospheric) at each reference station and a subsequent estimation of the rover position by interpolation methods within the network. If strong temporal and/or spatial gradients of the ionospheric ionization occur, the ambiguity resolution is not possible within a certain time limit. This effect may seriously degrade a reference network as is demonstrated in Figure 13.11 (see colour section) where the number of tracked, processed and solved satellites is shown for a European GPS reference network during the severe geomagnetic/ionospheric storm on 29 October 2003.

The figure indicates that the ambiguity solution may fail due to a severe space weather effect causing serious ionospheric perturbations. It is interesting to note that strong ionospheric perturbations are indicated in the high-latitude area at 08:00 UT (i.e., a few hours before the break down in the GPS service was observed). This fact shows the potential of forecasting performance degradation in GPS networks. The definition of specific ionospheric perturbation indices seems to be a promising approach for better and faster quantification of the strength and impact of ionospheric perturbations on GNSS [50].

The trouble usually arises due to a strong spatial de-correlation of TEC which increases with the growing horizontal gradients of the plasma distribution. A statistical estimation of horizontal TEC gradients reveals large-scale gradients of up to about 6 TECU/1,000 km under high solar activity conditions at an occurrence probability level of about 1%. Occasionally, during severe ionospheric storms this value may increase by a factor of 10 or even more [51].

As Figure 13.12 shows, the horizontal TEC gradients increase during enhanced geomagnetic activity which is represented here by 3-hourly ap indices. Thus, ionospheric storms and enhanced activity of TIDs may reduce the performance of RTK systems either by reducing their accuracy or by increasing the so-called fixing time which is required to obtain a unique solution of the positioning equations.

13.6 SUMMARY

Space weather phenomena may have a significant impact on GNSS signals. Fortunately, the GNSS technique itself provides a unique opportunity to monitor ionospheric key parameters continuously on regional and/or global scales in near-

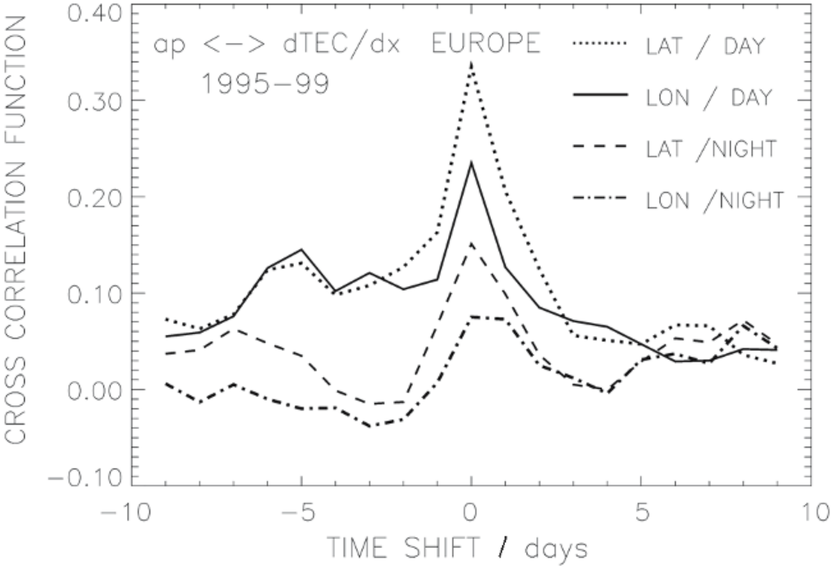


Figure 13.12. Cross correlation function of latitudinal and longitudinal TEC gradients with the 3-hourly geomagnetic index ap for day- and night-time conditions indicating that the correlation is principally stronger in the day-time.

real-time which may provide ionospheric corrections in augment navigation and positioning networks.

Considering complex space weather events such as the event beginning on 6 April 2000, it becomes clear that the ionosphere is modified in a complex and up to now unpredictable manner.

The close correlation between horizontal TEC gradients with geomagnetic activity indices gives the hope that TEC gradients may be forecast by predicted ionospheric and/or geomagnetic indices. Large-scale gradients degrade the performance of large area augmented systems for GNSS like WAAS and EGNOS.

If GNSS users are provided with additional ionospheric information, they can save higher accuracy and/or reliability of the measurements. Thus, permanent monitoring of the ionosphere and subsequent operational derivation of relevant information concerning the horizontal structure of the ionosphere and its dynamics is helpful for users in navigation, positioning, and surveying.

13.7 ACKNOWLEDGEMENTS

The authors wish to gratefully acknowledge the work of the European COST 296 Group and of the International SBAS–IONO group, which has provided substantial contributions to this chapter. Furthermore, we thank the International GPS Service

providing the opportunity for using high-quality GPS data on global scales over many years.

13.8 REFERENCES

[1] Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J., *GPS: Theory and Practice* (5th edition). Springer-Verlag, Wien, 2001.

[2] Brunner, F.K. and Gu, M., An improved model for the dual frequency ionospheric correction of GPS observations. *Manuscr. geodetica*, **16**, 205–214, 1991.

[3] Davies, K. and Hartmann, G.K., Studying the ionosphere with the global positioning system. *Radio Sci.*, **84**, 1695–1703, 1997.

[4] Sardón, E., Rius, A., and Zarraoa, N., Estimation of the receiver differential biases and the ionospheric total electron content from Global Positioning System observations. *Radio Science*, **29**, 577–586, 1994.

[5] Jakowski, N., TEC Monitoring by Using Satellite Positioning Systems. In: *Modern Ionospheric Science* (Eds, H. Kohl, R. Rüster, K. Schlegel). EGS, Katlenburg-Lindau, ProduServ GmbH Verlagsservice, Berlin, pp 371–390,1996.

[6] Wilson, B.D., Mannucci, A.J., and Edwards, C.D., Subdaily northern-hemisphere maps using an extensive network of GPS receivers. *Radio Sci.*, **30**, 639–648, 1995.

[7] Hernandez-Pajares, M., Juan, J.M., and Sanz, J., High-resolution TEC monitoring method using permanent ground GPS receivers. *Geophys. Res. Lett.*, **24**, 1643–1646, 1997.

[8] Saito, A., Fukao, S., and Miyazaki, S., High-resolution mapping of TEC perturbations with the GSI GPS network over Japan. *Geophys. Res. Lett.*, **25**, 3079–3082, 1998.

[9] Jakowski, N., Sardon, E., and Schlueter, S., GPS-Based TEC observations in comparison with IRI-95 and the European TEC Model NTCM2. *Adv. in Space Res.*, **22**, 803–806, 1998.

[10] Jakowski, N., Schlueter, S., and Sardon, E., Total electron content of the ionosphere during the geomagnetic storm on January 10, 1997. *J. Atmos. Sol.–Terr. Phys.*, **61**, 299–307, 1999.

[11] Jakowski, N., Generation of TEC Maps Over the COST Area Based on GPS Measurements. In: *Proc. 2nd COST 251 Workshop (Ed., A. Vernon), 30–31 March, 1998, Side,* pp. 51–57, 1998.

[12] Hernandez-Pajares, M., Juan, J.M., and Sanz, J., New approaches in global ionospheric determination using ground GPS data. *J. Atmos. Sol.–Terr. Phys.*, **61**, 1237–1247, 1999.

[13] Hajj, G.A. and Romans, L.J., Ionospheric electron density profiles obtained with the Global Positioning System: Results from the GPS/MET experiment. *Radio Sci.*, **33**, 175–190, 1998.

[14] Heise, S., Jakowski, N., Wehrenpfennig, A., Reigber, C., and Luehr, H., Sounding of the topside ionosphere/plasmasphere based on GPS measurements from CHAMP: Initial results. *Geophys. Res. Lett.*, **29**, No. 14, 10.1029/2002GL014738, 2002.

[15] Jakowski, N., Wehrenpfennig, A., Heise, S., Reigber, C., and Lühr, H., GPS radio occultation measurements of the ionosphere on CHAMP: Early results. *Geophys. Res. Lett.*, **29**, No. doi 10, 10.1029/2001GL014364, 2002.

[16] Jakowski, N., Wilken, V., Schlueter, S., Stankov, S., and Heise, S., Ionospheric space weather effects monitored by simultaneous ground and space based GNSS signals. *J. Atmos. Sol.–Terr. Phys.*, **67**, 1074–1084, 2005.

[17] Jakowski N., Fichtelmann, B., and Jungstand, A., Solar activity control of ionospheric and thermospheric processes. *J. Atmos. Terr. Phys.*, **53**, 1125–1130, 1991.

[18] Leonovich, L.A., Afraimovich, E.L., Romanova, E.B., and Taschilin, A.V., Estimating the contribution from different ionospheric regions to the TEC response to the solar flares using data from the international GPS network. *Ann. Geophysicae*, **20**, No. 12, 1935–1941, 2002.

[19] Jakowski, N., Schlueter, S., Heise, S., and Feltens, J., Satellite technology glimpses ionospheric response to solar eclipse, EOS. *Transactions. American Geophysical Union*, **80**, 51, 21 December, 1999.

[20] Mikhailov, A.V. and Schlegel, K., Equinoctial transitions in the ionosphere and thermosphere. *Ann. Geophysicae.*, **19**, 783–796, 2001.

[21] Fejer, B.G., The equatorial ionospheric electric fields: A review. *J. of Atmos. Terr. Phys.*, **43**, 377–386, 1981.

[22] Huttunen, K.E.J., Koskinen, H.E.J., Pulkkinen, T.I., Pulkkinen, A., Palmroth, M., Reeves, E., and Geoffrey, D., and Singer, H.J., April 2000 magnetic storm: Solar wind driver and magnetospheric response. *J. Geophys. Res.*, **107**, A12, pp. SMP 15–1, CiteID 1440, doi 10.1029/2001JA009154

[23] Jakowski, N., Wehrenpfennig, A., Heise, S., and Kutiev, I., Space weather effects on transionospheric radio wave propagation on 6 April 2000. *Acta Geod. Geoph. Hung.*, **37**, No. 2-3, 213–220, 2002.

[24] Jakowski, N., Heise, S., Wehrenpfennig, A., Schlueter, S., and Reimer, R., GPS/GLONASS-based TEC measurements as a contributor for space weather forecast. *J. Atmos. Sol.–Terr. Phys.*, **64**, 729–735, 2002.

[25] Foerster, M. and Jakowski, N., Geomagnetic storm effects on the topside ionosphere and plasmasphere: A compact tutorial and new results. *Surveys in Geophysics*, **21**, No. 1, 47–87, 2000.

[26] ITU-R P.531-8, Ionospheric propagation data and prediction methods required for the design of satellite services and systems, Geneva 2005.

[27] Fu, W.X., Han, S., Rizos, C., Knight, M., and Finn, A., Real-time ionospheric scintillation monitoring using GPS. *12th Int. Tech. Meeting of the Satellite Division of the U.S. Inst. of Navigation GPS ION'99, Nashville, Tennessee, 14-17 September 1999,* pp. 1461–1471,

[28] Basu, S., McClure, J.P., Basu, S., Hanson, W.B., and Aarons, J., Coordinated study of equatorial scintillation and *in situ* radar observations of night-time F region irregularities. *J. Geophys. Res.*, **85**, 5119, 1980.

[29] Basu, S., Kudecki, E., Basu, S., Valladares, C.E., Weber, E.J., Zengingonul, H.P., Bhattacharyya, S., Shehan, R., Meriwethwer, J.W., Biondi, M.A., *et al.*, Scintillations, plasma drifts and neutral winds in the equatorial ionosphere after sunset. *J. Geophys. Res.*, **101**, 26795–26806, 1996.

[30] Aarons, J., Gurgiolo, C., and Rodger, A.S., The effects of magnetic storm phases on F-layer irregularities below the auroral oval. *Radio Sci.*, **23**, No. 3, 309–319, 1988.

[31] Van Dierendonck, A.J. and Arbesser-Rastburg, B., Measuring ionospheric scintillation in the equatorial region over Africa, including measurements from SBAS geostationary satellite signals. *Proc ION GNSS Conf, 21–24 Sept 2004, Long Beach, U.S.A.*

[32] Beniguel, Y., Global Ionospheric Propagation Model (GIM): A propagation model for scintillations of transmitted signals. *Radio Sci.*, **37**, 400–412, 2002.

[33] Secan, J.R., Bussey, E.J., Fremouw, E.J., and Basu, S., An improved model of equatorial scintillation. *Radio Sci.*, **30**, 607–617, 1995.

[34] Secan, J.R., Bussey, E.J., and Fremouw, E.J., High-latitude upgrade to the wideband ionospheric scintillation model. *Radio Sci.*, **32**, 1567–1574, 1997.

[35] C/NOFS Report (AFRL/VS-TR-2003-1501), Communication/Navigation Outage Forecasting Saystem (C/NOFS) Science Plan, The C/NOFS Science Definition Team, (O. de La Beaujardiere, Editor), 2003.

[36] Skone, S., Kundsen, K., and de Jong, M., Limitations in GPS receiver tracking performance under ionospheric scintillation conditions. *Phys. Chem. Earth* (A), **26**, 613–621, 2001.

[37] Basu, S., Groves, K.M., Basu, S., Sultan, P.J., Specification and forecasting of scintillations in communication/navigation links: Current status and future plans. *J. Atmos. Sol.–Terr. Phys.*, **64**, 1745–1754, 2002.

[38] Klobuchar, J.A., Ionospheric time-delay algorithm for single-frequency GPS users. *IEEE Trans. Aerospace and Electr. Syst.*, AES-23, 325–331, 1987.

[39] Klobuchar, J.A., Ionospheric Effects on GPS. In: *Global Positioning System: Theory and Applications* (Eds, Parkinson and Spilker), Vol 1, AIAA Inc, Washington D.C., 1996, pp. 485–515.

[40] Hochegger, G., Nava, B., Radicella, S., and Leitinger, R., A family of ionospheric models for different uses. *Phys. Chem. Earth* (C), **25**, No. 4, 307–310, 2000.

[41] Leitinger, R., Radicella, S., and Nava, B., Electron density models for assessment studies: New developments. *Acta Geod. Geoph. Hung.*, **37**, 183–193, 2002.

[42] Rawer, K., *Wave Propagation in the Ionosphere*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1993.

[43] Leitinger, R. and Putz, E., Ionospheric refraction errors and observables. In: *Atmospheric Effects on Geodetic Space Measurements* (Ed., F.K. Brunner) (Monograph 12). School of Surveying, University of New South Wales, Australia, pp. 81–102, 1988.

[44] El-Alrini, M.B., Poor, W., Lejeune, R., and Conker, R., An introduction to Wide Area Augmentation System and its predicted performance. *Radio Sci.*, **36**, No. 5, 1233–1240, 2001.

[45] Gauthier, L., Ventura-Traveset, J., Toran, F., de Lesthievent C., and Bedu J.-Y., EGNOS operations and their planned evolution. *ESA Bulletin*, **124**, 56–61, November, 2005.

[46] Tomita, H., Harigae, M., and Hoshinoo, K., Flight evaluation of GPS aided inertial navigation avionics with MSAS augmentation (MSAS–GAIA). *Proc ION GPS/GNSS*, 2819–2827, 2003.

[47] Suryanarayana Rao, K.N. and Pal, S., The Indian SBAS System – GAGAN. *Proc. India–U.S. Conf. on Space Science, Applications and Commerce, Bangalore, June 2004.*

[48] Arbesser-Rastburg, B., Ionospheric corrections for satellite navigation using EGNOS. *Proc. URSI GA 2002, Paper 2140, Maastricht, 2002.*

[49] Datta-Barua, S., Walter, T., Konno, H., Blanch, J., and Enge, P., Verification of low latitude ionosphere effects on WAAS during October 2003 geomagnetic storm. *Proc. IoN Annual Meeting, Cambridge, MA, July 2005, http://waas.stanford.edu/˜wwu/papers/gps/PDF/DattaBaruaIONAM05.pdf*

[50] Jakowski, N., Stankov, S.M., Schlueter, S., and Klaehn, D., On developing a new ionospheric perturbation index for space weather operations. *Adv. Space Res.*, doi 10.1016/j.asr.2005.07.043, 2005.

[51] Luo, M., Pullen, S., Dennis, J., Konno, H., Xie, G., Walter, T., Enge P., Datta-Barua, S., and Dehel, T., LAAS onosphere Spatial Gradient Threat Model and impact of LGF and airborne monitoring. *Proc. ION GPS/GNSS 2003, Portland, September 2003,* pp. 2256–2274.

14

Forecasting space weather

Dimitris Vassiliadis

Forecasting space weather events presents the ultimate challenge to a space physics model. Not only should physical constraints be satisfied, but also practical issues such as timeliness, accuracy and reliability must be adequately addressed. Modern space weather forecasters and users rely on a great variety of space environment forecast systems, ranging from simple non-linear regressions and ending with fairly complex information-based (empirical), physical and hybrid models. Over the last two decades, model-based predictions of space environments have become increasingly sophisticated. Especially since the early 1990s when real-time data began to be available online, time-dependent inputs, and later data assimilation and related techniques such as Kalman filtering, have significantly improved prediction accuracy. Model verification is an important step in evaluating a space weather model's spectrum of forecast capabilities. Case studies are used to illustrate the above concepts. The chapter concludes with a summary of recent developments and future prospects.

14.1 INTRODUCTION

Space weather forecasting is the specification of the state of a space environment at a future time. It is one of the most challenging hypothesis-testing methods since, in addition to formulating and implementing a test of a model or theory, it is burdened with the complications of issuing a predictive statement with limited information and in advance of an event. A related concept is nowcasting – the specification of the present state of the environment. As seen in the preceding chapters, space weather comprises a large number of complex electrodynamic and plasma physical phenomena and their effects on human-made technologies and infrastructure. The complexity can be divided into three parts:

(1) The space environment itself (the region of geospace, heliospace, or planetospace).
(2) The physical effects which are manifested in that region,
(3) The technologies affected.

First, space weather models represent the activity of a particular region of geospace, and thus there are separate types of space weather modelling for the solar, heliospheric/interplanetary, magnetospheric, ionospheric and neutral atmosphere domains. Probably the greatest number and variety of technology assets affected by space weather are located within Earth's magnetosphere (satellites, re-entry vehicles, and the International Space Station). In terms of space weather economics, these assets have the highest costs, and therefore some of the highest priorities in forecasting (Table 14.1). A smaller number of assets are located on Earth's surface (electric power grids, pipeline systems, communication networks, and so on) and in extraterrestrial space (interplanetary spacecraft and lunar exploration bases). Second, within its domain, a space environment model may represent one or several effects of interest, which are usually electrodynamic and plasma physical phenomena. For instance, a space radiation model for the Van Allen belts should include several physically distinct components of radiation, such as cosmic rays, solar energetic particles, radiation-belt electrons, ion species, and so on (Heynderickx, 2002). Third, different technologies are vulnerable in different

Table 14.1. High-priority forecast and nowcast models for NOAA/SEC (2003) based on customer need. The enumeration is simply indicative, and no type of model is given priority over any other. Priorities vary with factors such as readiness, accuracy and lead time. Different aspects of each model or model area could have different priorities. Examples of forecast users are given. (After a presentation by T. Onsager, NOAA/SEC, September 2003.)

Priority	Model	Customer examples	Space weather area
1	Solar energetic particle forecast	Commercial airlines HF communication Satellite launch Manned space flight	Solar energetic particles
2	Regional geomagnetic activity forecast and nowcast	Electric power Commercial airlines HF communication	Geomagnetic activity
2	Relativistic electron forecast for International Space Station	Manned space flight	Radiation belt electrons
3	Ap prediction	Various military/ civilian users	Geomagnetic activity
3	Ionospheric disturbance forecast and nowcast	Navigation (GPS) Explorations/ Surveying	Ionospheric scintillation
4	Dst prediction	Various military/ civilian users	Geomagnetic activity

degrees to the same physical effect. For example, power grids and pipelines respond differently to the ground geoelectric field (Baker and Kappenman, 1996).

Each combination of space weather environment, physical effect and technology requires, in principle, a separate model, and presents a spectrum of challenges and opportunities in terms of data collection, model and theory development, real-time monitoring, and validation and verification (Figure 14.1, colour section). As a result, today's space weather models are at very different evolutionary stages (Daglis *et al.*, 2001).

Instead of the large diversity of models and cases, this chapter focuses on the underlying theoretical and practical issues of forecasting. We begin by showing that space weather phenomena can be predicted in terms of information dynamics, or in terms of electrodynamics and plasma physics.

14.1.1 Empirical and physical models: tracking information versus energy

The interaction of space weather systems can be effectively described in two different ways. Fundamentally, it can be viewed as an exchange of information. For instance, changes of interplanetary magnetic field (IMF) B_z polarity may not involve significant changes in the solar wind's internal energy balance; for instance, they may reflect the passage of a convected structure or the presence of turbulence. However, in the context of the interaction with Earth's magnetosphere, a southward turn of the IMF at the L_1 libration point means that magnetic reconnection will commence on the dayside equatorial region 30–60 minutes later. The exact amount of energy transferred during dayside reconnection is not known to as high a precision as the amplitude of the resulting ionospheric or magnetospheric currents. Thus, the IMF B_z amplitude is a useful input variable for a magnetospheric model, even if the model does not explicitly account for the energy transfer. Representing space environment interactions in terms of information dynamics leads to the development of information-based, or empirical, models.

Information-based models are typically the first to be developed for a given space weather environment. Simple models are non-linear regressions and superposed epoch analyses, and more generally averages of a field or other variable conditioned on measurement location and/or activity level. Models include semi-empirical representations of the solar wind expansion (Wang and Sheeley, 1990; Arge *et al.*, 2004), of the magnetospheric magnetic field (Tsyganenko, 1989, 1995; Tsyganenko and Sitnov, 2005), and of the high-latitude magnetospheric electric field (Weimer, 1996, 2001; Papitashvili *et al.*, 1998; Kappeman *et al.*, 2000).

Time-dependent information-based models define the state, or activity level, of the space environment. The state vector is comprised of several physical or proxy variables. For instance, in order to describe the essential, large-scale dynamics of the dynamic terrestrial magnetosphere, one uses a state vector such as:

$$x(t) = (\Phi_{\mathrm{lobe}}, E_{\mathrm{tail}}, I_{\mathrm{tail}}, E_{\mathrm{iono}}, I_{\mathrm{iono}}) \tag{14.1}$$

whose components represent two energy repositories (Φ: lobe flux; E_{tail}: cross-tail electric field) and two energy sinks (E_{tail}, I_{tail}: plasma sheet; E_{iono}, I_{iono}: ionosphere).

Proxies for these variables can be used (the PC index for the lobe flux; the AE index for the ionospheric current) and the observed dynamics are fit to differential or difference equations (for systems of nonlinear differential equations see Klimas, 1994; Horton and Doxas, 1996; 1998). The complementary approach is to start from observed time series, model their dynamics, and interpret it physically (Vassiliadis *et al.*, 1995, 1996). In those cases where only scalar or vector time series are available, the state vector is a lag, or delay, vector

$$x(t) = [x(t), x(t - \Delta t), x(t - 2\Delta t), \ldots, x(t - (m - 1)\,\Delta t)] \tag{14.2}$$

where $x(t - i\Delta t)$ is the i-th lag or delay, Δt is the time resolution, and m is the chosen dimension of the state vector. The convention is for smaller lags to appear, followed by larger lags. State space techniques and their physics are reviewed systematically elsewhere (Vassiliadis, 2006).

In physics-based models, on the other hand, space weather events are described as the transport of physical quantities (energy, momentum, mass, helicity, and so on). Physical models are approximations of the system at the electrodynamic, MHD, or kinetic level. MHD space weather models include descriptions of the corona and solar wind (Linker *et al.*, 1999; Odstrcil and Pizzo, 1999), and Earth's magnetosphere (Fedder and Lyon, 1995; Groth *et al.*, 2000). Kinetic models have been developed for the ring current (Fok *et al.*, 1995; Ebihara *et al.*, 2005) and the electron radiation belts (Boscher, 1996; Bourdarie *et al.*, 2005; Li *et al.*, 2001).

Realistic space weather models are hybrids of both the information-based and the physics-based approach. The development of information-based models, for instance, is guided by knowledge of the plasma physics of the system to determine variables of interest (such as moments of the plasma distributions, or relevant components of the IMF). On the other hand, physics-based models often rely on empirical approximations to represent external, non-linear or subgrid processes.

Forecast performances of information-based and physics-based models for the same system may vary widely. Empirical models are trained to reproduce specific sets of physical or proxy variables, and are not constrained by conservation laws as are physical models. For these reasons, empirical models tend to be faster and more accurate than physical models in predicting the set of variables on which they are trained. On the other hand, physical models can provide predictions on larger sets of variables. In addition, they can predict events outside the training domain in their phase space (provided they include the physics for such regions), whereas extrapolation of empirical models beyond their training domain is uncertain. Traditionally, empirical models have sometimes been perceived as 'rigid', in the sense that they are limited by the spatial, temporal and energy resolution of the measured data, and cannot be generalized to predict additional physical quantities. However, the wealth of new datasets and the development of hybrid models have reduced these limitations in several cases.

14.1.2 Model predictions and forecasts

The output of a model, given the conditions that lead to a space weather event, constitutes a prediction for the occurrence or a particular property of the event. (The evaluation of the prediction's accuracy will be discussed below.) Model predictions can be obtained for both future and historical events. In contrast, a forecast is a prediction about an event that has not occurred at the time the prediction is made. The main properties of a forecast are the following:

(1) *Timeliness* A forecast should be made as early in advance as possible to provide ample time for users to make operational changes. The time difference between the issue of a forecast and the actual event is called the lead time.
(2) *Accuracy* This is the agreement between the properties of the forecast and those of the actual event. Depending on the space weather application, the relevant property can be cast as an amplitude, a spectrum, a probability distribution, and so on.
(3) *Reliability* The forecast should ideally reach the user at all times under any conditions – the forecast system should be robust. Here the forecast system includes not only the numerical space environment model, but also any auxiliary systems for data acquisition, processing and communication, implemented in hardware and software.

14.1.3 Climatology and dynamics

Some of the most efficient models are obtained by simply averaging a representative number of relevant past measurements of the system's activity. Examples include the solar or interplanetary variation over a solar cycle, obtained from measurements over several recent cycles (Crooker, 2000; Richardson *et al.*, 2001), or the corresponding geomagnetic activity variation over similar timescales. Such models of the historically average activity are called climatological. A periodic climatology model for $x(t)$ has the form:

$$x(t) = a \cdot u(t) + b \tag{14.3}$$

where $u(t) = A\sin(\omega t)$ is an external driver with a particular periodicity (diurnal, seasonal, solar-cycle, and so on).

More sophisticated models add dynamic terms to the baseline provided by climatology. The simplest dynamical term is the persistence term, representing effects of the system's inertia. Adding a persistence term to model leads to:

$$x_{t+\Delta t} = (1-\varepsilon)\,x_t + \varepsilon(au_{t+\Delta t} + b) \tag{16.4}$$

where the value of ε is chosen to balance the persistence over the driver, and can be related to the time step, or resolution, Δt. For small Δt and $\varepsilon = \Delta t$, equation (14.4) is a discrete-time approximation of a differential equation for $x(t)$.

14.1.4 Input–output modelling

A key point in forecasting is that space weather environments are open; that is, they exchange mass, momentum and energy with one or more neighbouring plasmas. Therefore, plasma-physical 'box models' need to include time-dependent inputs such as electric and magnetic fields, injected beams, and/or inflowing plasmas before they can be developed as space weather tools. Unlike many other geophysical systems, however, the fluctuations in the driving variables are not small compared to their mean values, but, depending on the space environment, they can often significantly exceed them. The response and stability of the space environment to these changes requires a completely different type of modelling than that of closed or constant-input systems.

The open-system property means that a space weather model needs to keep track of the time history of its input variables. The main direction of information transfer is heliocentric – from the Sun to the edge of the heliosphere. However, there are many different pathways for these quantities to be transmitted, and information propagation speeds differ greatly among different environments and even within the same environment. For instance, in the interplanetary medium there are orders of magnitude of differences in propagation speed between electromagnetic radiation, energetic particles, waves and flows. Thus, a comprehensive model may need to keep track of very different time histories for each type of interplanetary input. The accuracy of solar wind and particle propagation is an important factor. The propagation starts at the measurement point – typically the L_1 libration point – to a reference location, such as the subsolar point on the dayside magnetopause at about $8\,R_E$ upstream. Recent advances in empirical modelling using minimum-variance-analysis methods have increased the propagation accuracy (Weimer *et al.*, 2003).

If input variables from more than one upstream region are available, each one offers distinct advantages and disadvantages. The further upstream the measurements are taken, the more advantageous they are in terms of a long prediction lead time. Solar-surface measurements provide the longest lead times in that respect. On the other hand, input variables from regions that are close to the space environment of interest will probably be more correlated with its activity rather than input variables from further upstream. Solar and solar wind perturbations are distorted as they propagate through interplanetary space. Clearly, the decision about which input variables to use will be based on their relative trade-offs.

14.1.5 An historical note

It is instructive to put the current developments in an historical context, because although space weather has been a rapidly developing field in the last two decades, it follows a similar path to other disciplines: meteorology and oceanography (see Chapter 2). Space forecasting efforts began in the mid-1960s, but the growth of the field was slow until the early 1990s when real-time data and powerful computing became available. In 1995, a national space weather plan was formally established in the US (Behnke and Tascione, 1995; Behnke *et al.*, 1997), followed by European and Asian national initiatives. Today, space forecast centres have been put in place by several national governments, academia, and private industry. Current space weather forecasting capabilities are comparable to the tropospheric weather science in the late 1960s, but evolve at a much faster pace. The end of this chapter features a brief account of several current forecasting efforts and the future outlook.

14.2 PREDICTIVE MODEL DEVELOPMENT: RING CURRENT DYNAMICS AND THE Dst INDEX

In this and the next section we use the Dst geomagnetic index as an example for modelling and forecasting a space weather environment (the ring current). We have chosen Dst because its time variations can be well captured by simple dynamical models with a clear physical interpretation. The index is designed to represent the magnetic effects of the ring current and is defined as a weighted average of the horizontal perturbation's north–south component, measured at four mid-latitude locations. The ring current is the dominant contributor to the mid-latitude magnetic activity (although not the only one[1]).

Southward turns of the IMF are followed by intensifications of the ring current. Central to Dst dynamics are the variations of the east–west rectified component of the interplanetary electric field, $(V_{\mathrm{SW}} \times B)_y = V_{\mathrm{SW}} B_{\mathrm{South}}$, where V_{SW} is the radial component of the solar wind velocity and B_{South} is the rectified IMF B_z component (0 when B_z is positive, and $-B_z$ when B_z is negative). GSM coordinates are used. Applied to the magnetotail, this interplanetary field component leads to enhanced plasma sheet convection earthward into the inner magnetosphere and thereby to an increase in the ring current ion population. As a starting point in modelling Dst, one can consider a linear regression between the two variables:

$$Dst(t) = c + b \cdot V_{\mathrm{SW}}(t)\, B_{\mathrm{South}}(t) \tag{14.5}$$

where time is measured in hours, and therefore the variations in the solar wind driver and the resulting plasma sheet convection can be considered as instantaneous. Increases in $V_{\mathrm{SW}} B_{\mathrm{South}}$ lead to intensifications of the ring current and decreases to Dst from its equilibrium value of approximately -50 nT. The regression is a simple climatology model.

The static model does not account for time variations of Dst during magnetic storms. For instance, when $V_{\mathrm{SW}} B_{\mathrm{South}}$ goes to zero, the |Dst| magnitude decreases, approximately as a slow exponential, reflecting the ion loss in the ring current (via magnetopause exit, wave–particle interactions, charge exchange, and so on). To a

---

[1] Several currents other than the ring current contribute to the measured Dst index. The effect of the most important one – the magnetopause current which responds to high interplanetary pressure – is removed from Dst in a preprocessing stage (Burton *et al.*, 1975). We will not discuss the distinction here, and assume that effects due to all currents except the ring current are negligible.

large extent, balancing the interplanetary driving term with a loss term and adding a persistence term can capture these dynamics. The result is a first-order differential equation (Burton *et al.*, 1975):

$$\frac{dDst(t)}{dt = bV_{\mathrm{SW}}(t)\, B_{\mathrm{South}}(t) - \frac{Dst(t)}{\tau}} \tag{14.6}$$

where the term $bV_{\mathrm{SW}}(t)\, B_{\mathrm{South}}(t)$ represents the injection rate of plasma sheet ions into the ring current. The decay time τ represents the effects of the loss processes mentioned. The Burton *et al.* (1975) model is a simple, yet effective model of the Dst time variations with physically interpretable terms. A comparison of the model with the observed Dst is shown in Figure 14.2 (see colour section). The timescale τ of the model ranges from a few hours to a fraction of a day (Burton *et al.*, 1975).

In order to measure the degree of success in reproducing the observed variation, the time series of the model-predicted index, $\hat{D}st(t)$, is compared to observations, $Dst(t)$. (Section 14.5 reviews several techniques and related issues.) Here we mention a standard technique: the correlation coefficient:

$$C_{(\hat{D}st,Dst)} = \frac{1}{T}\frac{1}{\sigma_{\hat{D}st}\sigma_{D_{st}}}\int_0^T (\hat{D}st(t) - \bar{\hat{D}}st)(Dst(t) - \bar{D}st)\, dt \tag{14.7}$$

where $\bar{X}$ and σ_X are the average and standard deviation of variable X, respectively. This correlation takes values in the range 60–70%. The percent variance of Dst explained by this type of model is approximately the square of the correlation, in the range 35–50%.

14.3 ENHANCING THE MODEL

Having the preceding example as a baseline model, we now consider additional complexities.

14.3.1 Time dependence

One way to generalize the regression (14.5), in order to both improve on forecast accuracy and account for more physics represented, is to include a longer history of the recent solar wind input rather than a single term. Starting from (5) we write the current state of the system in terms of the history of the solar wind input:

$$Dst(t) = \int_0^{\infty} H(\tau)\, V_{\mathrm{SW}}(t-\tau)\, B_{\mathrm{South}}(t-\tau)\, d\tau \tag{14.8}$$

The coupling between Dst and $V_{\mathrm{SW}} B_{\mathrm{South}}$ is represented by the impulse response function $H(\tau)$ which is convolved with the input. If $H(\tau)$ is known, this so-called finite impulse response (FIR) model can be used to predict the linear dynamics of Dst. The fundamental property of the impulse response function is that if the solar wind input is an impulse (similar to a Δ function) the estimated geomagnetic amplitude is $Dst(t) = H(t)$.

The impulse response function can be obtained either analytically, by applying a Laplace transform to a known model such as (14.6), or numerically by direct inversion of (14.8) (Clauer, 1986). Typically, the dynamics is unknown and the response function needs to be solved directly from experimental time series data. For a time series of length N, equation (14.8) is written

$$Dst(t) = \sum_{i=0}^{T} H(i\Delta t)\, V_{\mathrm{SW}}(t - i\Delta t)\, B_{\mathrm{South}}(t - i\Delta t) \tag{14.9}$$

where we keep track of the solar wind history up to a time T, and Δt is the time resolution (Iyemori *et al.*, 1979). The memory time T corresponds to the longest time scale in $V_{\mathrm{SW}}B_{\mathrm{South}}$ that can determine Dst. Inverting (14.9) as a multi-linear regression (Press *et al.*, 1992) gives

$$H = [VB_s VB_s^T]^{-1}[Dst VB_s^T] \tag{14.10}$$

where Dst is an N-dimensional column vector (the Dst time series) and $V_{\mathrm{SW}}B_{\mathrm{South}}$ is an N xT matrix obtained from the corresponding $V_{\mathrm{SW}}B_{\mathrm{South}}$ measurements. Here the superscript T denotes the transpose of a matrix. The impulse response function of Dst has a peak at $\tau = 1$ hour, followed by a rapid decrease to zero.

FIR models have been useful in analysis of geomagnetic and particle-flux time variations. FIR models have been developed for various geomagnetic indices and solar wind/IMF inputs (Bargatze *et al.*, 1985; Clauer, 1986; Trattner and Rucker, 1990) and the relativistic electron flux at geosynchronous orbit (Nagai, 1988; Baker *et al.*, 1990) as well as other inner-magnetospheric regions (Vassiliadis *et al.*, 2002). The Baker *et al.* model is used at NOAA/SEC as a forecasting tool (*http://www.sec.noaa.gov/refm/doc/REFMDoc.html*).

14.3.2 Multi-input models and input ranking

As mentioned, input–output modelling is necessary for geospace weather modelling, and as models become more realistic the number of input variables increases (for examples of Dst, see Trattner and Rucker, 1989; Temerin and Li, 2002). Identification of the most relevant input variables is based on the physics of the environment, information–theoretical criteria, and practical considerations (such as availability or reliability of inputs). In magnetospheric and ionospheric applications, the most relevant solar and interplanetary inputs include the IMF components (primarily B_z) and absolute magnitude, the solar wind radial speed and density, UV radiation, and solar energetic particle fluxes. Even for a single interaction, there are several ways to approximate the effective input function as a parametrization of interplanetary variables (Gonzalez *et al.*, 1990). Depending on the particular applications, the order of these input variables may be different; also, other variables may need to be added to the list. Each input represents a different type of coupling between solar/interplanetary energy sources and the geospace sinks. The primary interactions are magnetic reconnection, viscous interaction, radiation forcing, and so on.

Solar and interplanetary activity is highly correlated during the development and propagation of geoeffective structures; for example, high-speed streams (McPherron *et al.*, 2004). Therefore, few variables can be considered as mutually independent. Magnetic field components, plasma flows and energetic particle fluxes follow characteristic time variations in the formation and propagation of geoeffective structures such as interplanetary shocks, coronal mass ejections and streams. This poses the question: what are the most significant inputs for a particular space environment?

We can assess a geoeffectiveness of a solar/interplanetary input, or of a given type of structure, by a metric such as the correlation coefficient (14.7) (to be discussed in some detail in Section 14.5). The dependence of the geoeffectiveness function on location or particle energy can be used to identify the dominant physical processes. A recent study compared the geoeffectiveness of solar, interplanetary and magnetospheric variables in determining the high-energy electron flux in the radiation belts (Vassiliadis *et al.*, 2005). The electron flux is determined by a number of processes such as viscous interaction and excitation of ULF waves which leads to enhanced radial diffusion, magnetic reconnection which leads to injections and/or *in situ* acceleration depending on the magnetospheric location, ionospheric and plasmaspheric loss effects, and so on. Figure 14.3 (see colour section) shows a comparison of the geoeffectiveness of seventeen such variables as a function of L shell for fixed electron energy. Here the geoeffectiveness is represented by the correlation coefficient between model predictions and observations of the electron flux $j_e(t;L)$:

$$C_{(\hat{j}_e,j_e)}(L) = \frac{1}{T}\frac{1}{\sigma_{\hat{j}_e}\sigma_{j_e}}\int_0^T (\hat{j}_e(t;L) - \bar{\hat{j}}_e(L))(j_e(t;L) - \bar{j}_e(L))\,dt$$

similar to equation (14.7). In the plot, variables have been grouped together according to the similarity of the geoeffectiveness profiles and indicate three types of coupling: (a) hydrodynamic (including viscous) interactions; (b) magnetic reconnection and related geomagnetic index activity; and (c) electron-loss-enhancing (ionosphere/plasmasphere expansion) or driver-mitigating processes. Note that the profiles differ depending on the L shell. Thus we can select the most important input variables for each L range of interest in the radiation belts. Furthermore, detailed structure identification is possible from the observed time series of fields, flows or fluxes. A superposed-epoch method has been developed that can give the average time profile of a high-speed stream in velocity, field, and other variables at a stream interface (McPherron *et al.*, 2004). The time profiles have been used as templates in comparison with datastreams and automatic identification of stream interfaces.

14.3.3 Feedback and non-linearity

These two properties of plasma systems are distinct but closely related, and in actuality they often occur together. Feedback is generally associated with instabilities (linear or non-linear).

14.3.3.1 Feedback

Plasma instabilities are prime examples of positive feedback mechanisms. Instabilities lead to rapid (exponential for linearized dynamics) growth in system energy. We have already seen a case of negative feedback in model (14.6) in terms of the ring current loss term. A more general form of (14.6) including several feedback and coupling terms is:

$$Dst(t) = \sum_{i=1}^{m} a_i Dst, i(t - i\Delta t) + \sum_{j=0}^{l} b_j \cdot V_{\text{SW}}(t - j\Delta t)\, B_{\text{South}}(t - j\Delta t) \qquad (14.11)$$

14.3.3.2 Non-linearity

In a plasma system, non-linearity can be viewed as a symmetry breaking; namely, of the invariance of the dynamics for different levels of activity. Typically the breaking occurs for extreme (much higher or lower than average) levels of activity. There are many ways in generalizing a linear system such as (6) to a non-linear system. Below we discuss three of them.

14.3.3.3 Coupling coefficients are functions of the input

In those cases where the system is driven by an external source, a direct way to make the model dynamics non-linear is to parametrize the dynamics in terms of the input level.

For instance, in a non-linear version of the Dst model (14.6), the loss rate τ has been modelled as a function of solar wind input $V_{\text{SW}} B_{\text{South}}$ (O'Brien and McPherron, 2000). The coefficients b and τ are obtained by fitting to historical data of Dst and $V_{\text{SW}} B_{\text{South}}$.

In a modern and flexible methodology, one approximates the form of the non-linearity by a neural network or other iterative approach. A simple network, such as a multilayer perceptron, has a hierarchical input–output structure where the outputs of layer n are the inputs to layer $n + 1$. The inputs of the first layer are the inputs to the system (in this case $V_{\text{SW}} B_{\text{South}}(t)$ and lags thereof) and the output of the last layer is the system output (in this case, the geomagnetic activity at time t, $Dst(t)$) (Wu and Lundstedt, 1996, 1997). The i-th element (neuron) of the n-th layer is determined as follows:

$$x_i^{(n+1)} = g\left(\sum_j w_{ij}^{(n+1)} x_j^{(n)} - \mu_i^{(n+1)} \right) \qquad (14.12)$$

where the activation function $g(.)$ maps the weighted sum of the layer inputs to the output, $x_i^{(n+1)}$. The function $g(.)$ is non-linear, such as the hyperbolic tangent and more generally a radial basis function. In order to calculate the activation functions for a given network architecture, one uses iterative methods such as back-propagation (Hertz *et al.*, 1991) rather than the simple regression used to solve (14.9). Neural networks have been used in several applications in radiation-belt modelling (Koons and Gorney, 1991; Wu *et al.*, 2000) and geomagnetic modelling (Gleisner *et al.*, 1997; Gavrishchaka and Ganguli, 2001; Sutcliffe, 2001).

14.3.3.4 Local phase-space dynamics

So far we have discussed ‘global’ methods because the same functional form applies to the entire phase space. Sparse measurements and, more generally, non-uniform coverage of the phase space can hamper the effectiveness of these methods. Neural networks with two or more layers can reproduce any smooth function, but the number of data needed rises rapidly with non-linear features. Under those conditions it is more useful to develop local models, or ‘maps’ for individual neighbourhoods of the phase space, and then combine them in an ‘atlas’ model. A local-linear model has the form similar to (14.11):

$$Dst(t) = \sum_{i=0}^{m} a_i^{(NN)} Dst, i(t - i\Delta t) + \sum_{j=0}^{l} b_j^{(NN)} \cdot V_{\mathrm{SW}}(t - j\Delta t)\, B_{\mathrm{South}}(t - j\Delta t) \quad (14.13)$$

but here the coefficients $s_i^{(NN)}$, $b_j^{(NN)}$ are determined from a small subset of Dst measurements, which have the same recent history as the current state whose future we want to predict, rather than the entire dataset (Vassiliadis *et al.*, 1999). The number NN of the historic measurements used in each incarnation of (14.13) represents the degree of non-linearity in the local model: if it goes to infinity (or the size of the entire database), we recover the linear model (14.11). Comparisons between local models and linear or non-linear FIR models of the type (14.9) show that the former are much more accurate (Vassiliadis *et al.*, 1995).

14.3.3.5 Modelling subsystems

A third way to introduce non-linear couplings is by identifying physical subsystems, modelling each one individually, and then synthesizing the individual forecasts to estimate the activity of the entire system.

The Dst index, for instance, is determined by the symmetric ring-current time variations, but also those of the asymmetric ring current, and the magnetopause and tail currents (including the substorm current wedge). In the Burton *et al.* (1975) model for Dst (14.6), the model of the ring current magnetic signature was built from separate fits to the data for active ($V_{\mathrm{SW}} B_{\mathrm{South}} \neq 0$) and quiet conditions. The geomagnetic effect of the magnetopause current was modelled separately. In a more comprehensive treatment, the geomagnetic effects of five currents contributing to the index were modelled separately based on driving by the solar wind and IMF (Temerin and Li, 2002), using intrinsic growth and decay times with models that resemble equation (14.6). In that model, the estimated Dst activity was the sum total of the geomagnetic effects of all these currents, but one can envision similar models with strong coupling between subsystems.

14.3.4 Higher dimensions

While indices of activity, when accurately calculated, are invaluable for space weather, since they describe succinctly the global or regional level of activity, users are typically interested in local forecasts. Therefore, field models have been developed from either empirical or primarily physical models.

14.3.4.1 Parametrization by spatial scales

A linear FIR model of the log-flux (logarithm of the flux) j_e which is parametrized by the L shell has the form:

$$j_e(t; L = \text{const.}) = \int_0^\infty H(\tau; L = \text{const.})\, V_{\text{SW}}(t - \tau)\, d\tau \tag{14.14}$$

Note that the log-flux and the impulse response function are parametrized by L. In addition, the response lag takes both negative and positive values. Negative lags are acausal, and indicate that the flux increases are due to other interplanetary (such as IMF) or internal (for example, ionospheric) activity than V_{SW}. If that activity is uncorrelated with V_{SW}, the impulse response is zero for $\tau < 0$; otherwise peaks at negative lags will indicate where the correlation of the other activity is highest and can be used to identify the source of the activity.

The response $H(\tau; L)$ has three prominent peaks: two positive, P_0 and P_1, and one negative, V_1 (Figure 14.4, colour section). Peak P_1 occurs in the region $L = 4$–7.5 including the geosynchronous orbit. The flux in P_1 responds coherently to solar wind velocity increases such as those in high-speed streams. Peak P_0, on the other hand, occurs at $L = 3$–4. The flux corresponds to increases in V_{SW}, but also to increases of IMF B_z, which produces dayside reconnection, followed by enhanced convection and rapid injections deeper in the magnetosphere than P_1. This response is excited by the passage of CMEs (including magnetic clouds and other ejecta) rather than high-speed streams. Therefore, the peaks represent two different modes of response of the inner magnetosphere (Vassiliadis *et al.*, 2003). The negative peak, V_1, occurs in the same spatial region as P_1, but earlier in time. It corresponds to the quasiadiabatic diffusion and loss brought about by the variation of the ring current close to $L = 5.5$.

The response of the flux as a function of L shell thus shows a complex dynamics, which characterizes the radiation belts. In addition to the solar wind velocity, other interplanetary and magnetospheric variables are important in determining the radiation belt electron flux (Vassiliadis *et al.*, 2005).

14.3.4.2 Kinetic models

A recent quasi-empirical diffusion model of the electron flux (Li *et al.*, 2001) has its starting point at the diffusion and loss of the electron phase space density:

$$\frac{\partial f_e(t; L)}{\partial t} = L^2 \frac{\partial}{\partial L}\left(\frac{D_{LL}}{L^2}\frac{\partial f_e}{\partial L}\right) - \frac{f_e}{\tau_{\text{loss}}} \tag{14.15}$$

The phase space density is defined in terms of the observed flux J as $f = J/p^2$, where p is the relativistic momentum of the particles. Diffusion occurs as electrons interact with low-frequency waves whose power is significant in the region close to and below geosynchronous orbit. Importantly, parameters D_{LL} and τ_{loss} are functions of time-dependent solar wind velocity and IMF B_z, as well as of other interplanetary inputs (Li *et al.*, 2001). An example of the model's prediction capability is shown in Figure 14.5 (see colour section). Equation (14.15) represents the daily flux at geosynchronous orbit at high fidelity (see *http://lasp.colorado.edu/space_weather/xlf3/xlf3.html*).

A complementary approach is to investigate the effects of a specific interplanetary parameter on the energetic electron flux at different altitudes. Probably the single most significant interplanetary parameter for electron flux levels is the solar wind bulk speed V_{SW}. Its increases are typically followed by increases in the relativistic electron flux at geosynchronous orbit, as a simple correlation shows (Paulikas and Blake, 1979). The relation between geosynchronous flux and interplanetary velocity was further refined by Baker *et al.* (1990), who used a linear filter (14.8) approach to show that the response lagged by 2–3 days on average, which is the time it takes for ULF waves to develop (Rostoker *et al.*, 1998) and electron diffusion to occur (Li *et al.*, 2001), as seen above. However, diffusion alone does not explain the flux variability at other L shells.

14.3.4.3 MHD models

As large-scale simulations have become faster and more efficient, they too are used increasingly for (near) real-time modelling of space environments. Heliospheric and magnetospheric simulations involve multiple, and in some cases adaptive, grids to solve the partial differential equations of magnetohydrodynamics. Boundary conditions (solar surface, solar wind, or ionosphere) are time-dependent and realistic. Kinetic effects such as the ring current and conductivities obtained from experiment or empirical modelling are also included in some simulations. Notable models include the MAS inner heliospheric model (Linker *et al.*, 1999), the Enlil inner/mid-heliospheric model (Odstrcil and Pizzo, 1999), the LFM magnetospheric model (Lyon *et al.*, 2001), the BATS-R-US adaptive-grid comprehensive helio-spheric/magnetospheric model (Groth *et al.*, 2004; Song *et al.*, 1999), and many others. Magnetospheric models are used to drive test-particle and/or diffusion models for the radiation belts and the ring current; for instance, the Comprehensive Ring Current Model (Fok *et al.*, 1995). The output of a such a radiation belt electron model by Fok *et al.*, in terms of the electron phase space density, is shown in Figure 14.6 (see colour section).

14.4 DATA ASSIMILATION AND KALMAN FILTERING

Since the mid-1990s, real-time space environment data has become publicly available on the Internet. In actuality, it takes about 5–20 minutes for *in situ* magnetospheric and solar wind measurements to be telemetred to Earth, processed, and publicly made available. This is a relatively small delay compared to substorm timescales (0.5–3 hours), or storm timescales (tens of hours/days). The data availability enabled the development of models that could ingest and use the measurements to improve forecast quality.

As a result of these developments, data assimilation has become possible. Data assimilation is the insertion of measurements into models with the aim of improving forecast quality. There are several methods for implementing data assimilation, a central one being the Kalman filter. Below we briefly discuss the discrete Kalman filter, which is the optimal state estimator for systems with linear dynamics. Its extension to non-linear systems (extended Kalman filter), is quite analogous in structure, but beyond the scope of this chapter.

14.4.1 The Kalman filter

We consider a space environment with dynamics given by:

$$x_k = Ax_{k-1} + Bu_{k+1} + w_{k-1} \tag{14.16}$$

where A is the system dynamics and B is the coupling to an external source u_k. For simplicity the time is assumed discrete with time resolution $\Delta t = 1$. Except for the noise term w_{k-1}, equation (14.16) is analogous to (14.6). The state x_k cannot be observed directly, but instead measurements y_k are obtained via a measurement function H:

$$y_k = Hx_k + v_k \tag{14.17}$$

The noise terms for the system dynamics, w_k, and for the observation, u_k, are normally distributed with zero mean and covariances Q and R, respectively.

Our estimate of the state is $\hat{x}_k$, so its error is $e_k = \hat{x}_k - x_k$. The goal is to improve our estimate $\hat{x}_k$ of the space environment state using the available measurements y_k; more specifically, to reduce the uncertainty in $\hat{x}_k$, expressed as the error covariance:

$$P_k = E[e_k e_k^T]$$

The Kalman filter is an optimal solver for this problem. It consists of a correction to the state estimate

$$\hat{x}_k \leftarrow \hat{x}_k + K(y_k - H\hat{x}_k^-) \tag{14.18}$$

where the estimate is adjusted by the difference between the actual measurement y_k and the prediction $H\hat{x}_k$. The difference is modulated by the Kalman gain, K. By minimizing the error covariance P_k we derive the gain:

$$K_k = \frac{P_k H^T}{HP_k H^T + R}$$

The estimate state and its covariance can then be written using (14.18):

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_k + K_k(y_k - H\hat{x}_k^-)$$

$$P_k = (1 - K_k H)(AP_{k-1}A^T + Q)$$

14.4.2 Parameter estimation in a radiation-belt model

The Kalman filter can be adapted to a wide variety of space weather environments and applications such as parameter estimation. As we have seen, a key space-weather region is the inner magnetosphere, and more specifically the electron radiation belts. Real-time measurements from geostationary satellites are readily available. An early study showed the potential of these measurements for data assimilation (Moorer and Baker, 2000).

Rigler *et al.* (2004) examined the variability of the radiation-belt state, and explored ways to describe it in a Kalman filter formalism. In that application the system state was defined as the response to an input. Therefore, obtaining the state is tantamount to parameter estimation. Rigler *et al.* (2004) rewrote equation (14.14) as

$$j_{e,t} = \hat{H}_t \cdot V_{\mathrm{SW},t}$$

where $j_{e,t}$ is the electron log-flux at time t, produced by the solar wind-magnetosphere coupling $\hat{H}_t$ (a vector; the hat indicates that it is estimated rather than observed) in a dot product with the lag vector $V_{\mathrm{SW},t}$ (see equation (14.2)). The Kalman filter for the log-flux system is:

$$\begin{aligned} K_t &= \frac{P_t V_{\mathrm{SW},t}^T}{V_{\mathrm{SW},t} P_t V_{\mathrm{SW},t}^T + R} \\ \hat{H}_{t+1} &= \hat{H}_t + K_t (j_{e,t} - \hat{H}_t V_{\mathrm{SW},t}) \\ P_t &= (1 - K_t V_{\mathrm{SW},t})(P_{t-1} + Q) \end{aligned} \tag{14.9}$$

The model $\hat{H}_t$, estimated from equation (14.19), differs substantially from the average (least-squares) FIR model H of equation (14.14) (Rigler *et al.*, 2004). Differences exist both for high-activity intervals (passages of high-speed streams, CMEs, and so on) and for quiescent intervals (such as, prolonged northward B_z). At the same time, the time-dependent model gives substantially lower errors for the one-step-ahead forecasts (Figure 14.7, colour section). The prediction efficiency increases dramatically, and in a low-L-shell region ($\sim$3) it exceeds 75% both for single-input and multi-input models (Rigler *et al.*, 2004). The reduction of prediction error is expected, since the model flexibly adapts to changing solar wind conditions; however, the amount of error reduction is highly significant. Extensions of this approach to more complex systems such as diffusion models is ongoing (J. Koller, private communication).

14.4.3 Ionospheric data assimilation

Significant progress has already been made in ionospheric modelling using data assimilation techniques with several ionospheric weather applications.

One of the longest-running efforts has been the high-latitude model for Assimilative Mapping of Ionospheric Electrodynamics (AMIE) (Kamide *et al.*, 1981; Richmond and Kamide, 1988). The AMIE algorithm ingests radar, satellite and ionosonde measurements of electric fields and densities, ground magnetometer measurements of magnetic fields, and other datastreams. More recently this effort has developed into a real-time model, rtAMIE (Ridley *et al.*, 2000).

The Global Assimilation of Ionospheric Measurements (GAIM) effort, led by the University of Utah, has produced a physics-based ionosphere–plasmasphere model which is adjusted in real-time using data assimilation (Schunk *et al.*, 2003). The assimilation model is a Kalman filter which ingests a diverse set of measurements: total electron content (TEC), ionosonde electron density profiles, occultation data, tomography chain data, line-of-sight UV emissions, and so on (Sojka *et al.*, 2001; Schunk *et al.*, 2004). The output of the model is the electron density distribution from 90 km to 35,000 km (geosynchronous orbit), as well as several other derived electron parameters. In its specification mode, GAIM accurately reconstructs the ionospheric electron density.

14.5 MODEL VERIFICATION

After a space weather model has completed a stage of development, it undergoes tests so that its accuracy and forecast potential can be evaluated. The two main testing stages are validation and verification. In validation we measure how well a model does what it is programmed to do, including how numerically accurate it is; and in verification we compare the model to actual measurements (Doggett, 1996). The rest of this section discusses the basics of verification.

Model prediction accuracy is measured via statistical functions, or metrics, in either an absolute or a relative (comparative) way. A standard absolute metric is the linear correlation coefficient between observed activity x_i and model prediction $\hat{x}_i$:

$$C_{\hat{x},x} = \frac{1}{N}\frac{1}{\sigma_{\hat{x}}\sigma_x}\sum_{i=1}^{N}(\hat{x}_i - \langle\hat{x}\rangle)(x_i - \langle x\rangle) \tag{14.20}$$

where N is the size of the forecast sample, and σ_i represents the corresponding standard deviations for observations and forecasts. (We encountered an application of (14.20) for Dst in equation (14.7).)

A more detailed diagnostic than the correlation coefficient is the prediction error, or the difference between model forecast and the actual system activity. Usually this is a root-mean-square error,

$$e_{\mathrm{rms}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}$$

which is compared with the sample standard deviation of the observations, $e_{\mathrm{rms}}/\sigma_x$. An example of the normalized prediction error as a function of the model's nonlinearity and number of degrees of freedom is shown in Figure 14.8.

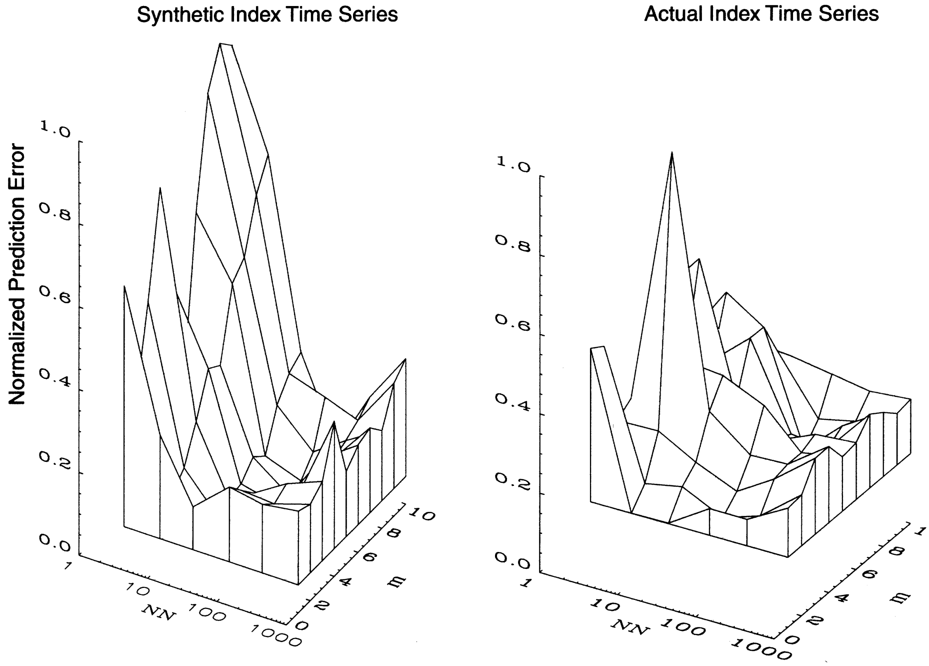


Figure 14.8. Model optimization using the normalized prediction error for AL geomagnetic index data. (Left) actual data; (right) synthetic data from the Klimas *et al.* (1994) model. The error is plotted versus the number of neighbours (NN) representing non-linearity, and the model order (after Vassiliadis *et al.*, 1995).

A third, often-used, diagnostic is the prediction efficiency – the percent variance of the observations that can be explained by the model forecasts:

$$PE = 1 - \frac{1}{N\sigma_x^2}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 = 1 - \frac{e_{\mathrm{rms}}^2}{\sigma_x^2} \tag{14.21}$$

When different models are developed, it becomes increasingly important to express the accuracy of one model versus that of another. The skill of a model with respect to a reference model is the accuracy of its forecasts relative to the accuracy of the reference model forecasts. The relative accuracy is expressed by metrics called skill scores. A standard skill score is a function of the root-mean-square error of the model compared to a similar function of the error of a reference model:

$$SS = 1 - \frac{e_{\mathrm{rms}}^2}{(e_{\mathrm{rms}}^{(\mathrm{ref})})^2} \tag{14.22}$$

Simple but important reference model are the persistence model (for example, equation (14.4) with $\varepsilon = 0$), periodic or recurrent models, and climatology models. In the special case when the climatology model is the average activity $\bar{x}$, then the above skill score becomes the prediction efficiency (14.21).

14.6 SUMMARY AND OUTLOOK

While the diversity of space weather environments is high, models and forecasts share many similar characteristics. A key consideration in space weather modelling is that these environments are open systems exchanging mass, momentum and energy with their environment. Representing all physically relevant, independent inputs correctly is necessary for capturing the complete variability of the system. The actual development of accurate and reliable forecast models relies on the physics of the environment and information–theoretical principles. Physical and empirical models are useful elements of the description of a space weather environment, as the example of Dst prediction has shown.

14.6.1 Forecast providers

Modern space weather forecasts are provided by a variety of government, academic and private sources. The distribution of providers affects the establishment of forecast standards, protocols, and the overall efficiency of the information transfer.

Historically, national and continental agencies have provided the most comprehensive nowcast and forecast suites. These include NOAA's Space Environment Center, ESA's Space Weather Services, Japan's Space Environment Information Service, and Australia's IPS Solar and Space Services among others. China's National Space Administration has actively evolved plans for space weather monitoring. Specialist providers – almost exclusively of nowcasts rather than forecasts – are the Los Alamos National Laboratories supplying real-time geosynchronous energetic particle fluxes, NASA centres (such as, Marshall SFC's Polar/UVI group providing near-real-time data and images; the Space Weather Bureau which focuses both on research and on education and public outreach; Goddard SFC's solar and magnetospheric/ionospheric research groups; JPL's Ionospheric and Atmospheric Remote Sensing), NCAR/HAO, UCAR (the Space Physics and Aeronomy Research Collaboratory), and others.

Academic and private providers offer more specialized nowcasts and forecasts. Academic space weather providers include the Universities of Michigan, Iowa, Alaska, Rice University, NJIT, and others. In addition, private companies have entered the field, ranging from aerospace leaders such as Lockheed–Martin's research laboratories and JHU's APL, to small ventures such as Solar–Terrestrial Dispatch and Space Environment Technologies.

Today, space weather forecasting is in a stage of rapid transition. A number of physics and information–theoretical techniques are being developed, while new spacecraft missions, with orders-of-magnitude larger datastreams than the current ones, are anticipated to significantly elevate both the prediction accuracy and physical understanding of the near-Earth space environment.

14.7 ACKNOWLEDGEMENTS

Work on this chapter was supported by grants NASA/LWS NAG5-12431, NSF/ATM NSWP-0418857, and NSF/CISM. The author thanks D. Baker, I. Daglis, A. Klimas, B. McPherron, A. Pulkkinen, J. Rigler and B. Weigel for discussions on modelling and forecasting.

14.8 REFERENCES

Arge, C.N., J.G. Luhmann, D. Odstrcil, C.J. Schrijver, and Y. Li, Stream structure and coronal sources of the solar wind during the May 12th, 1997 CME, *J. Atmos. Sol.–Terr. Phys.*, **66**(15–16), 1295–1309, 2004.

Baker, D.N., and J.G. Kappenman, Uninterrupted electric power, *Science*, **273**, 5272, 165–168, 1996.

Baker, D.N., R.L. McPherron, T.E. Cayton, and R.W. Klebesadel, Linear prediction filter analysis of relativistic electron properties at 6.6 RE, *J. Geophys. Res.*, **95**, A9, 15,133–15,140, 1990.

Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones, Jr., Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res.*, **90**, A7, 6,387–6,394, 1985.

Behnke, R., and T.F. Tascione, *The National Space Weather Program: The Strategic Plan*, Office of the Federal Coordinator for Meteorological Services and Supporting Research, FCM-P30-1995, Washington, DC, 1995.

Behnke, R., *et al.*, *The National Space Weather Program: The Implementation Plan*, Office of the Federal Coordinator for Meteorological Services and Supporting Research, FCM-P31-1997, Washington, DC, 1997.

Boscher, D., S. Bourdarie, and T. Beutier, Dynamic modeling of trapped particles, *IEEE Trans. Nuclear Sci.*, **43**, 416–425, 1996.

Bourdarie, S., R.H.W. Friedel, J. Fennnell, S. Kanekal, and T.E. Cayton, Radiation belt representation of the energetic electron environment: Model and data synthesis using the Salammbo radiation belt transport code and Los Alamos geosynchronous and GPS energetic particle data, *Space Weather*, **3**, 4, Art. No. S04S01, 2005.

Burton, R.K., R.L. McPherron, and C.T. Russell, Empirical relationship between interplanetary conditions and Dst, *J. Geophys. Res.*, **80**, 31, 4204–4214, 1975.

Clauer, C.R., The technique of linear prediction filters applied to studies of solar wind–magnetosphere coupling, in: *Solar Wind–Magnetosphere Coupling*, Y. Kamide and J.A. Slavin (eds.), 39–57, Terra Scientific, 1986.

Crooker, N.U., Solar and heliospheric geoeffective disturbances, *J. Atm. Sol.–Terr. Phys.*, **62**, 12, 1071–1085, 2000.

Daglis, I.A., *Space Storms and Space Weather Hazards*, Kluwer, Dordrecht, Boston, 2001.

Doggett, K. (ed.), *The Evaluation of Space Weather Forecasts. Proceedings of a Workshop at Boulder, Colorado, June 19–21, 1996*, NOAA/SEC, Boulder, 1996.

Ebihara, Y., M.-C. Fok, S. Sazykin, M.F. Thomsen, M.R. Hairston, D.S. Evans, F.J. Rich, and M. Ejiri, Ring current and the magnetosphere–ionosphere coupling during the superstorm of 20 November 2003, *J. Geophys. Res.*, **110**, A9, Art. No. A09S22, 2005.

Fedder, J.A., and J.G. Lyon, The Earth's magnetosphere is 165 RE long: self-consistent currents, convection, magnetospheric structure, and processes for northward interplanetary magnetic field, *J. Geophys. Res.*, **100**, 3623, 1995.

Fok, M.-C., T.E. Moore, J.U. Kozyra, G.C. Ho, and D.C. Hamilton, 3-dimensional ring current decay model, *J. Geophys. Res.*, A6, 9619–9632, 1995.

Gavrishchaka, V.V., and S.B. Ganguli, Optimization of the neural-network geomagnetic model for forecasting large-amplitude substorm events, *J. Geophys. Res.*, A106, 4, 6247–6257, 2001.

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

Gonzalez, W.D., A unified view of solar wind-magnetosphere coupling functions, *Planet. Space Sci.*, **38**, 5, 627–632, 1990.

Groth, C.P.T., D.L. DeZeeuw, T.I. Gombosi, and K.G. Powell, Global three-dimensional MHD simulation of a space weather event: CME formation, interplanetary propagation, and interaction with the magnetosphere, *J. Geophys. Res.*, **105**, A11, 25053–25078, 2000.

Hertz, J., A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, 1991.

Heynderickx, D., Radiation belt modelling in the framework of space weather effects and forecasting, *J. Atm. Sol.–Terr. Phys.*, **64**, 16, 1687–1700, 2002.

Horton, W., and I. Doxas, A low-dimension energy-conserving state space model for substorm dynamics, *J. Geophys. Res.*, **101**, A12, 27,223–27,237, 1996.

Horton, W., and I. Doxas, A low-dimensional dynamical model for the solar wind driven geotail-ionosphere system, *J. Geophys. Res.*, **103**, A3, 4561–4572, 1998.

Iyemori, T., H. Maeda, and T. Kamei, Impulse-response of geomagnetic indexes to interplanetary magnetic field, *J. Geomag. Geoelectr.*, **31**, 1, 1–9, 1979.

Kamide, Y., A.D. Richmond, and S. Matsushita, Estimation of ionospheric electric-fields, ionospheric currents, and field-aligned currents from ground magnetic records, *J. Geophys. Res.*, **86**, NA2, 801–813, 1981.

Kappenmann, J.G., W.A. Radasky, J.L. Gilbert, and I.S. Erinmez, Advanced geomagnetic storm forecasting: A risk management tool for electric power system operations, *IEEE Trans. Plasma Sci.*, **28**, 6, 2114–2121, 2000.

Klimas, A.J., D.N. Baker, D. Vassiliadis, and D.A. Roberts, Substorm recurrence during steady and variable solar-wind driving – evidence for a normal mode in the unloading dynamics of the magnetosphere, *J. Geophys. Res.*, **99**, A8, 14,855–14,861, 1994.

Koons, H.C., and D.J. Gorney, A neural network model of the relativistic electron flux at geosynchronous orbit, *J. Geophys. Res.*, **96**, 5549–5556, 1991.

Li, X., M. Temerin, D.N. Baker, G.D. Reeves, and D. Larson, Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements, *Geophys. Res. Lett.*, **28**, 9, 1887–1890, 2001.

Lindsay, G.M., C.T. Russell, and J.G. Luhmann, Predictability of Dst index based upon solar wind conditions monitored inside 1 AU, *J. Geophys. Res.*, **104**, A5, 10,335–10,344, 1999.

Linker, J., Z. Mikic, D.A. Biesecker, R.J. Forsyth, W.E. Gibson, A.J. Lazarus, A. Lecinski, P. Riley, A. Szabo, and B.J. Thompson, Magnetohydrodynamic modeling of the solar corona during whole sun month, *J. Geophys. Res.*, **104**, 9809, 1999.

Lyon, J.G., R.E. Lopez, M.J. Wiltberger, and C.C. Goodrich, Comparison of global MHD simulation results with actual storm and substorm events, in: Validation of magnetospheric models, *Adv. Space Res.*, **28**, 12, 1701–1706, 2001.

McPherron, R.L., G. Siscoe, and N. Arge, Probabilistic forecasting of the 3-h ap index, *IEEE Trans. Plasma Sci.*, **32**, 4, 1,425–1,438, Part 1, 2004.

Moorer, D.F., and D.N. Baker, Specification of energetic magnetospheric electrons and ions, paper presented in: Chapman conference on Space Weather: Progress and challenges in research and applications, Clearwater, Florida, March 20–24, 2000.

Nagai, T., 'Space weather forecast': prediction of relativistic electron intensity at synchronous orbit, *Geophys. Res. Lett.*, **15**, 425–428, 1988.

O'Brien, T.P., and R.L. McPherron, An empirical phase space analysis of ring current dynamics: Solar wind control of injection and decay, *J. Geophys. Res.*, **105**, A4, 7707–7719, 2000.

Odstrcil, D., and V.J. Pizzo, Three-dimensional propagation of coronal mass ejections (CMEs) in a structured solar wind flow, 1, CME launched within the streamer belt, *J. Geophys. Res.*, **104**, 483, 1999.

Papitashvili, V.O., C.R. Clauer, T.L. Killeen, B.A. Belov, S.A. Golyshev, and A.E. Levitin, Linear modeling of ionospheric electrodynamics from the IMF and solar wind data: Application for space weather forecast, Solar–Terrestrial Relations: Predicting the effects on the near-Earth environment, *Adv. Space Res.*, **22**, 1, 113–116, 1998.

Paulikas, G.A., and J.B. Blake, Effects of the solar wind on magnetospheric dynamics: energetic electrons at the sunchronous orbit, 180-, in *Quantitative Modeling of Magnetospheric Processes*, W.P. Olson (ed.), Geophysical Monograph 21, American Geophysical Union, Washington, DC, 1979.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.V. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Second edition, Cambridge University Press, Cambridge, 1993.

Richardson, I.G., E.W. Cliver, and H.V. Cane, Sources of geomagnetic storms for solar minimum and maximum conditions during 1972–2000, *Geophys. Res. Lett.*, **28**(13), 2569–2572, 2001.

Richmond, A.D., and Y. Kamide, Mapping electrodynamic features of the high-latitude ionosphere from localized observations – technique, *J. Geophys. Res.* A93, **6**, 5741–5759, 1988.

Ridley, A.J., G. Crowley, and C. Freitas, An empirical model of the ionospheric electric potential, *Geophys. Res. Lett.*, **27**, 22, 3675–3678, 2000.

Rostoker, G., S. Skone, and D.N. Baker, On the origin of relativistic electrons in the magnetosphere, *Geophys. Res. Lett.*, **25**, 3701–3704, 1998.

Schunk, R.W., L. Scherliess, and J.J. Sojka, Recent approaches to modeling ionospheric weather, in *Space Weather 2000, Adv. Space Res.*, **31**, 4, 819–828, 2003.

Schunk, R.W., L. Scherliess, J.J. Sojka, and D.C. Thompson, Development of a physics-based reduced state Kalman filter for the ionosphere, *Radio Sci.*, **39**, 1, Art. No. RS1S04, 2004.

Sojka, J.J., D.C. Thompson, R.W. Schunk, T.W. Bullett, and J.J. Makela, Assimilation ionosphere model: Development and testing with combined ionospheric campaign caribbean measurements, *Radio Sci.*, **36**, 2, 247–259, 2001.

Song, P., D.L. DeZeeuw, T.I. Gombosi, C.P.T. Groth, and K.G. Powell, A numerical study of solar wind–magnetosphere interaction for northward interplanetary magnetic field, *J. Geophys. Res.*, **104**, A12, 28,361–28,378, 1999.

Sutcliffe, P.R., The development of a regional geomagnetic daily variation model using neural networks, *Ann. Geophys.*, **18**, 1, 120–128, 2000.

Temerin, M., and X.L. Li, A new model for the prediction of Dst on the basis of the solar wind, *J. Geophys. Res.*, **107**, A12, Art. No. 1472, 2002.

Trattner, K.J., and H.O. Rucker, Linear prediction-theory in studies of solar wind–magnetosphere coupling, *Ann. Geophys.*, **8**, 11, 733–738, 1990.

Tsyganenko, N.A., A Magnetospheric Magnetic Field Model with a Warped Tail Current Sheet, *Planet. Space Sci.*, **37**, 5–20, 1989.

Tsyganenko, N.A., Modeling the Earth's magnetospheric magnetic field confined within a realistic magnetopause, *J. Geophys. Res.*, **100**, A4, 5599–5612, 1995.

Tsyganenko, N.A., and M.I. Sitnov, Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms, *J. Geophys. Res.*, **110**, A3, Art. No. A03208, 2005.

Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, A description of the solar wind–magnetosphere coupling based on nonlinear filters, *J. Geophys. Res.*, **100**, A3, 3495–3512, 1995.

Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, The nonlinearity of models of the vBSouth–AL coupling, *J. Geophys. Res.*, **101**, A9, 19779–19787, 1996.

Vassiliadis, D., A.J. Klimas, J.A. Valdivia, and D.N. Baker, The Dst geomagnetic response as a function of storm phase and amplitude and the solar wind electric field, *J. Geophys. Res.*, **104**, A11, 24957–24976, 1999.

Vassiliadis, D., System identification, modeling, and predictions for space weather environments, *IEEE Trans. for Plasma Sci.*, **28**, 6, 1944–1955, 2000.

Vassiliadis, D., A.J. Klimas, J.A. Valdivia, and D.N. Baker, The nonlinear dynamics of space weather, *Adv. Space Res.*, **26**, 1, 197–207, 2000.

Vassiliadis, D., A.J. Klimas, B.-H. Ahn, R.J. Parks, A. Viljanen, and K. Yumoto, High-latitude electrodynamics from a multi-array nonlinear geomagnetic model, in: *Space Weather Study using Multipoint Techniques*, L.-H. Lyu (ed.), Pergamon, 2002.

Vassiliadis, D., R.S. Weigel, A.J. Klimas, S.G. Kanekal, and R.A. Mewaldt, Modes of energy transfer from the solar wind to the inner magnetosphere, *Phys. Plasmas*, **10**, 2, 463–473, 2003.

Vassiliadis, D., S.F. Fung, and A.J. Klimas, Solar, interplanetary, and magnetospheric parameters for the radiation belt energetic electron flux, *J. Geophys. Res.*, **110**, A4, Art. No. A04201, 2005.

Vassiliadis, D., Systems theory for geospace plasma dynamics, *Rev. Geophys.* (in print), 2006.

Wang, Y.M., and N.R. Sheeley, Solar-wind speed and coronal flux-tube expansion, *Ap. J.*, **355**(2), 726, Part 1, 1990.

Weimer, D.R., A flexible, IMF dependent model of high-latitude electric potentials having 'space weather' applications, *Geophys. Res. Lett.*, **23**, 18, 2549–2552, 1996.

Weimer, D.R., An improved model of ionospheric electric potentials including substorm perturbations and application to the Geospace Environment Modeling November 24, 1996, event, *J. Geophys. Res.*, A106, 1, 407–416, 2001.

Weimer, D.R., D.M. Ober, N.C. Maynard, M.R. Collier, D.J. McComas, N.F. Ness, C.W. Smith, and J. Watermann, Predicting interplanetary magnetic field (IMF) propagation delay times using the minimum variance technique, *J. Geophys. Res.*, **108**, A1, Art. no. 1026, 2003.

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

Wu, J.-G., and H. Lundstedt, Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks, *J. Geophys. Res.*, **102**, A7, 14,255–14,268, 1997.

Wu, J.-G., L. Eliasson, H. Lundstedt, A. Hilgers, L. Andersson, O. Norberg, Space environment effects on geostationary spacecraft: analysis and predictions, in*: Space Weather: physics and applications, Adv. Space Res.*, **26**, 1, 31–36, 2000.

Index

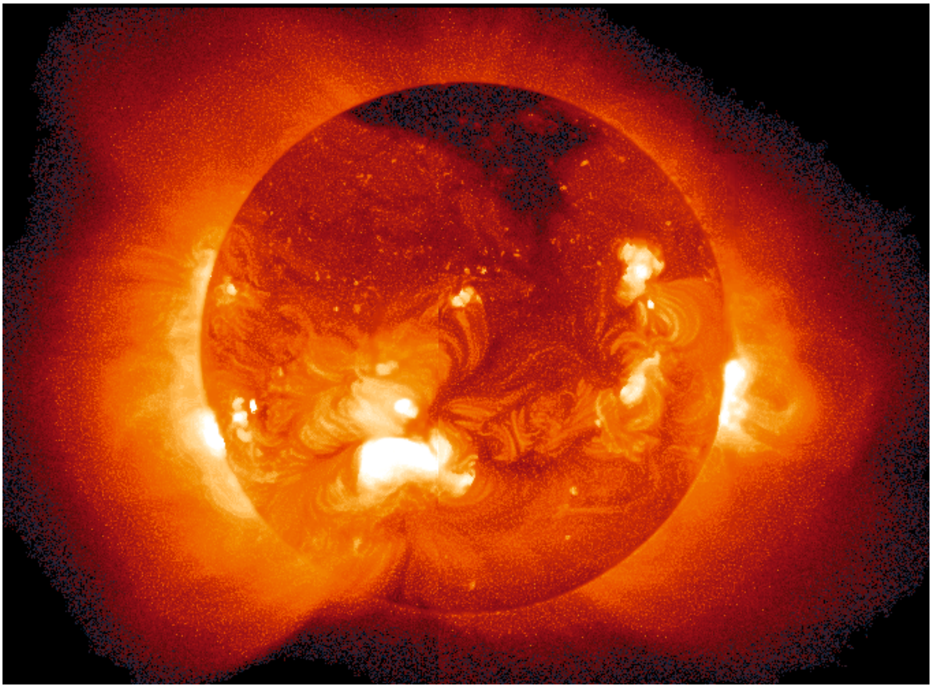


Figure 3.4. Soft X-ray image of the solar corona taken on May 8, 1992 using the soft X-ray telescope (SXT) onboard the Japanese/U.S. satellite Yohkoh. Courtesy: Yohkoh/SXT Consortium.

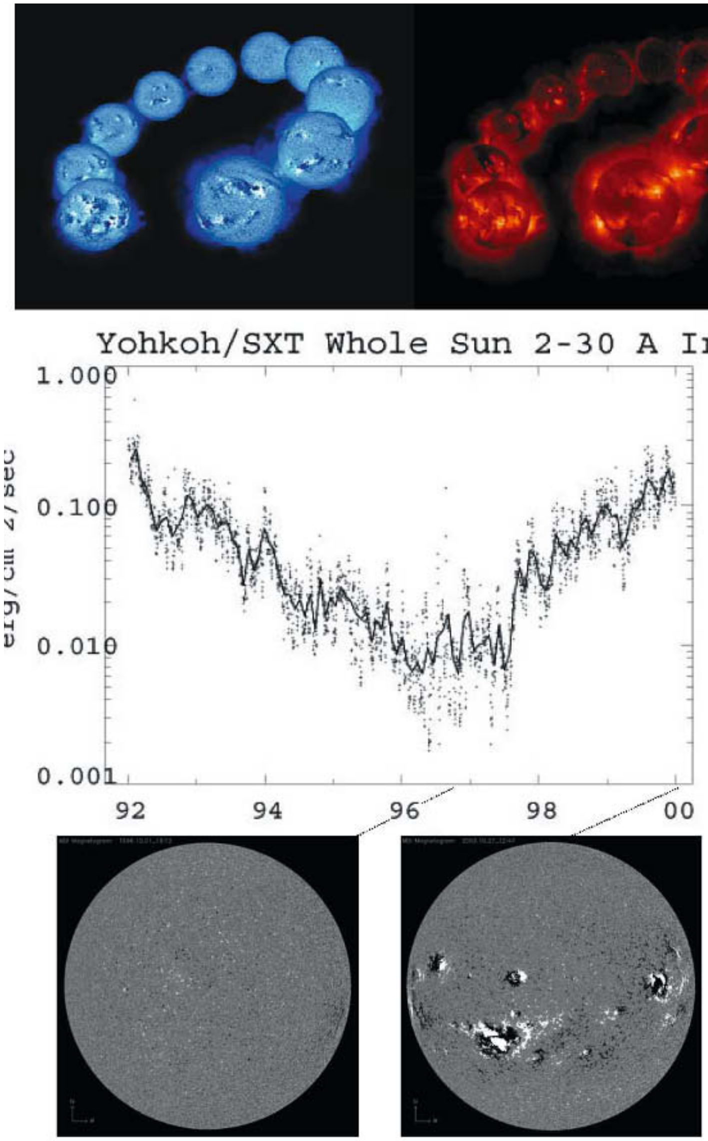


Figure 3.5. Top left: variation of the longitudinal component of the photospheric magnetic field from the time after solar maximum in cycle 22, near 1992, until the next one in cycle 23, around 2000, as observed by the Kitt Peak National Observatory (KPNSO). Top right: Variability of the Sun's X-ray corona for the same time as imaged by Yohkoh/SXT. Center plot: X-ray intensity–time profile measured by Yohkoh/SXT at 2–30 Å; the points are daily average irradiance values and the solid curve is a plot for average values over time periods of a solar rotation. Courtesy: Loren Acton, Montana State University. Bottom: SoHO/MDI magnetograms taken in October 1996 (left) and in October 2000 (right). Dark photospheric regions correspond to negative magnetic polarities, white areas to positive ones. Courtesy: KPNO, Yohkoh and SoHO consortia.

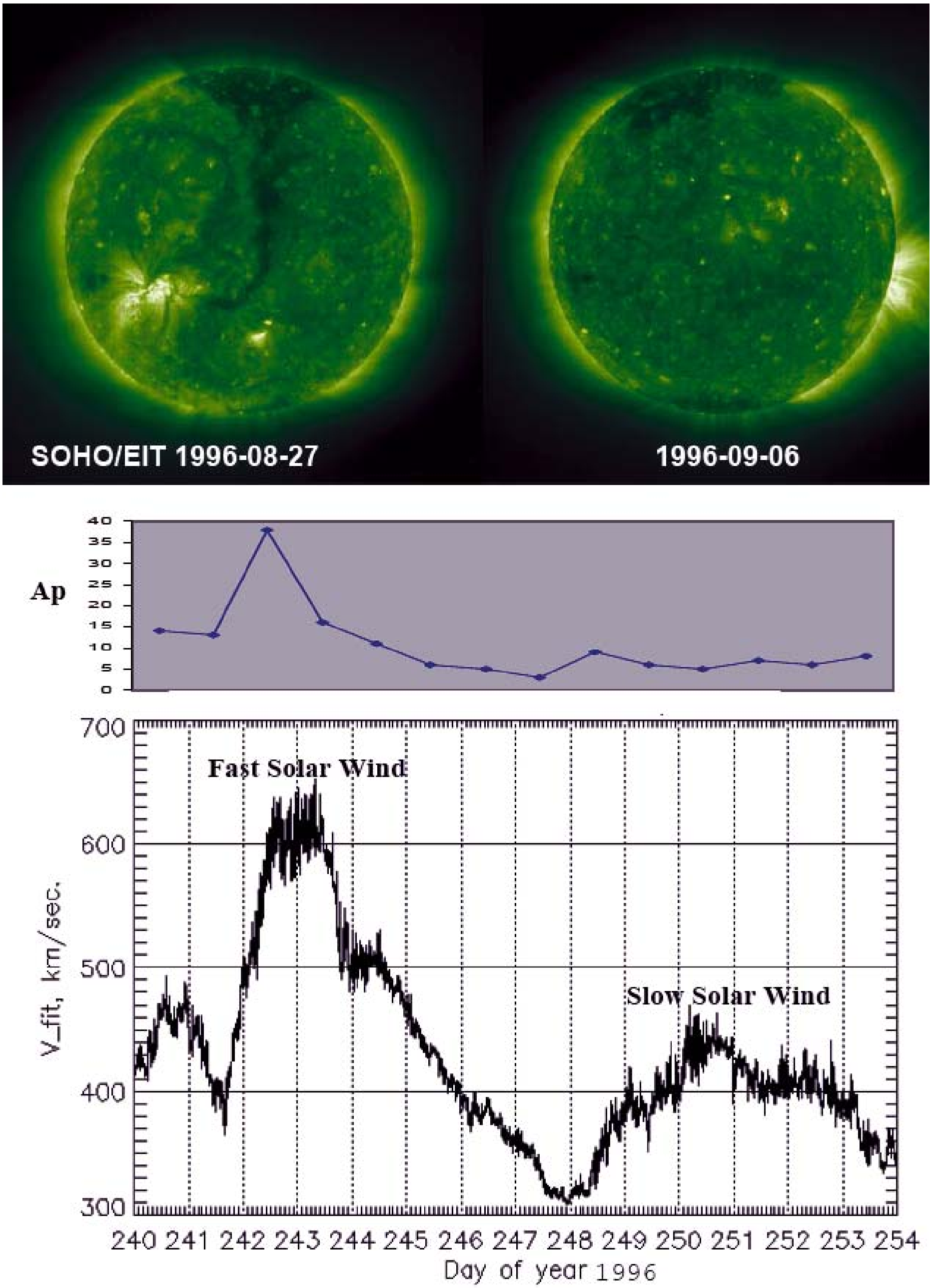


Figure 3.8. Top: structure of the solar corona imaged by SoHO/EIT at 195 Å on August 27 and September 6, 1996. Middle: intensity of the geomagnetic Ap index during the time period of the solar wind measurements. Bottom: solar wind speed measured by the WIND satellite from August 27 until September 10, 1996. The time periods of passages of fast and slow solar wind streams are marked.

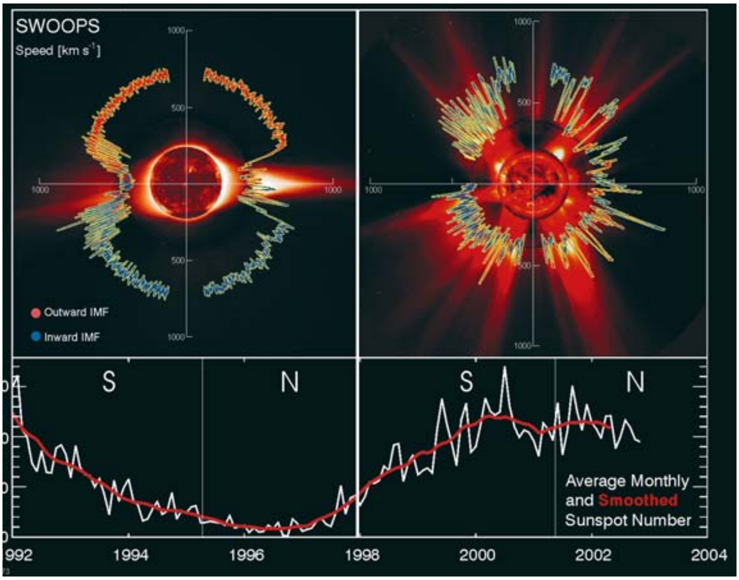


Figure 3.10. Ulysses measurements of the solar wind speed and IMF polarity in the 3-D heliosphere at times near solar minimum (top left) and maximum (top right) – that is, during its first and second orbit – plotted on top of the sunspot number diagram from 1992 until 2004. In the polar diagrams, the solar wind speed increases with distance from the coordinate center. The blue and red colors denote negative (inward-directed) and positive (outward-directed) polarities of the IMF as measured by Ulysses. The insetted SoHO/EIT/LASCO and Mauna Loa K-coronameter images show the different structures of the corona during the two orbits. Solar wind data are from the Ulysses SWOOPS experiment, magnetic field data from the Ulysses VHM/FGM experiment. From McComas *et al.* (2003).

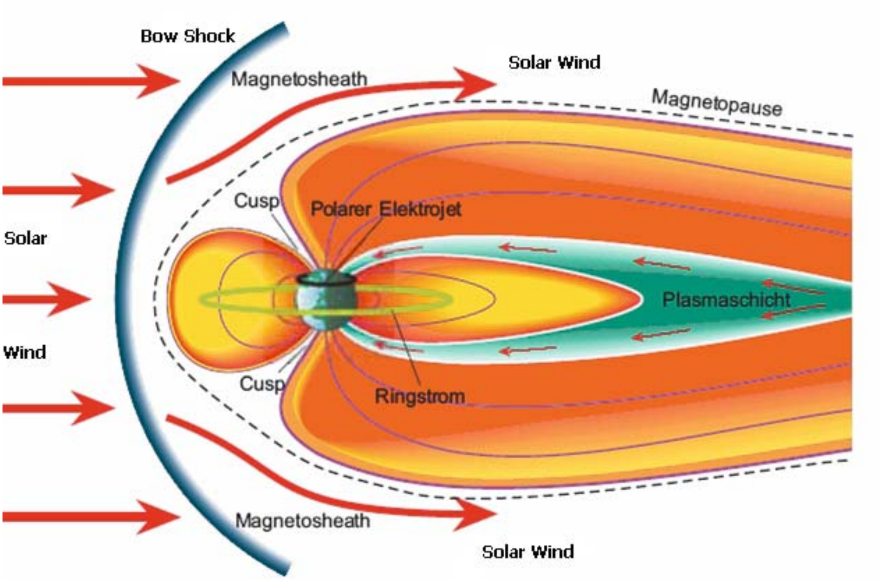


Figure 3.11. The solar wind flow around the Earth's magnetosphere, its structure due to the solar impact and major plasma regimes and current systems. Courtesy: Schlegel (2001).

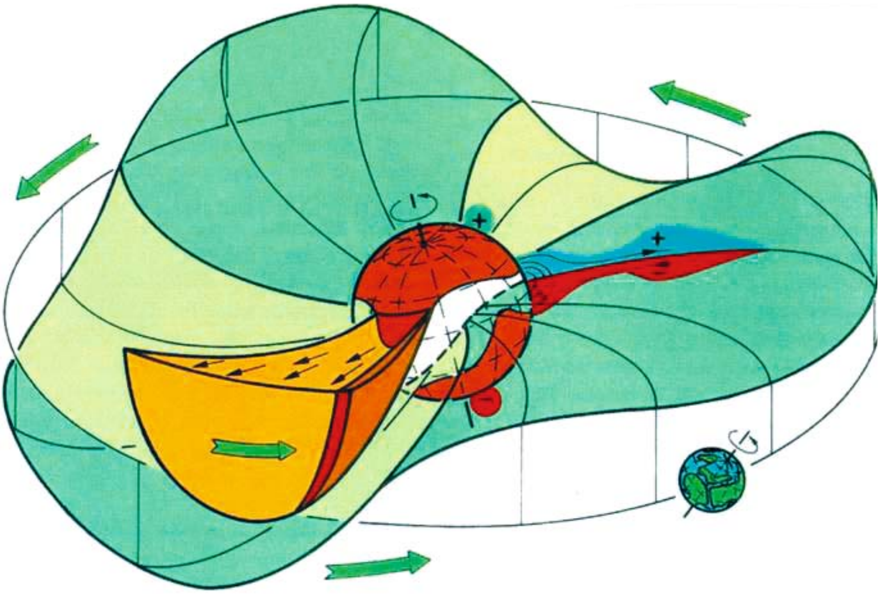


Figure 3.14. Sketch of the inner heliosphere – the solar ballerina model – as proposed by Alfvén (1977). The Sun's poles are occupied by large coronal holes of opposite magnetic field polarity, typical for odd cycles. On both sides of the heliospheric current sheet (HCS), solar wind with opposite magnetic field polarity will be observed. A co-rotating interaction region where a fast solar wind stream interacts with a low-speed one is cross-hatched. Note how the rotating Sun with its warped HCS dances across the Earth.

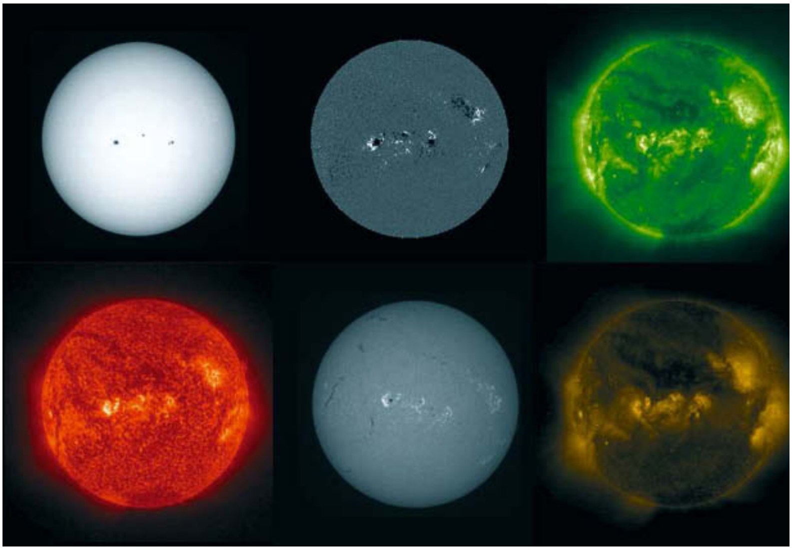


Figure 3.19. Top left to bottom right: white-light solar image taken on November 9, 2005 by the GONG project, SoHO/MDI magnetogram, SoHO/EIT 195 Å image, EIT 304 Å image, Catania H_α image and SoHO/EIT 284 Å image. Courtesy: SoHO/EIT/MDI Consortium, GONG Project, Catania Astrophysical Observatory. The images were retrieved using the Solar Weather Browser (Copyright © 2004, 2005 The Royal Observatory of Belgium) at *http://sidc.be/SWB*

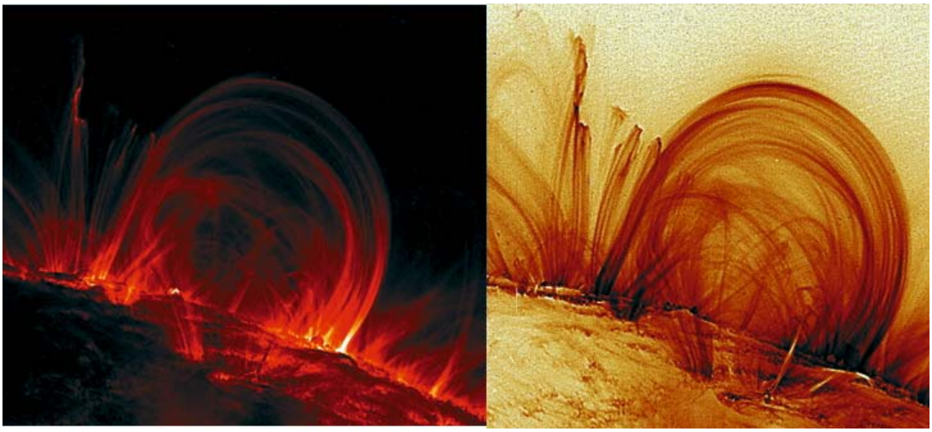


Figure 3.21. Fine structure of the solar corona as observed by TRACE. Courtesy: TRACE Consortium.

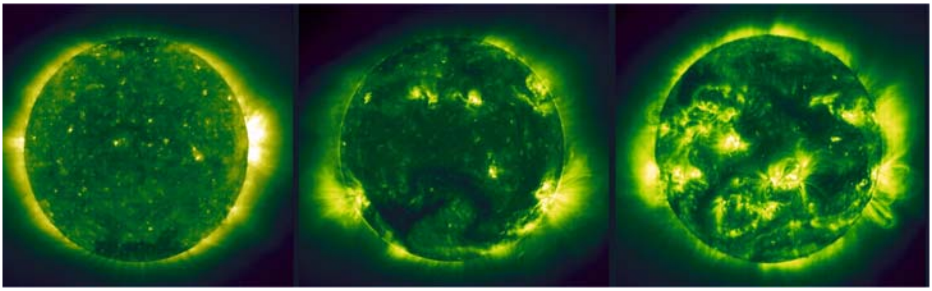


Figure 3.23. Left to right: changing structure of the solar corona from solar activity minimum to maximum as seen in 195 Å images taken by SoHO/EIT in 1996, 1998 and 1999. Courtesy: SoHO/EIT Consortium.

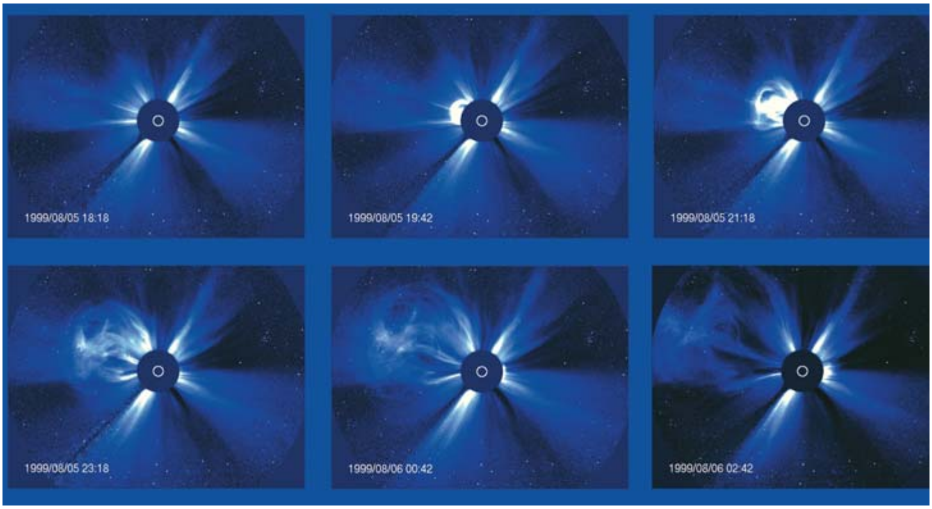


Figure 3.24. A fast solar coronal mass ejection (CME) observed at the east limb on August 5, 1999 by SoHO/LASCO in the field of view from 6–30 R_S. The CME reached a speed of about 1000 km/s. Courtesy: SoHO/LASCO Consortium.

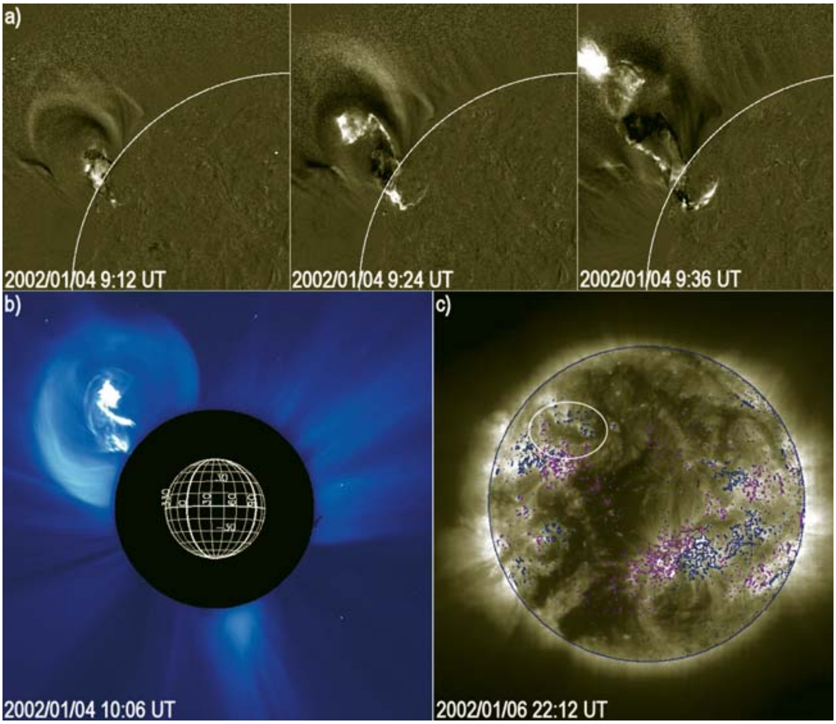


Figure 3.26. Top row: running difference images taken by SoHO/EIT at 195 Å on January 4, 2002 showing the eruptive rise of a prominence at the Sun's east limb. The images are differential in intensity with respect to the previous one. Bottom left: structure of the prominence-associated CME seen by SoHO/LASCO about 50 minutes later. Bottom right: contours of a SoHO/MDI magnetogram superposed on an EIT 195 Å image showing the identified source region of the CME in the low corona and photosphere as marked by the ellipse. From Cremades and Bothmer (2004).

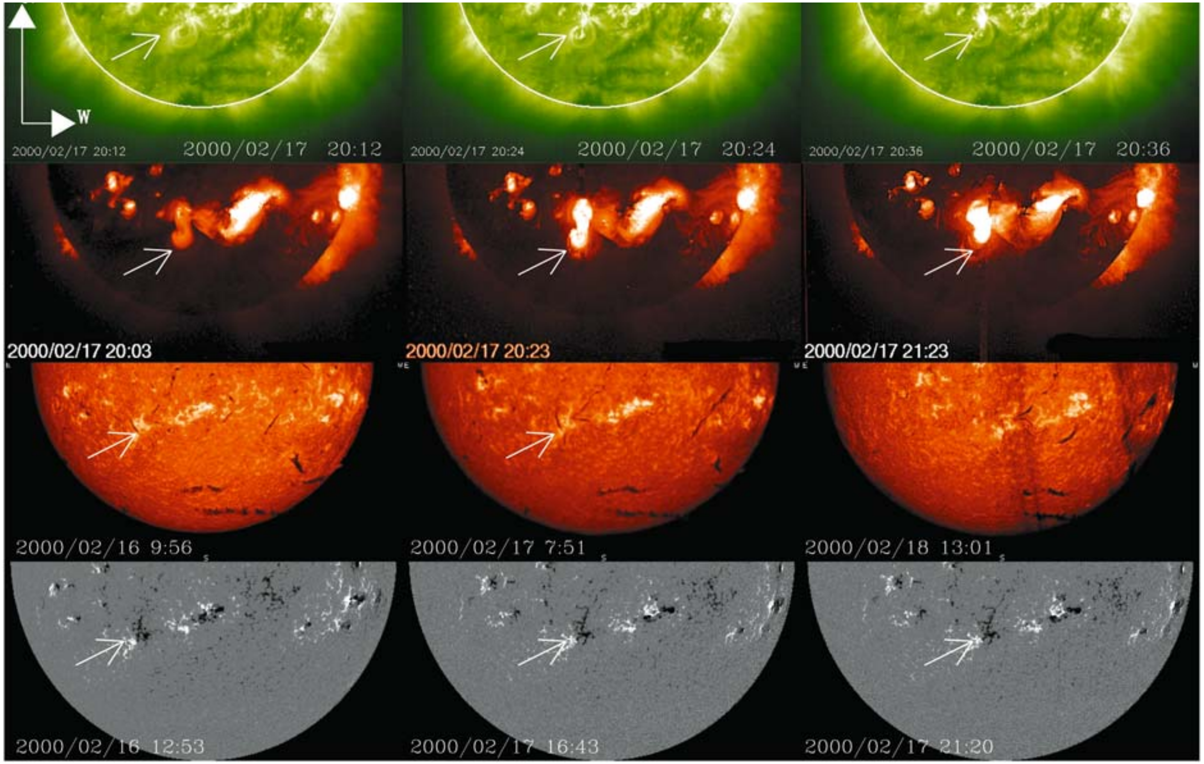


Figure 3.29. Multi-wavelength observations showing the source region of the halo CME observed by SoHO/LASCO on February 17, 2000. Top images: SoHO/EIT 195 Å images showing the development of a post-eruptive arcade. Second panel: Yohkoh soft X-ray images showing a bright sigmoidal structure in the CME's source region. Third panel: H_α images showing the disappearance of the associated filament. Bottom panel: SoHO/MDI observations of the magnetic field structure in the underlying photospheric source region of the CME. From Tripathi, Bothmer and Cremades (2004).

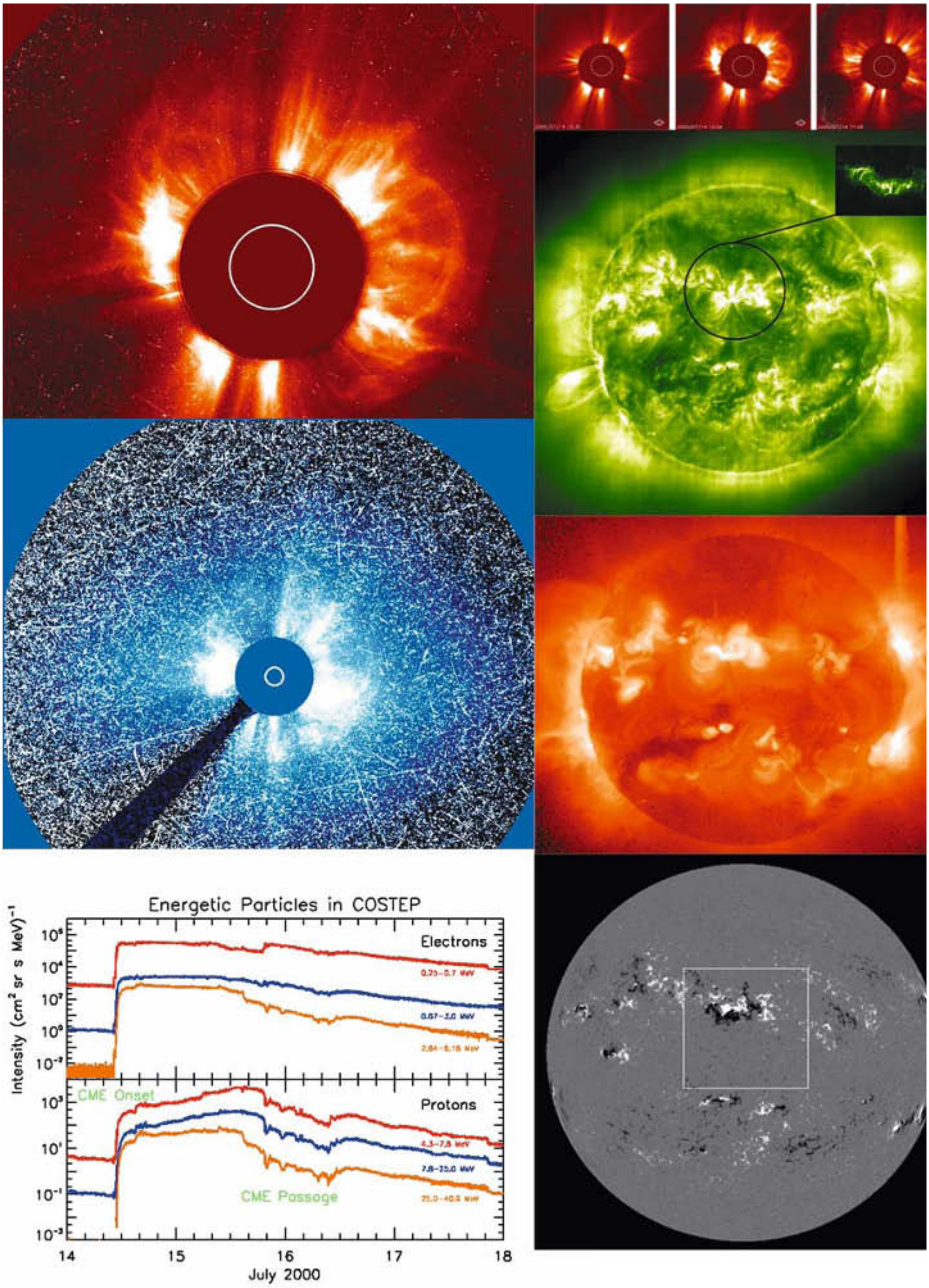


Figure 3.37. Mosaic of remote-sensing and *in situ* observations from SoHO for the CME on July 14, 2000. Left, top to bottom: SOHO/LASCO C2 image, SoHO/LASCO C3 image, MeV electron and proton fluxes measured by SoHO/COSTEP. Right, top to bottom: Sequence of C2 images, EIT 195 Å image and TRACE inset showing the EUV post-eruptive arcade, Yohkoh/SXT image, SoHO/MDI image. Courtesy: SoHO and Yohkoh Consortia.

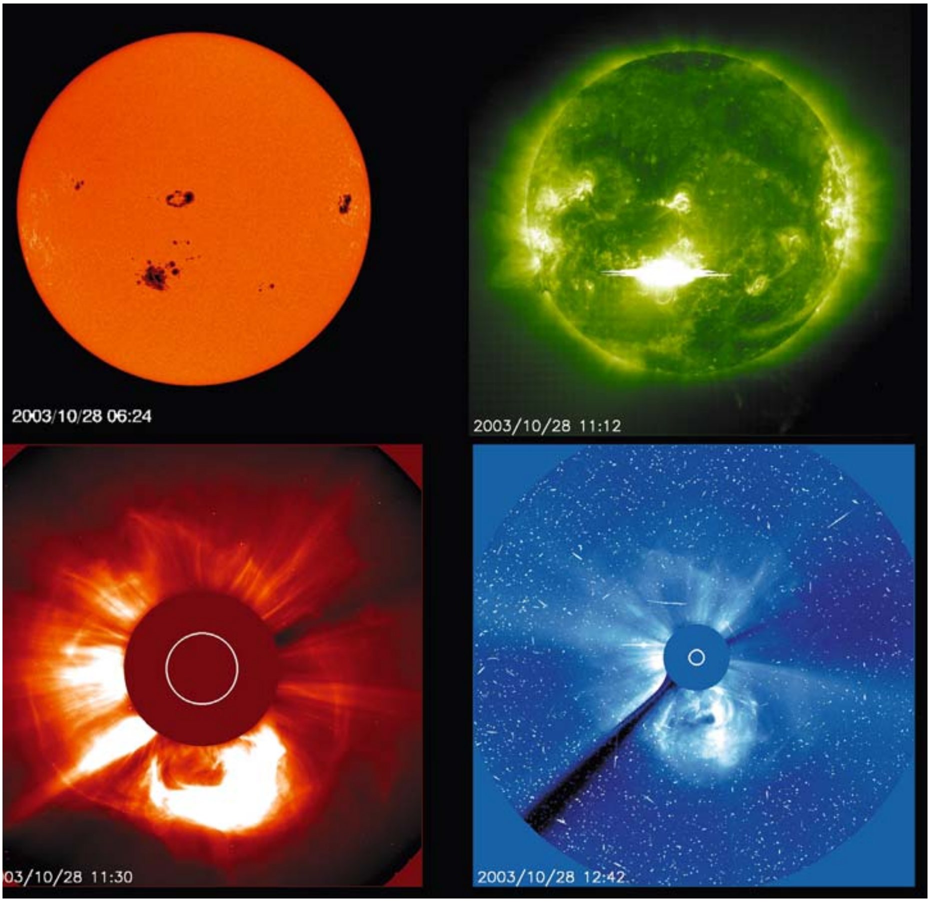


Figure 3.40. SoHO multi-wavelength observations of the superfast (>2000 km/s) CME on October 28, 2003. Top images: (left) MDI continuum image; (right) low corona and flare imaged by EIT at 195 Å. Bottom images: (left) LASCO/C2 image of the CME; (right) LASCO/C3 image of the CME. Courtesy: SoHO MDI/EIT/LASCO Consortia.

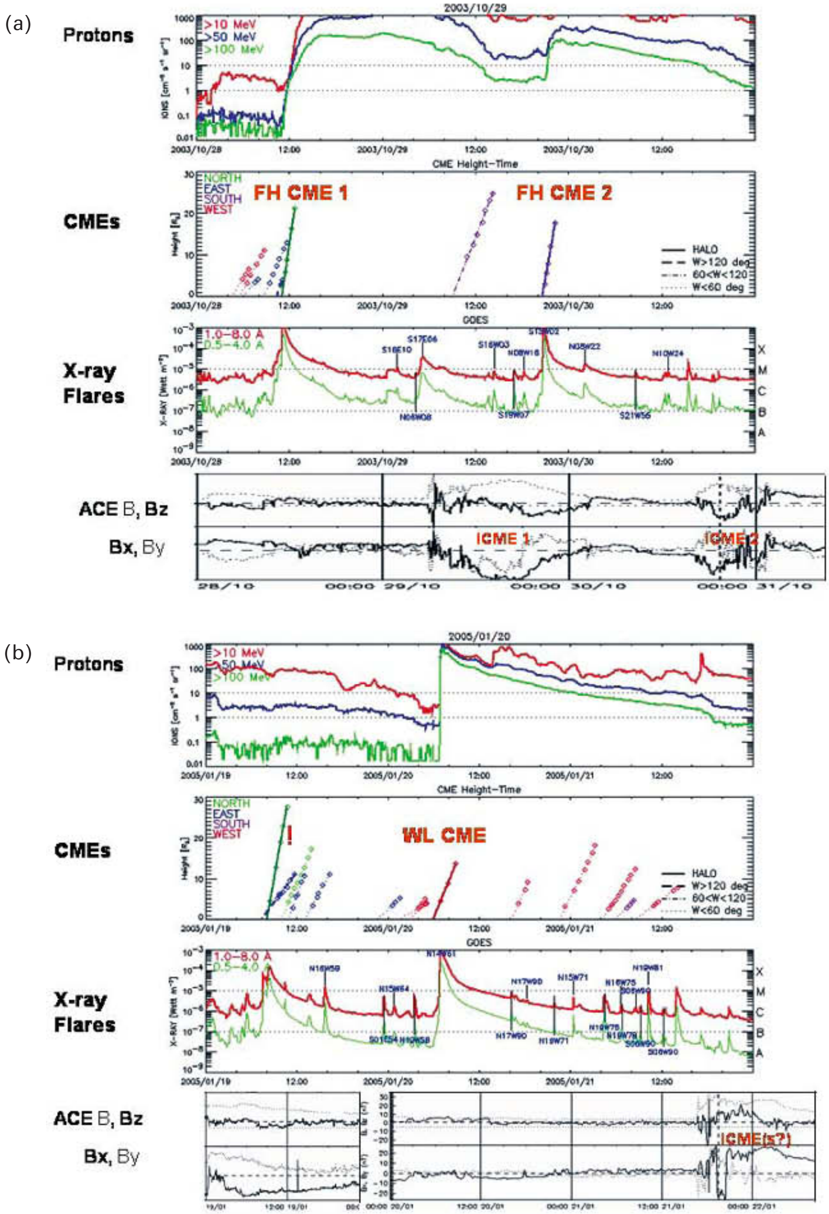


Figure 3.41. (a) From top to bottom: proton flux measured in different energy channels by the GOES satellite on October 28–30, 2003. Height–time diagram labeling the two superfast front-side halo (FH) CMEs observed by SoHO/LASCO. Intensity–time profile of the soft X-ray flux measured by GOES. IMF parameters measured by ACE. The two ICMEs at 1 AU are labeled. (b) Top to bottom: same parameters as in (a) plotted for the SEP event on January 20, 2005. In this case the CME relationship could not be uniquely determined. Plots adapted from the *SoHO LASCO CME Catalog* at *http://cdaw.gsfc.nasa.gov/CME_list/*

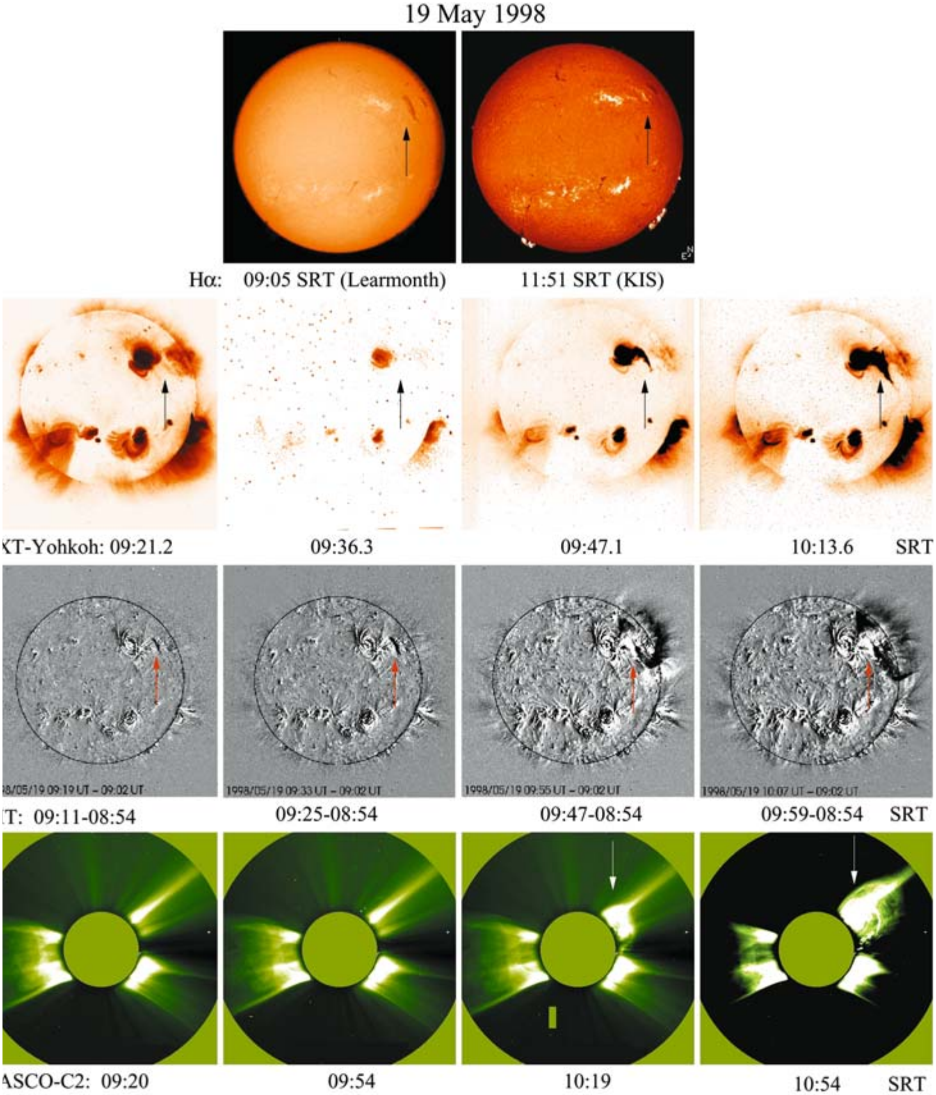


Figure 3.45. Observations of the filament eruption and CME on May 19, 1998. Arrows point to the filament position (panels 1–3). The time is given in Solar Release Time (SRT). Top panel: H_α images (courtesy: Learmonth Solar Observatory and Kiepenheuer-Institut für Sonnenphysik) shortly before and after the eruption. Second panel: Yohkoh/SXT images. Darkness corresponds to enhanced emission. Third panel: SoHO/EIT 195 Å running difference images showing the filament and the manifestation of an EIT wave after its eruption. Brightness corresponds to enhanced emission. Bottom panel: CME observed by LASCO/C2. Note that the erupted filament occurs inside the CME as a long narrow structure. North is to the top and east to the left. From Klassen *et al.* (2002).

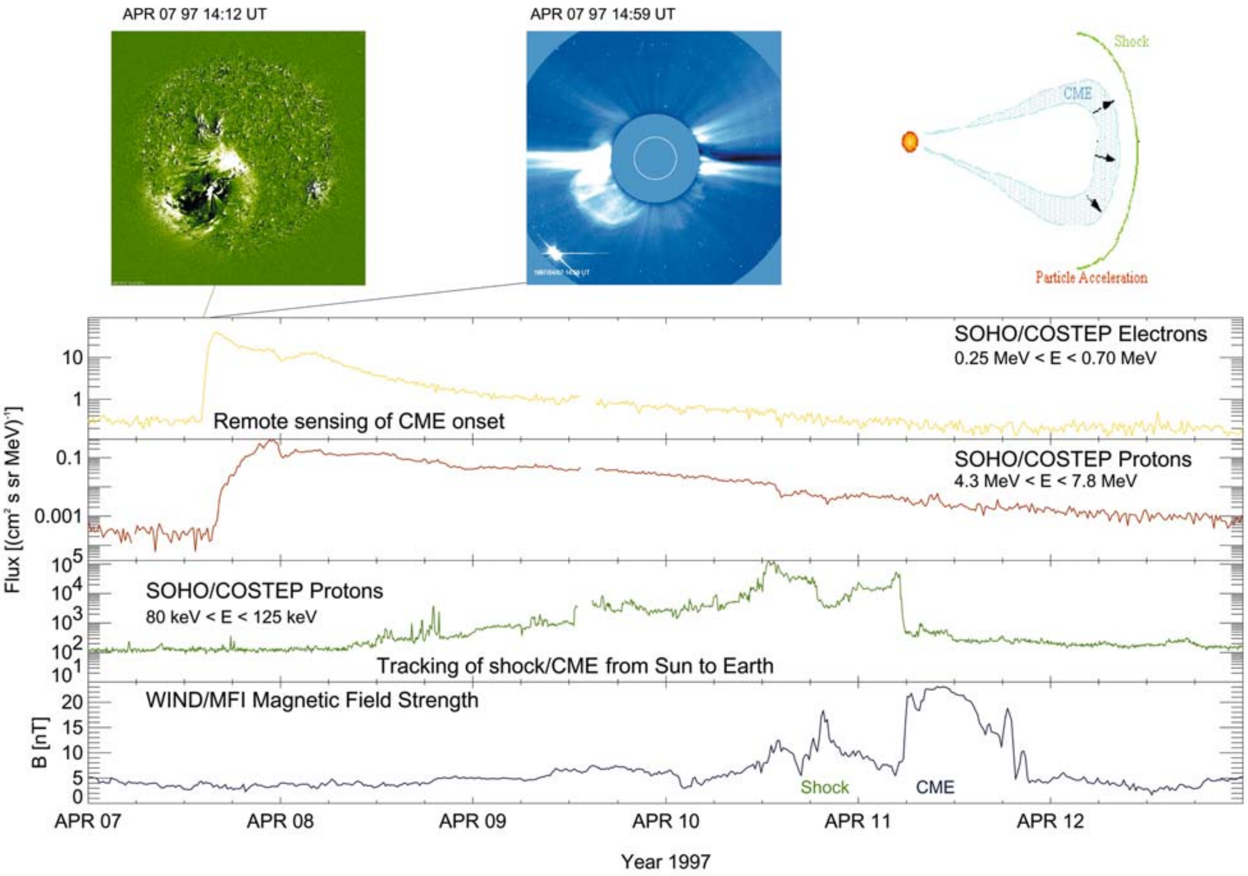
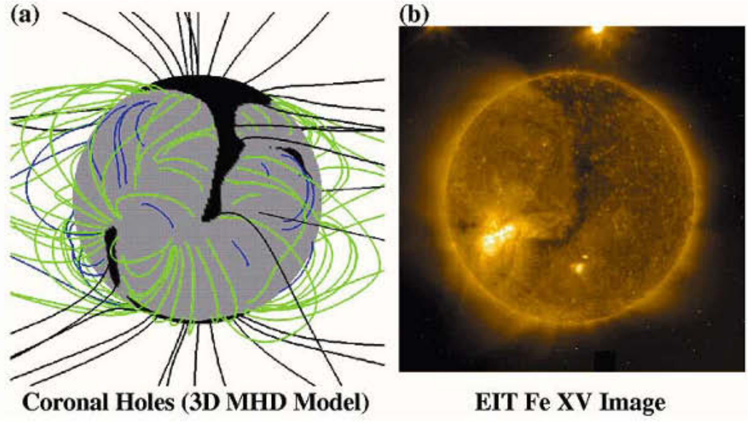


Figure 3.47. SoHO/EIT/LASCO/COSTEP EUV, white-light, electron and proton observations for the CME event on April 7, 1997. The shock-associated CME (more precisely the ICME) passed Earth's orbit on April 10/11, as shown by the WIND magnetic field and SoHO/COSTEP 80–125-keV proton measurements. From Bothmer *et al*. (1997b).



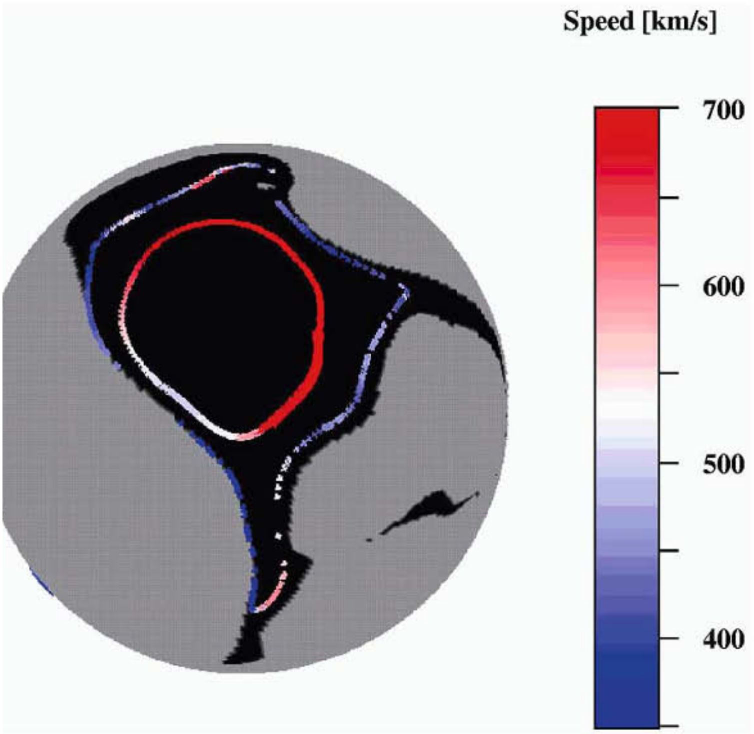


Figure 3.56. Top: (a) coronal holes (open field regions, colored black) from an MHD model of the solar corona for August 27, 1996. Closed field regions are colored gray. (b) EIT Fe XV image for the same day. Bottom: mapping of solar wind velocity measurements back to the Sun. Red marks highest solar wind speeds. Coronal hole boundaries are modeled as for the top left image but shown as viewed from above the north pole of the Sun. From Balogh and Bothmer *et al*. (1999). Modeling by Linker *et al*. (1999).

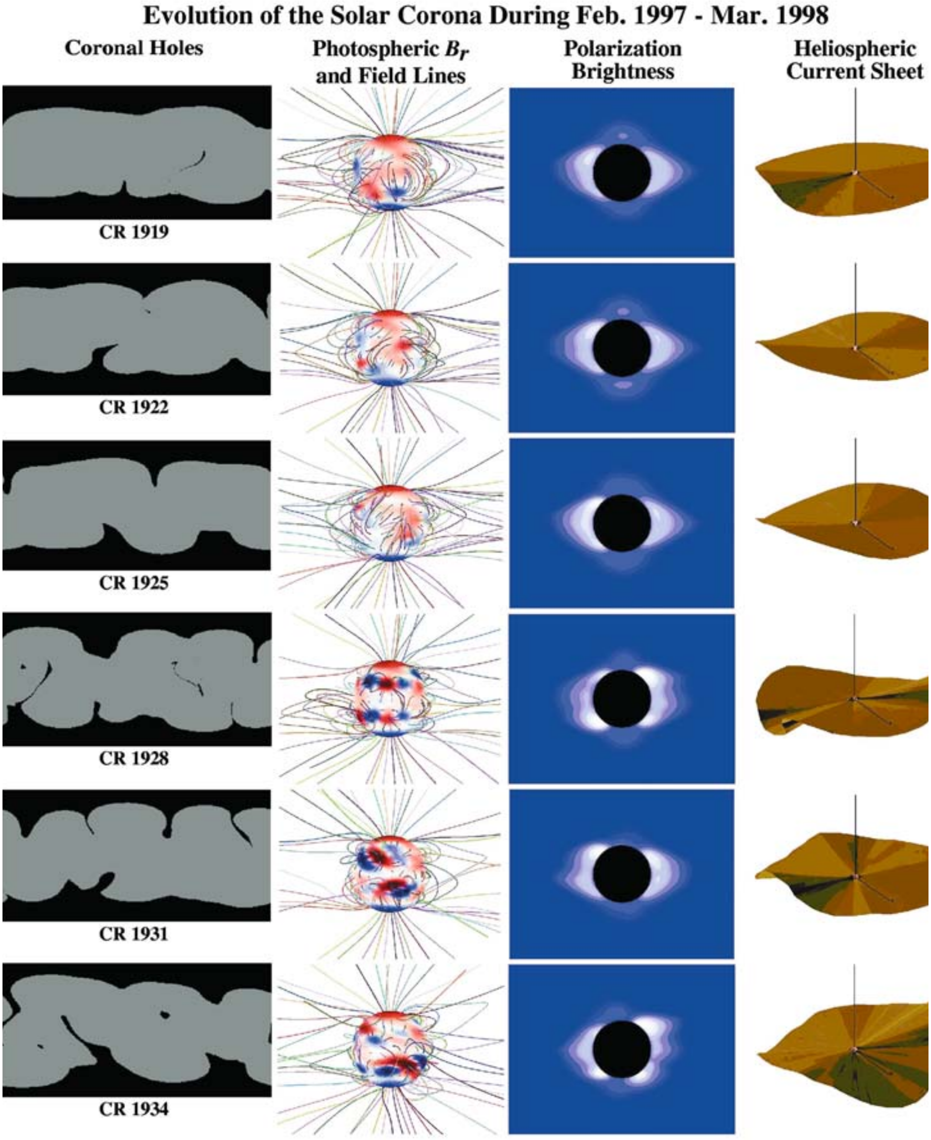


Figure 3.57. The changing structure of the solar corona during the period February 1997–March 1998 (Carrington rotations 1919–1934), as illustrated by coronal hole maps (longitude vs. latitude), with gray/black indicating closed/open field regions, field line traces with the radial magnetic field shown on the surface of the Sun (blue areas denote field lines directed into, red areas field lines out of the Sun, respectively), polarization brightness, and the shape of the heliospheric current sheet. This time period represents the rising phase of solar cycle 23. The photospheric magnetic field was set as a time-dependent boundary condition on the 3-D MHD simulation using Kitt Peak synoptic maps (Mikic *et al.*, 1999). From Balogh and Bothmer *et al.* (1999).

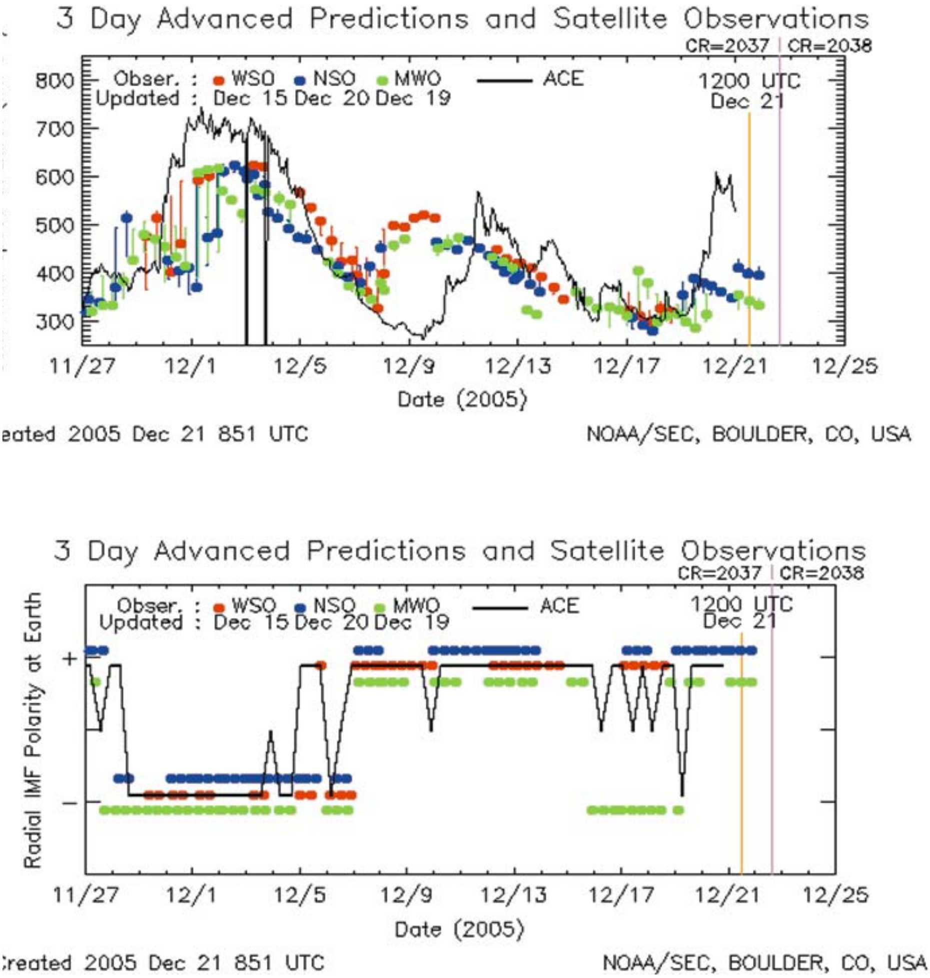


Figure 3.58. Comparison of predicted and measured solar wind velocities (top panel) and IMF polarities (bottom panel) at 1 AU in December 2005. *In situ* data are provided by the ACE satellite. WSO (Wilcox Solar Observatory), NSO (National Solar Observatory), MWO (Mount Wilson Observatory) denote the different magnetic field data used as input for the model calculations (Wang, Sheeley, Arge). Courtesy: NOAA/SEC, U.S.A.

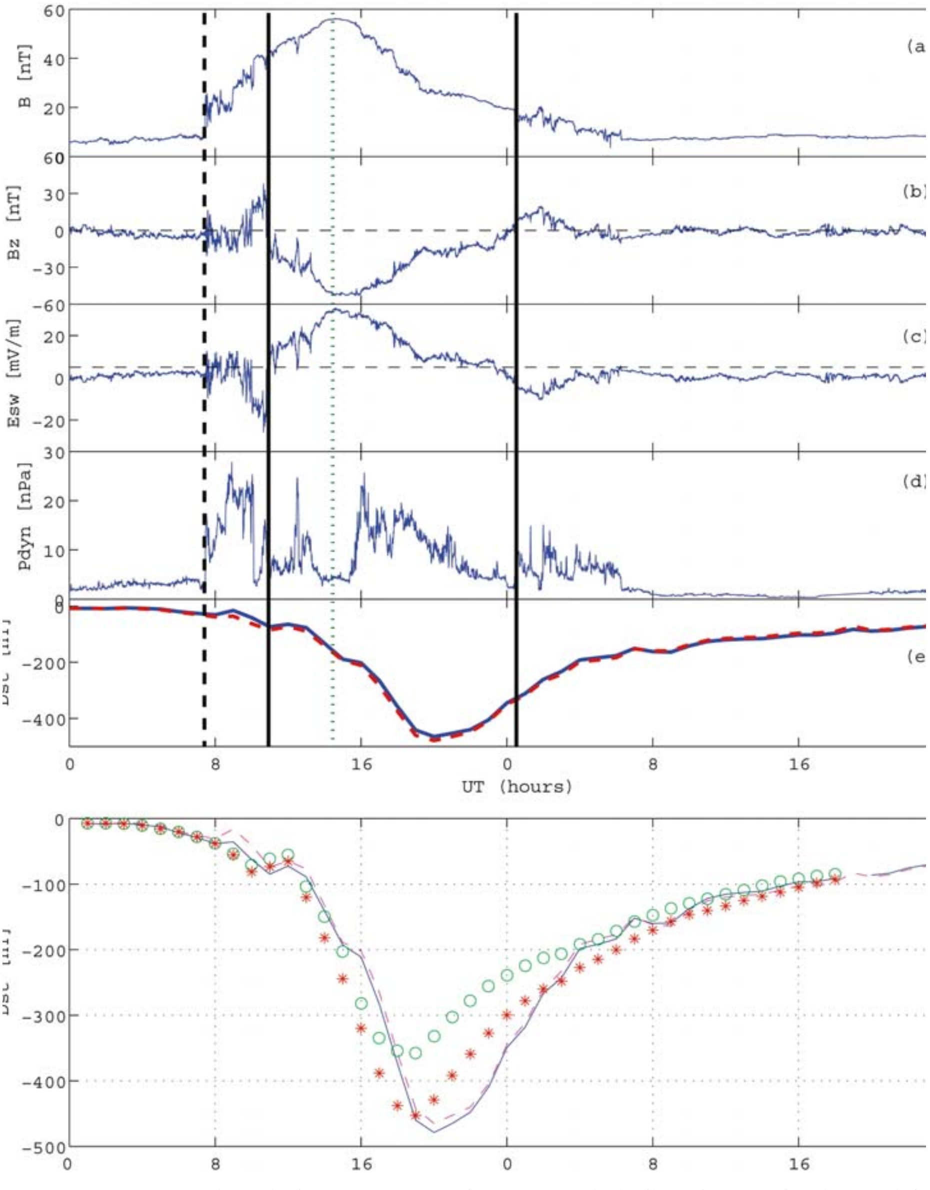


Figure 3.59. Top: solar wind parameters and geomagnetic indices for a 2-day interval from November 20 to 21, 2003 measured by ACE. The dashed lines indicate the shock, the two solid lines label the ICME (magnetic cloud). Bottom: measured and predicted Dst development for November 20–21, 2003. The blue solid line is the 1-hour Dst index and the purple dashed line is Dst*. The green open circles show the predicted Dst* using the O'Brien and McPherron (2000) model and the red stars the predicted Dst* using the Wang *et al*. (2003) model. From Huttunen *et al*. (2005).

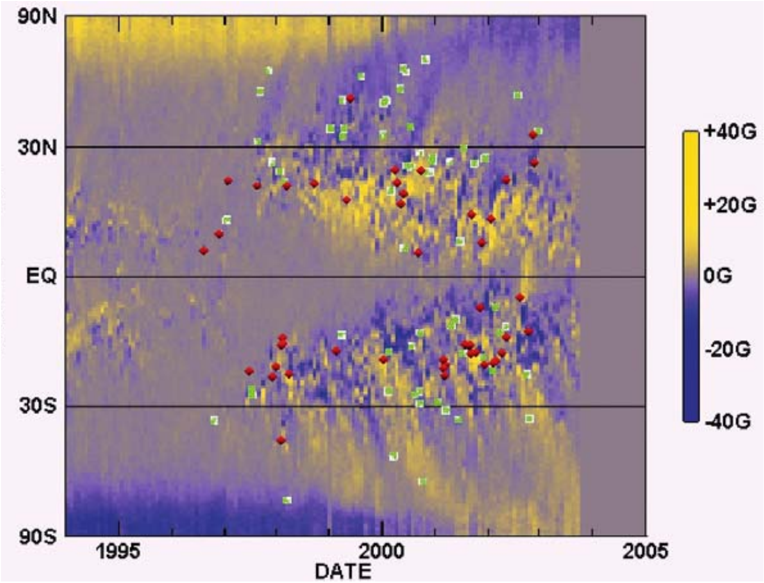


Figure 3.60. Source regions of structured CMEs in 1996–2002 displayed together with the evolution of the longitudinal component of the photospheric magnetic field. Active regions are marked dark blue, decaying ones are labeled bright blue. The blue color of the photospheric field corresponds to negative magnetic polarity, the yellow to positive polarity. From Cremades, Bothmer and Tripathi (2006).

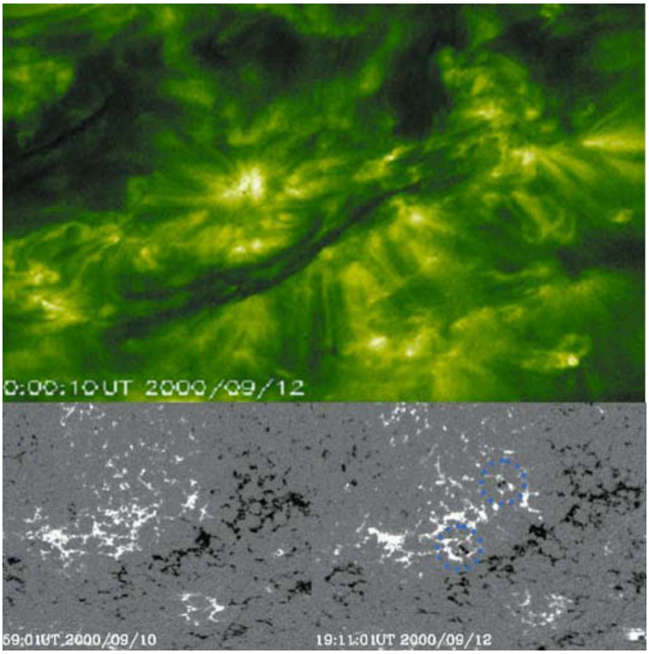


Figure 3.62. A large solar filament on September 12, 2000 at a time shortly before its eruption imaged by SoHO/EIT at 195 Å and the evolution of the photospheric magnetic field in its source region as observed by SoHO/MDI. The erupting filament was associated with a CME with a speed of ~1600 km/s as observed by SoHO/LASCO. Blue circles mark the locations where new magnetic flux emerged close to the filament channel. Note that the bottom panel shows two MDI images at different times. From Bothmer (2006), after Bothmer and Tripathi (2006).

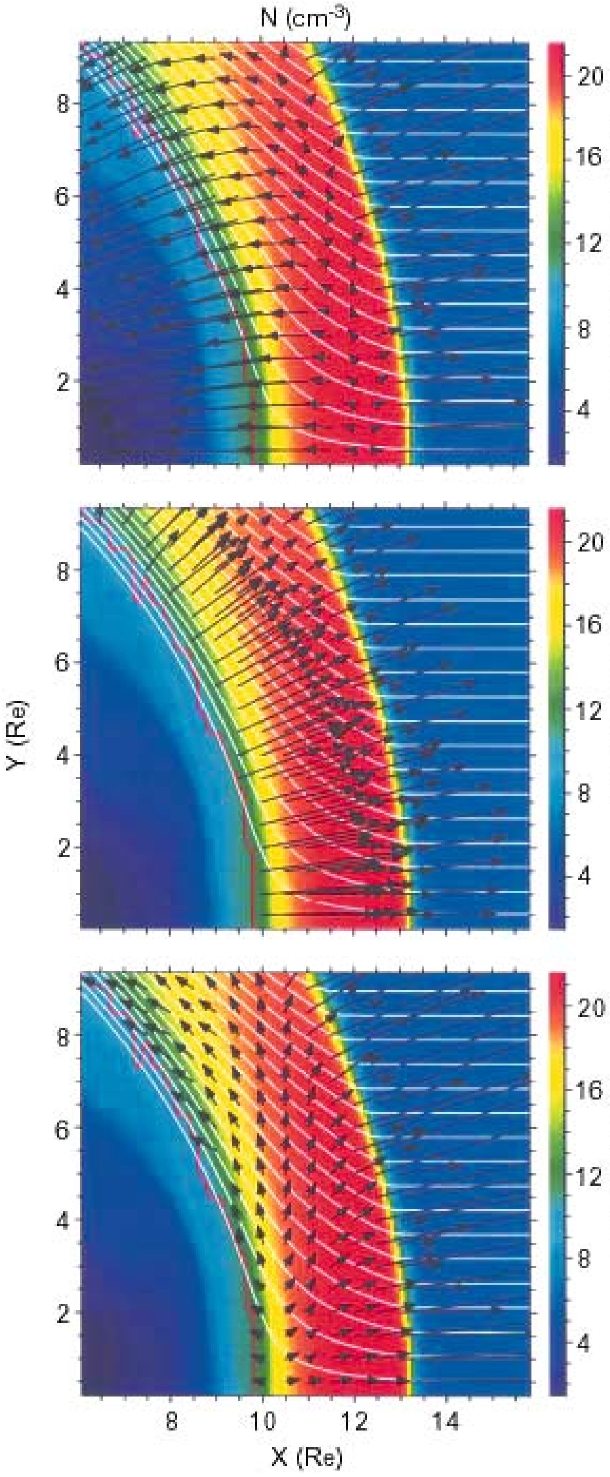


Figure 4.4. The pressure gradient force (top), the magnetic force (middle), and their sum (bottom) for an MHD simulation of the solar wind interaction with a compressible magnetosphere. The MHD solution produces a magnetosheath density low near the magnetopause, in contrast to the gas dynamic solution.

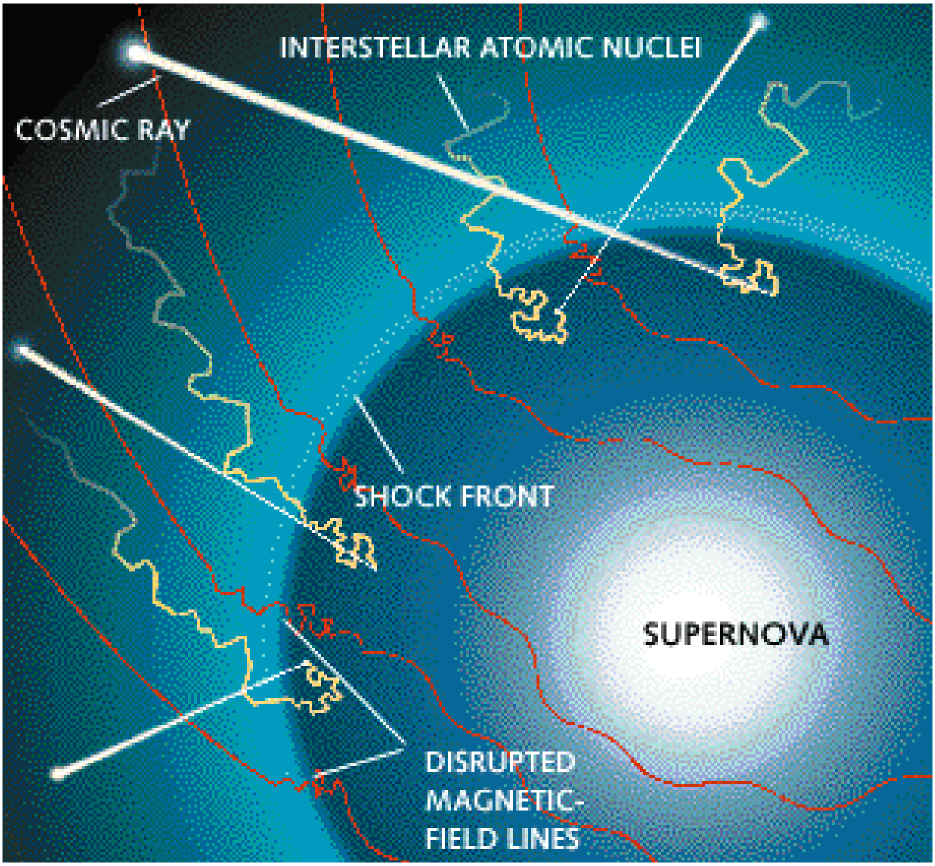


Figure 5.5. The supernova shock front acts as an accelerator for cosmic ray particles. (From Cronin, Gaisser and Swordy, 1997.)

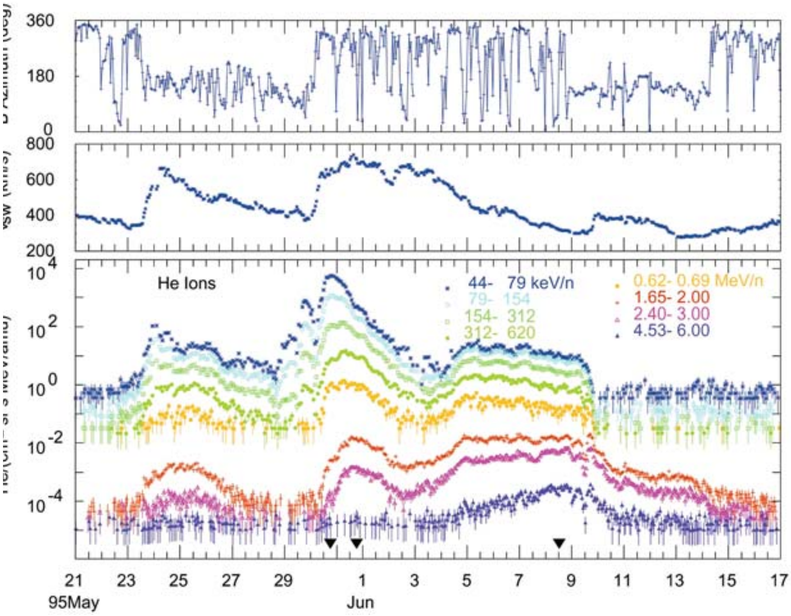


Figure 5.9. Intensity-time profiles for He ions (lower panel), magnetic azimuth (middle panel) and solar wind speed (upper panel). A particle event is associated with a CIR passage on 24 May, and another event occurs in association with a CIR on 30 May. (From Reames, 1999.)

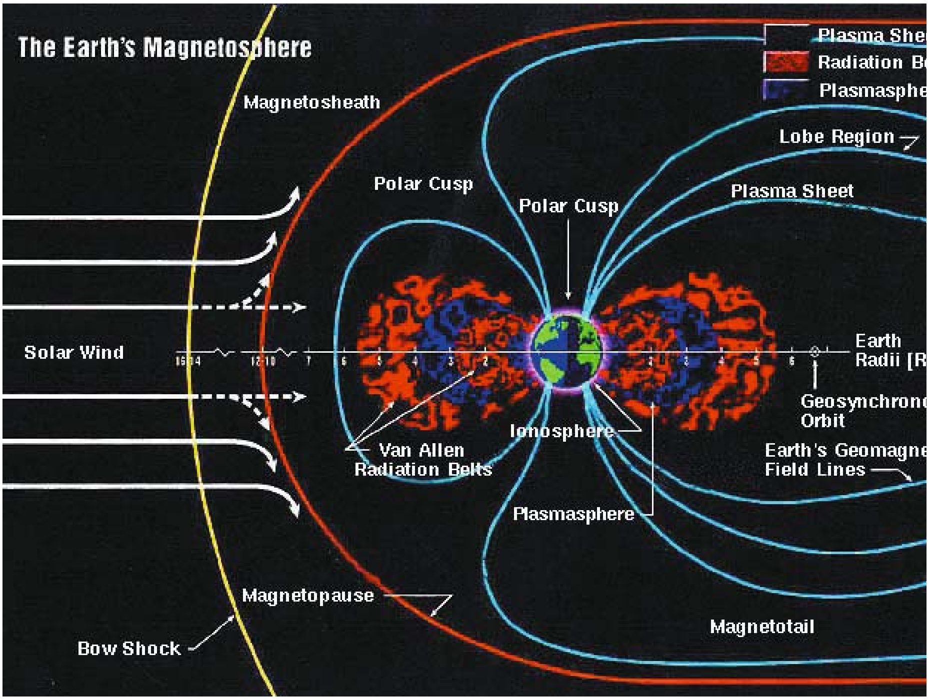


Figure 5.12. Schematic illustration of Earth's magnetosphere, emphasizing the different plasma regions. (Space Environments & Effects Program, NASA Marshall Space Flight Center; see *http://see.msfc.nasa.gov/*.)

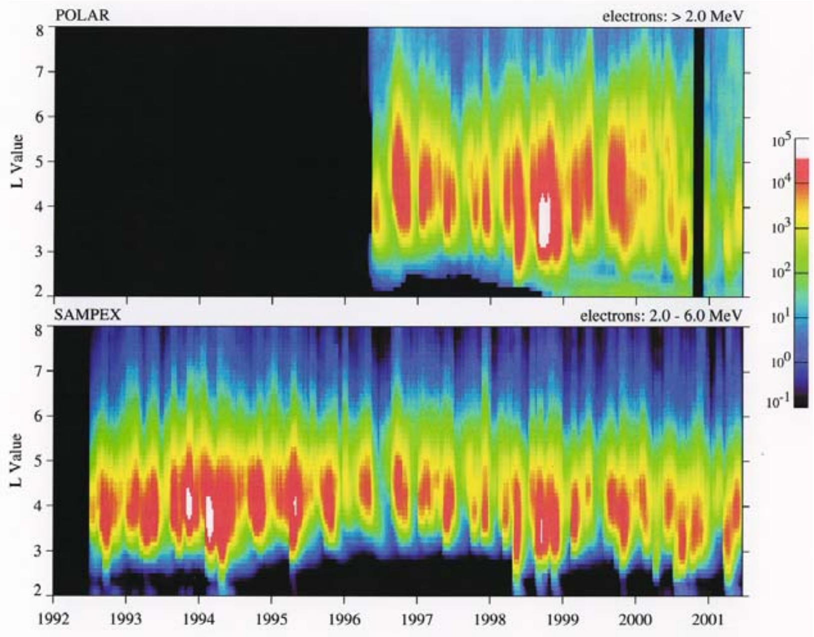


Figure 6.7. Colour-coded intensities of electrons with $E > 2\,\text{MeV}$ (colour bar to right of figure). The McIlwain L-parameter is the vertical scale and time (in years) is the horizontal axis. The lower panel shows available POLAR data and the upper panel shows SAMPEX data. A 27-day smoothing filter has been run over all data. (Adapted from Baker and Li, 2003.)

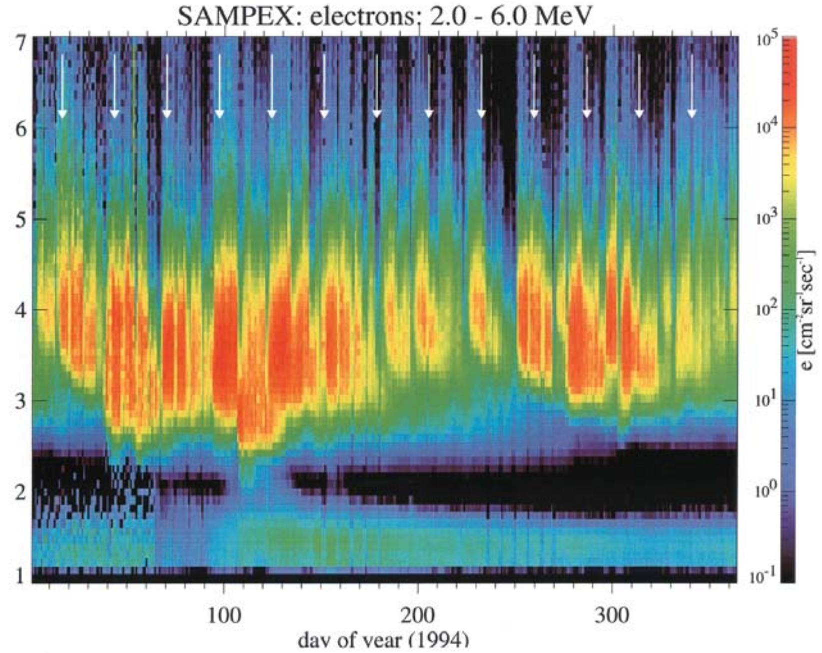


Figure 6.9. Colour-coded intensities of electrons ($E = 2$–$6\,\text{MeV}$) measured by SAMPEX during 1994. Fluxes are indicated by colour bar. Data are plotted as L-value (vertical axis) versus time. White arrows show 27-day recurrence times. (From Baker and Li, 2003.)

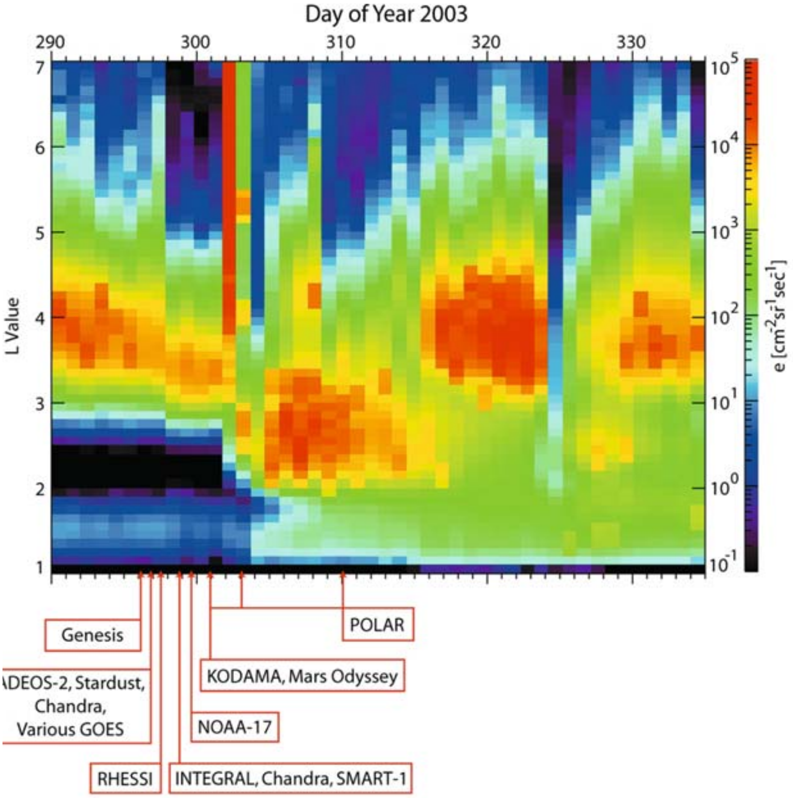


Figure 6.10. A detailed plot of electron flux in the colour-coded format of L-value versus time (see Figure 6.7, previous page). The period covered is from Day 290 to Day 335 (17 October–15 December) 2003. The times of several spacecraft operational anomalies are shown by the red arrows and boxes at the bottom of the figure.

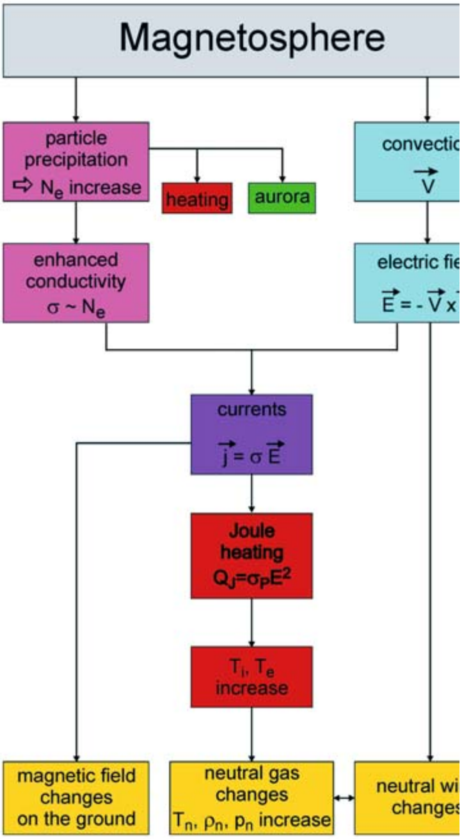


Figure 7.2. Flow chart of space weather effects in the ionosphere and atmosphere.

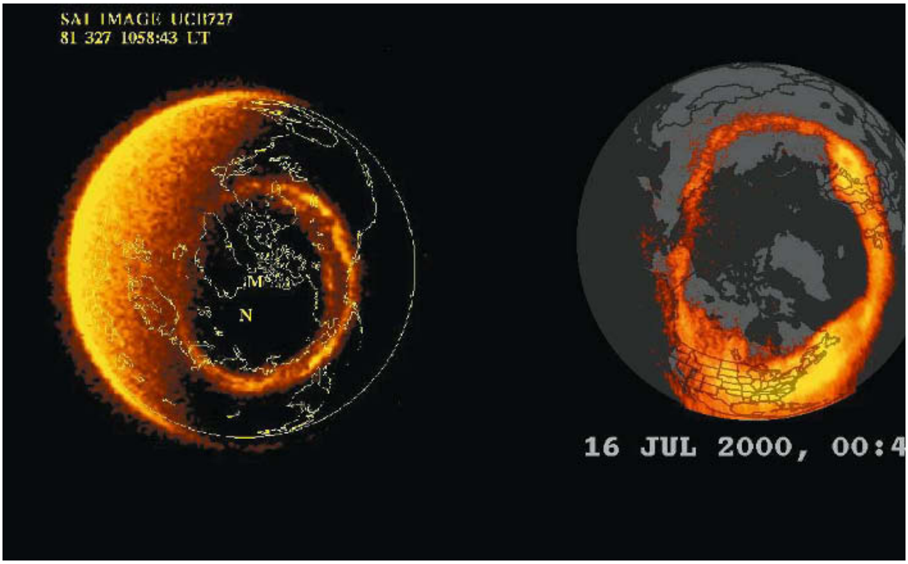


Figure 7.15. Auroral oval during quiet (left) and disturbed (right) conditions.

Figure 7.17. Four auroral displays with different colours and forms. (Upper left) over the EISCAT UHF antenna in Sodankyl, Finland; (upper right and lower left) over Alaska; (lower right) near Cologne, Germany (mid-latitude).

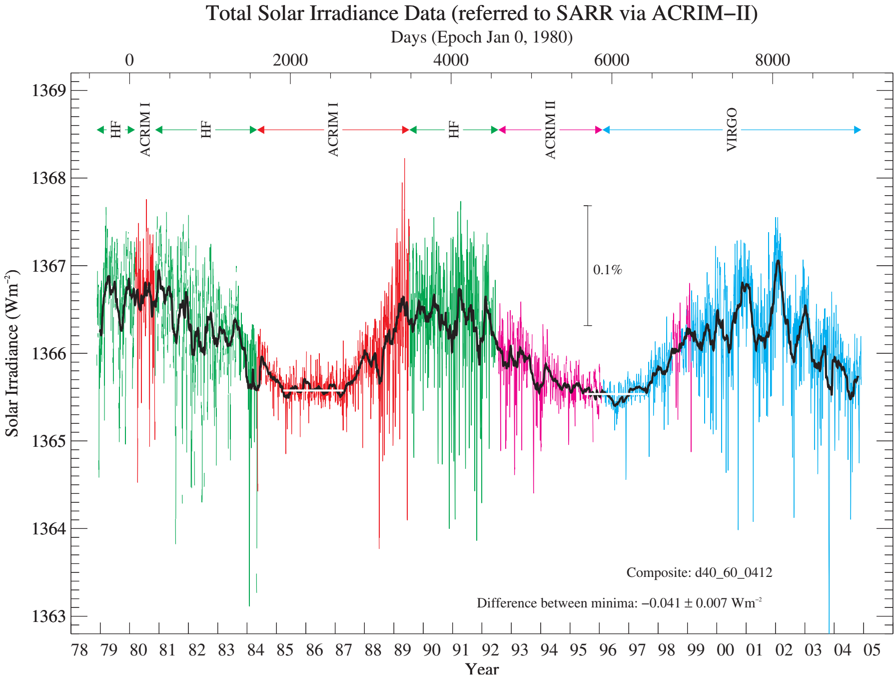


Figure 8.2. Composite total solar irradiance (TSI) with indication of which time series are used at different times (Fröhlich and Lean, 1998).

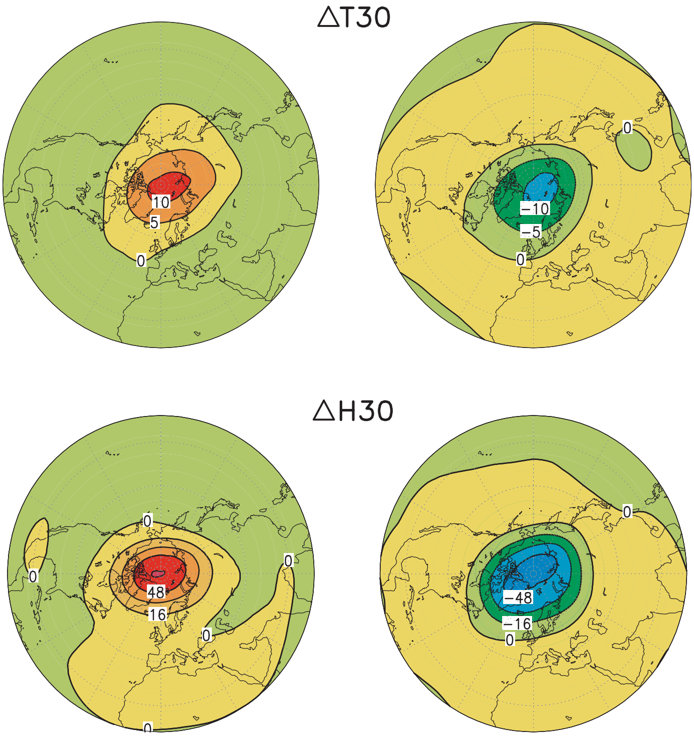


Figure 8.6. (Top) Deviations of 30-hPa monthly mean temperatures, February (K). (Bottom) Deviations of 30-hPa monthly mean heights (decameters). Both are deviations from the February mean of the years 1965–1974. (Left) for the four Warm Events 1966, 1977, 1973, 1987. (Right) For the four Cold Events 1965, 1974, 1976, 1997. (Labitzke and van Loon, 1999.)

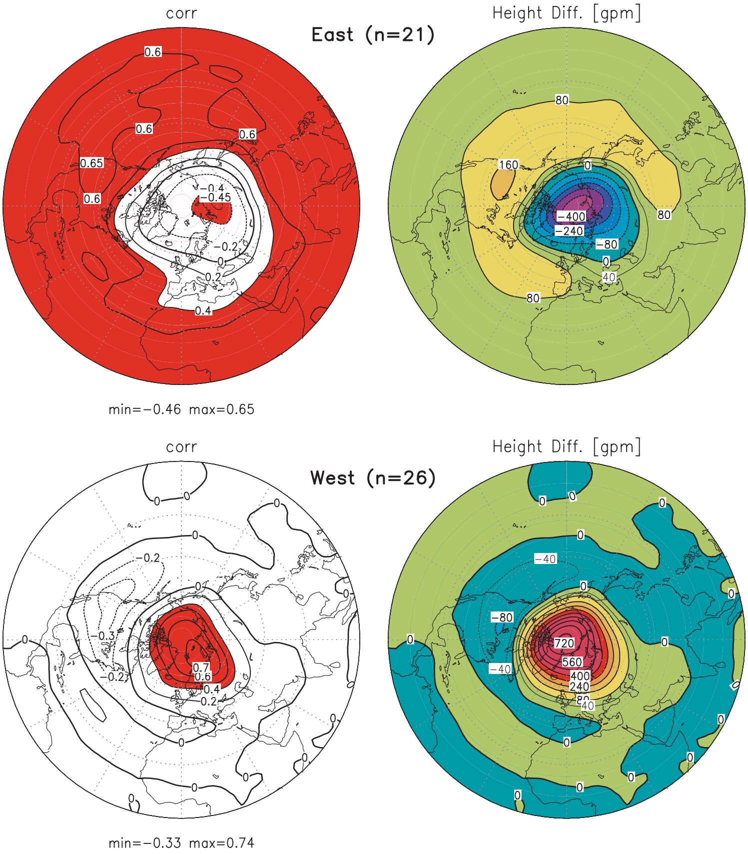


Figure 8.7. Northern hemisphere. (Left) Correlations between the 10.7-cm solar flux (the 11-year solar cycle) and 30-hPa heights in February, shaded for emphasis where the correlations are above 0.4; (upper panel) years in the east phase of the QBO; (lower panel) years in the west phase of the QBO. (Right) Respectively, height differences (geopot. m) between solar maxima and minima. (NCEP/NCAR re-analyses, period: 1958–2004; Labitzke 2002, Figure 7, updated.)

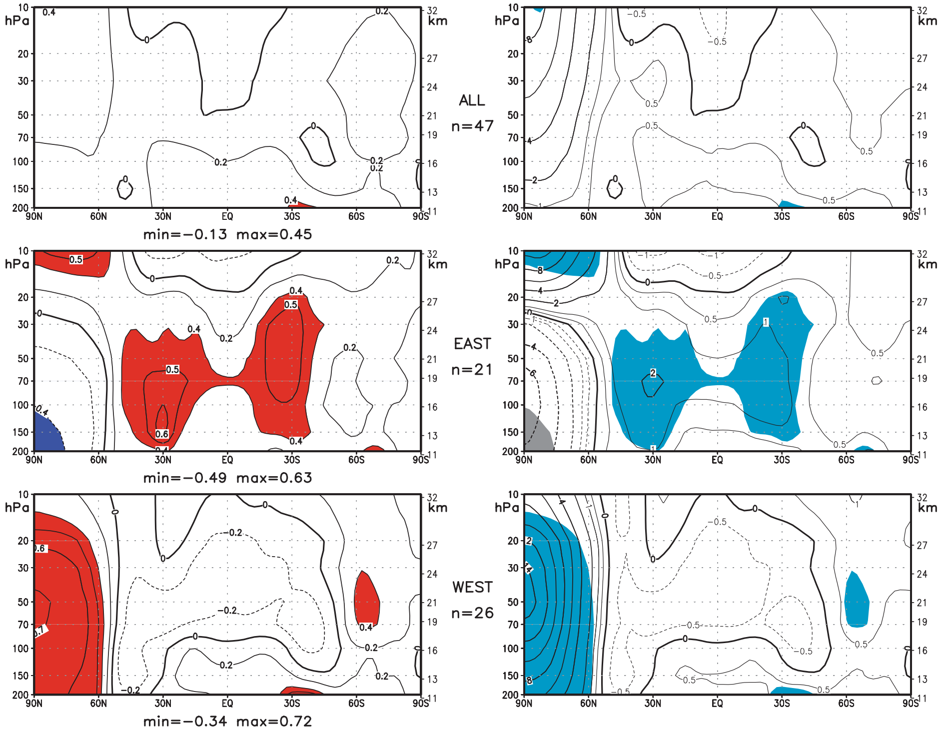


Figure 8.8. Vertical meridional sections between 200 and 10 hPa (11 and 32 km) of (left) the correlations between the detrended zonally averaged temperatures for February and the 10.7 cm solar flux (shaded for emphasis where the correlations are larger than 0.4); and (right) the respective temperature differences (K) between solar maxima and minima, shaded where the corresponding correlation on the left-hand side are above 0.4. (Upper panels) all years; (middle panels) only years in the east phase of the QBO; (lower panels) only years in the west phase of the QBO. (NCEP/NCAR reanalyses, 1958–2004; Labitzke, 2002, updated.)

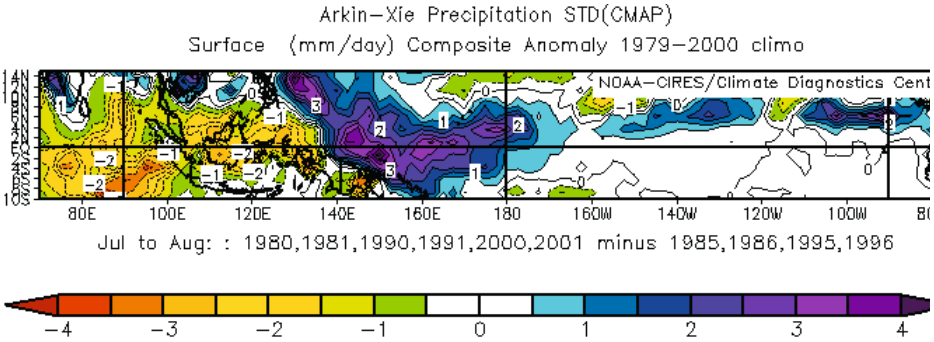


Figure 8.16. Difference in rainfall in July–August. The three solar maxima after 1979 minus the two solar minima, between 10° S and 14° N, and from 70° E to 175° W. (van Loon *et al.*, 2004.)

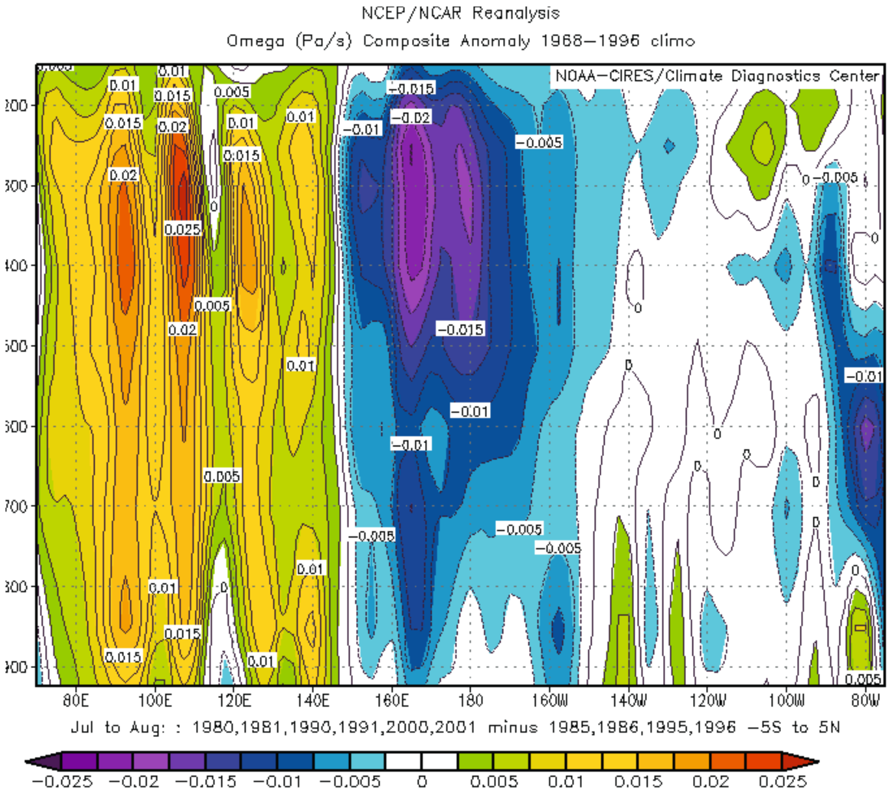


Figure 8.17. Vertical section (from 925 hPa to 200 hPa) of the difference in vertical motion (Pa s^{-1}) between the three solar maxima and two minima after 1979. From 70° E to 75° W, and between 5° S and 5° N, in July–August. (van Loon *et al.*, 2004.)

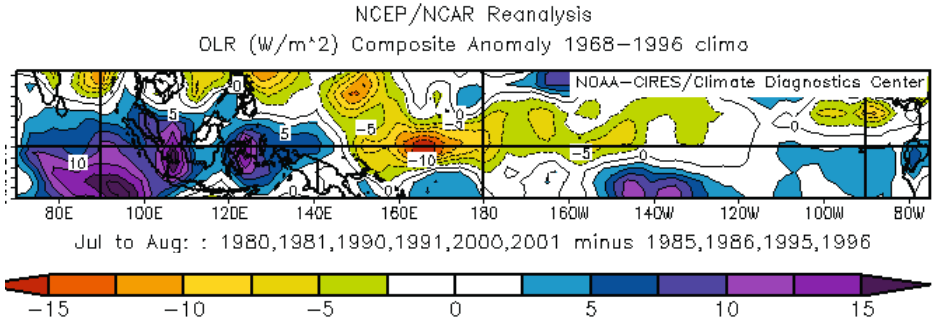


Figure 8.18. The difference after 1979 between solar maxima and minima of outgoing long-wave radiation (Wm^{-2}) in the areas of bipolarity: 10° S to 14° N, between 70° E and 175° W. (van Loon *et al.*, 2004.)

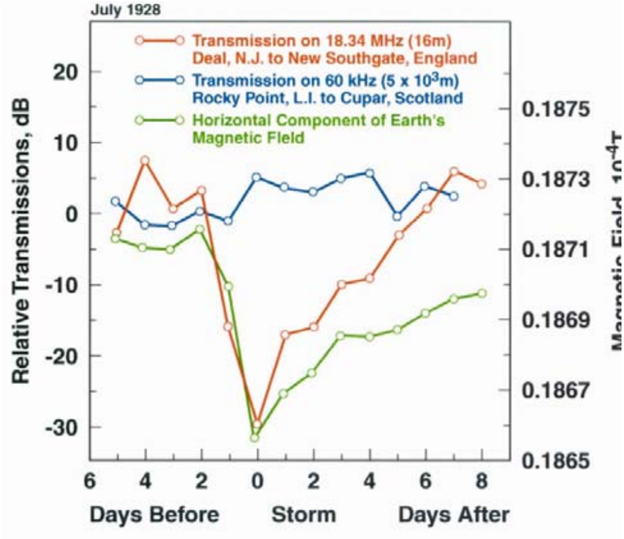


Figure 9.5. Transatlantic wireless transmissions from the Eastern US to the UK on two frequencies before and during a magnetic storm event in July 1928. Also shown are the values of the horizontal component of the Earth's magnetic field. (Anderson, 1929.)

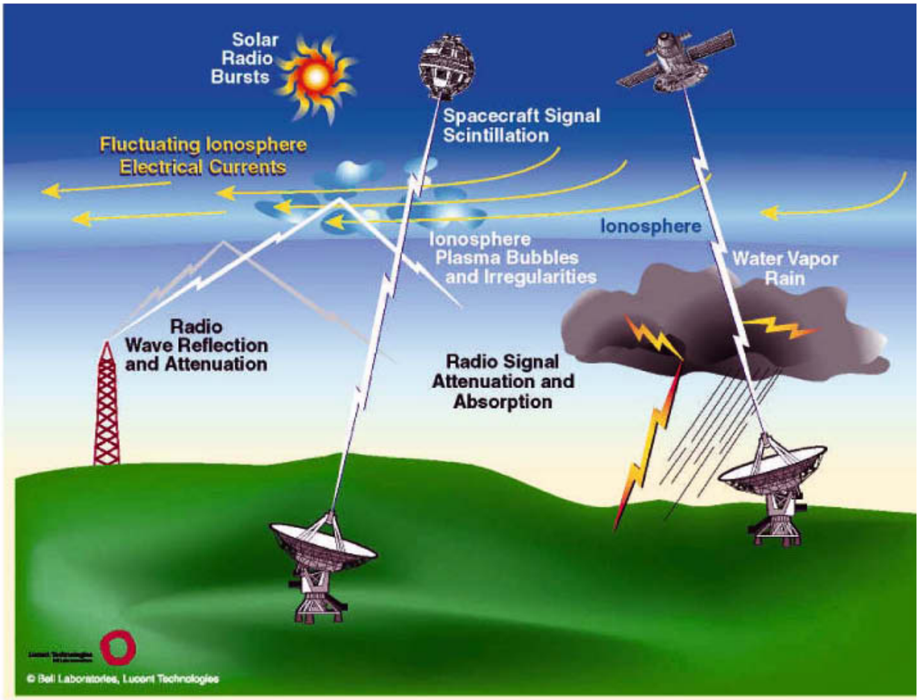


Figure 9.7. Some of the effects of space weather on communications systems that are deployed on the Earth's surface and in space, and/or whose signals propagate through the space environment.

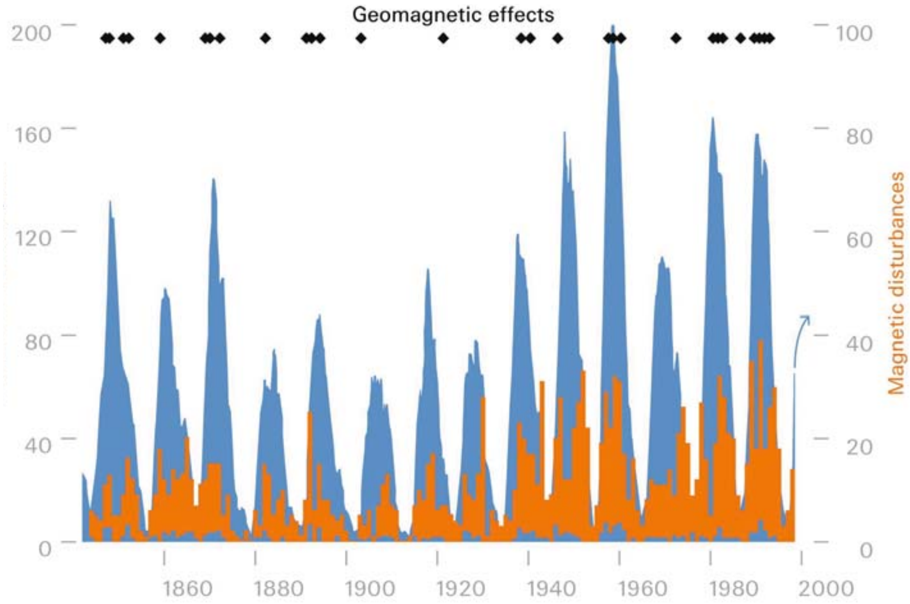


Figure 10.1. Statistically GIC phenomena follow the eleven-year sunspot number depicted by the curve. The bars show the number of geomagnetic disturbances. (The two colours correspond to different intensity levels.) The small white diamonds at the top of the diagram show the times of significant GIC effects. (Boteler *et al*., 1998; Jansen and Pirjola, 2004; Pirjola *et al*., 2005.)

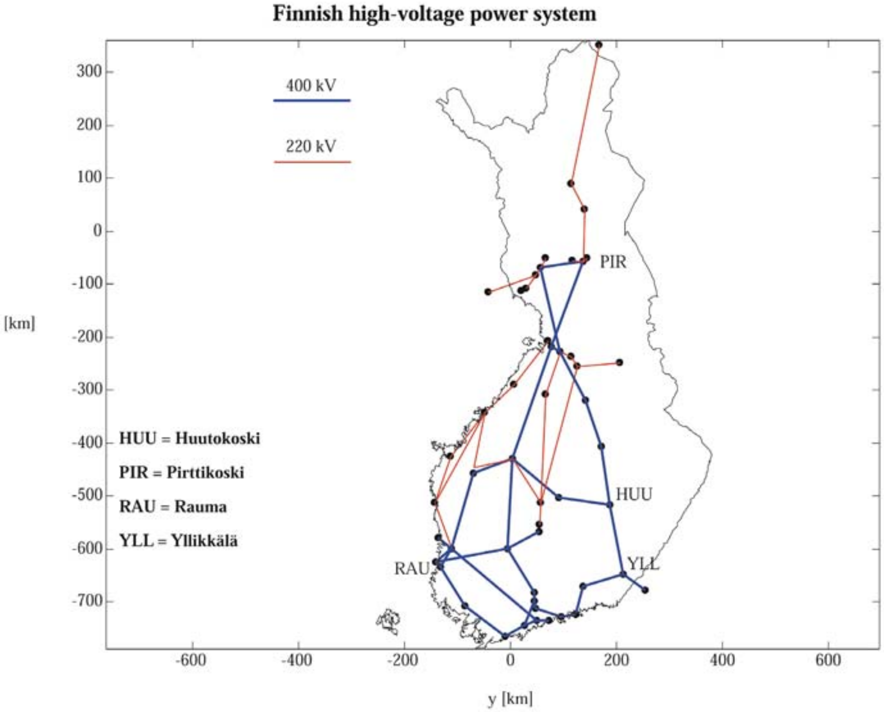


Figure 10.3. Finnish 220-kV and 400-kV electric power transmission systems. The 400-kV grid has connections to Sweden in northern Finland, but they are not shown in the figure because the transmission lines are interrupted by series capacitors preventing the flow of GIC. However, all lines in Finland are shown, even though some of them also have capacitors.

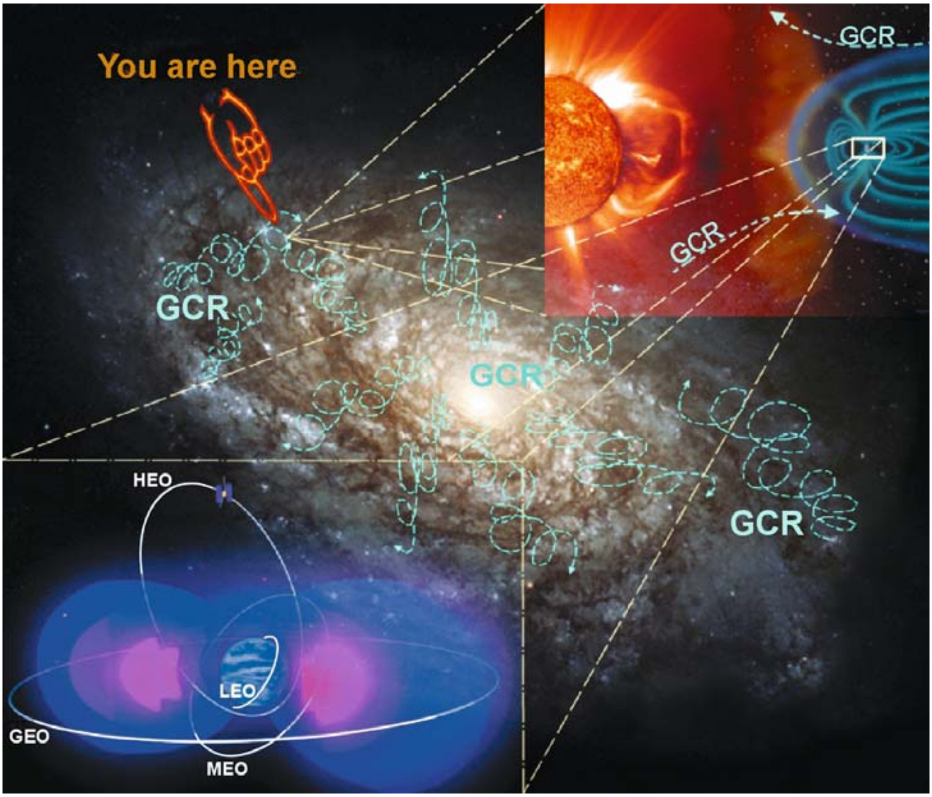


Figure 11.1. The three major components of ionizing space radiation: galactic cosmic rays (GCR), solar particle-radiation, and geomagnetically trapped radiation.

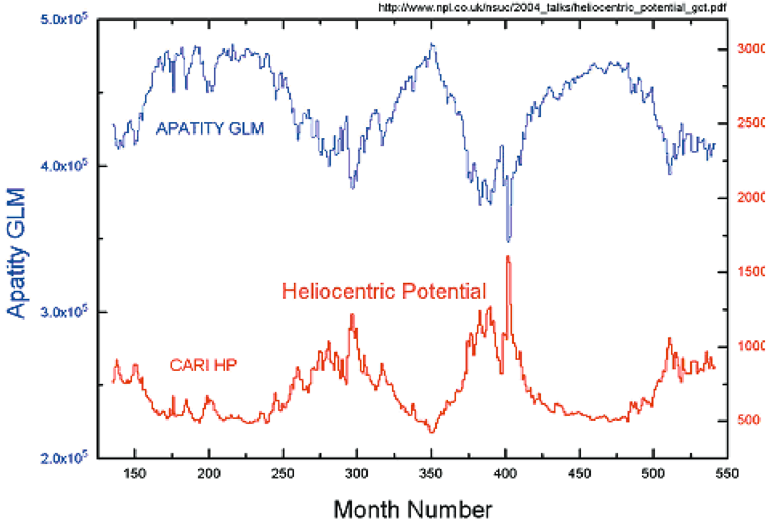


Figure 11.16. Temporal covariation of the heliocentric potential with the neutron flux in the Apatity monitor.

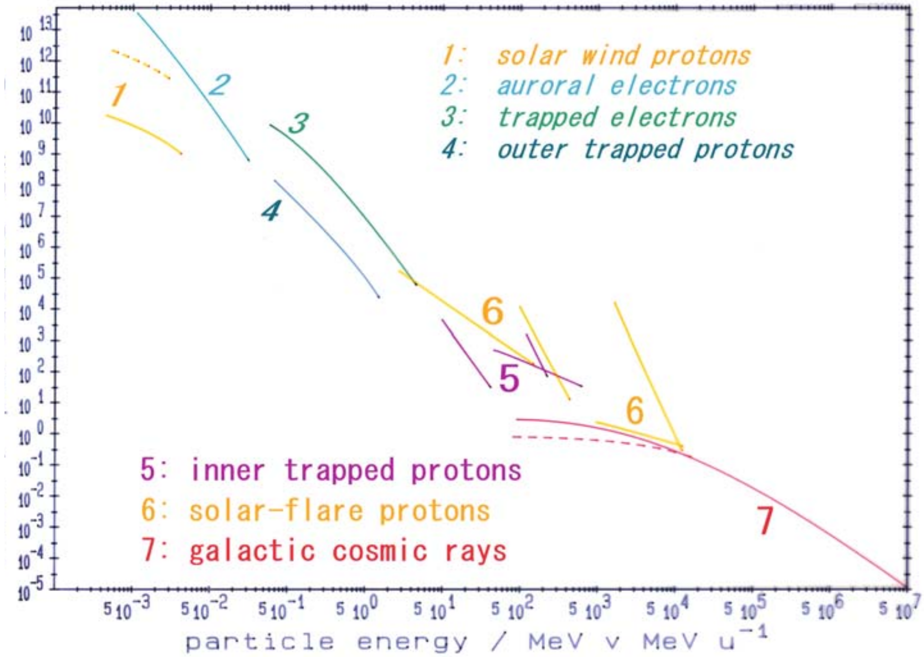


Figure 11.18. Synoptic view of particle and energy spectra of space radiation.

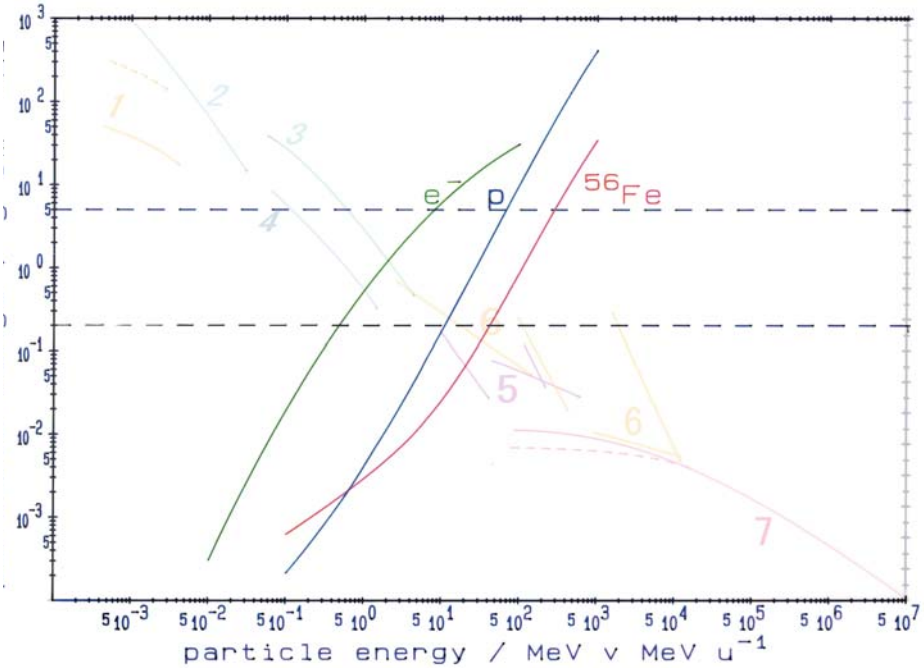


Figure 11.19. Selection of space radiation components relevant for radiation protection due to their range in shielding material.

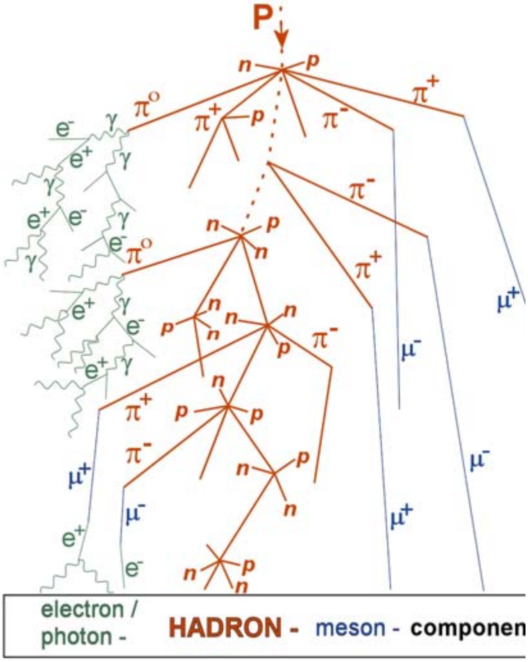


Figure 11.20. The typical components of cosmic-ray air showers, arranged according to their radiobiological quality.

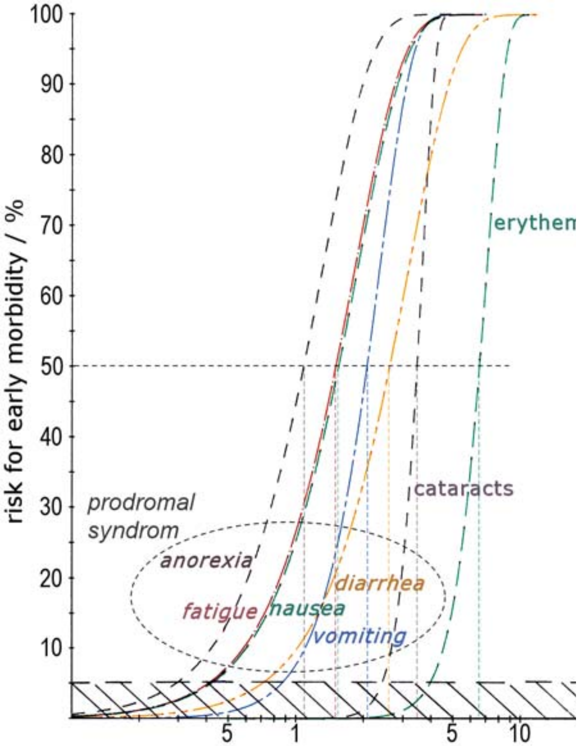


Figure 11.29. Early human morbidity as a function of acute whole-body exposure to sparsely ionizing radiation.

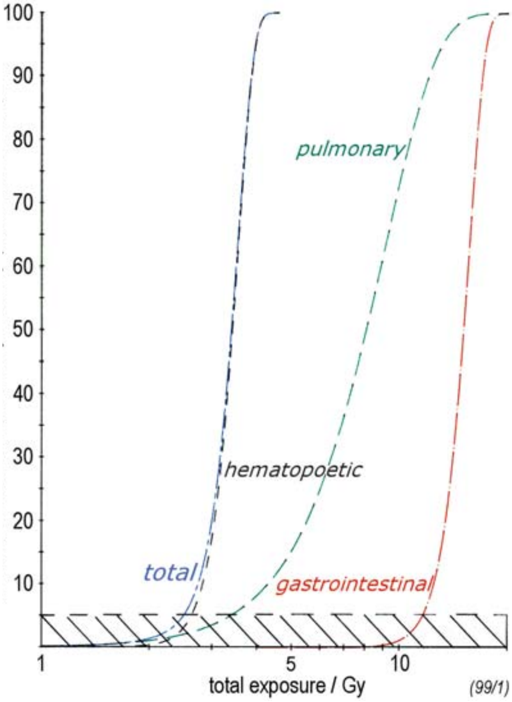


Figure 11.30. Early human mortality as a function of acute whole body exposure to sparsely ionizing radiation.

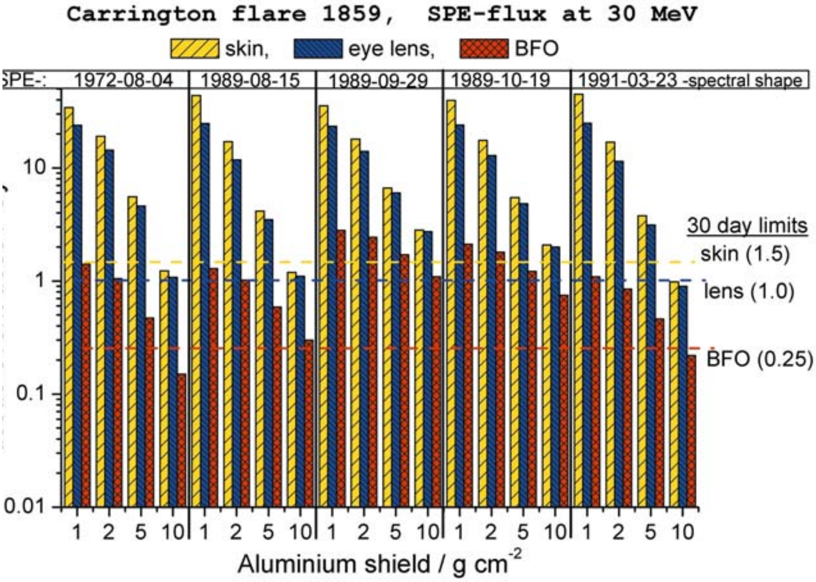


Figure 11.38. Total event doses in critical tissues for spectra of known large SPE, as matched to the >30 MeV proton flux of the Carrington 1859 solar flare. Present NASA limits for deterministic effects from acute exposures are specified in Sv. The comparison presumes an RBE of 1 for protons and the associated effects.

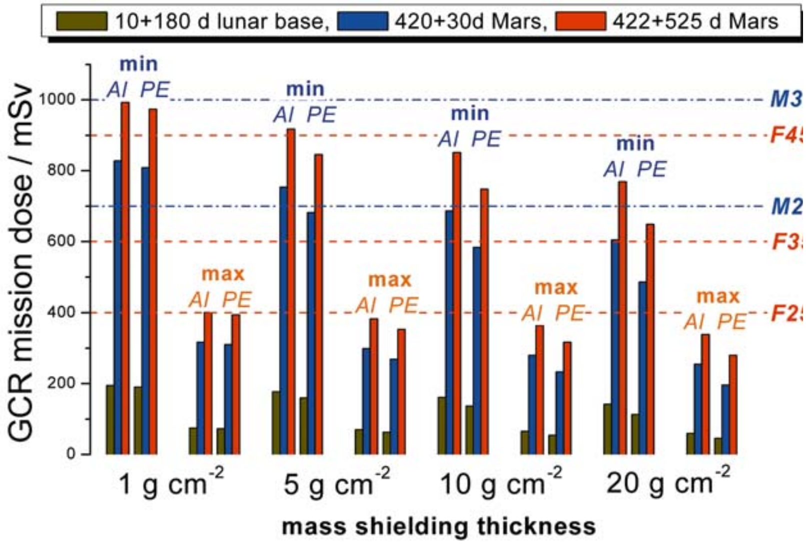


Figure 11.39. Estimated BFO mission doses from galactic cosmic rays for long-term missions outside the magnetospheric shield during solar minimum (min) and maximum (max) phases. Shields are either aluminium (Al) or polyethylene (PE) as a hydrogen-rich material. Horizontal lines specify NASA radiation protection career limits established for low Earth orbits for male (M-) or female (F-) crew-members of age 25, 35 or 45 at mission beginning.

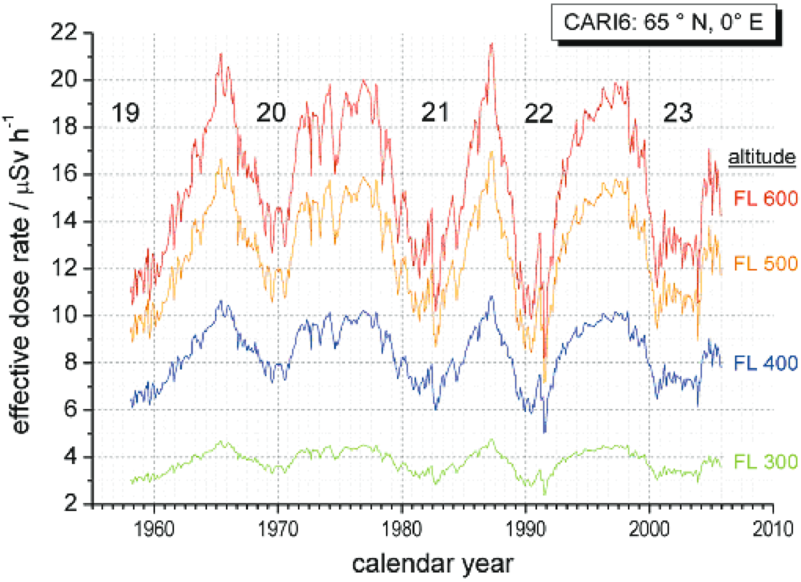


Figure 11.43. Temporal variation with the solar cycle of air crew exposures for typical ranges of cruising altitudes of commercial and corporate jets in the polar plateau region (CARI calculations).

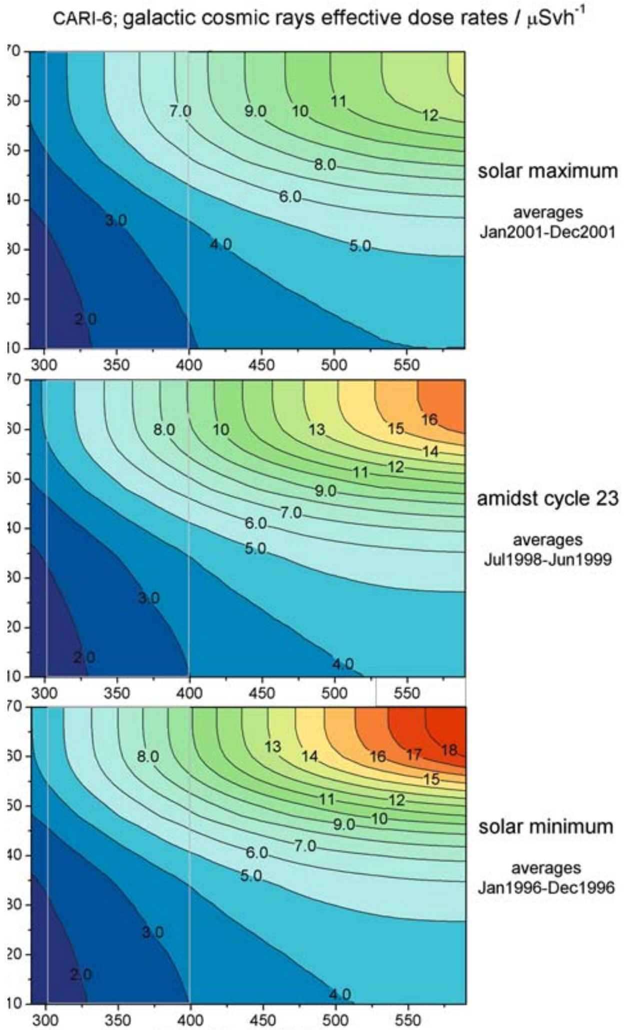


Figure 11.44. Altitude, latitude and solar cycle dependence of GCR-related exposure to ionizing atmospheric radiation (AIR), as calculated by the CARI programme. The range of commercial jet cruising altitudes is marked.

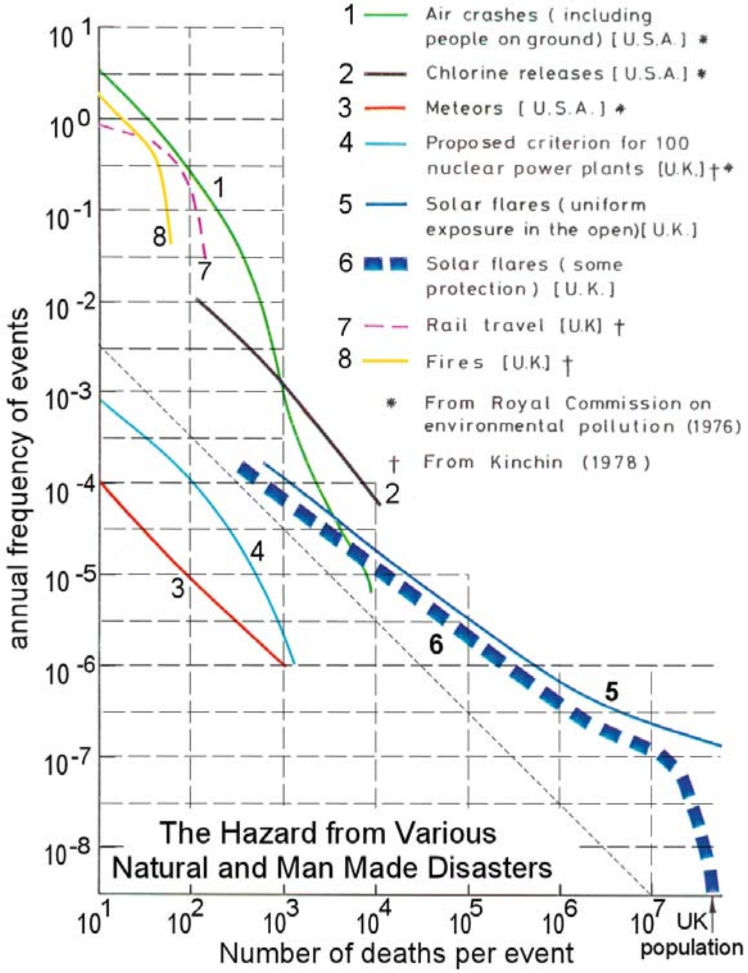


Figure 11.45. Comparison of estimated probabilities for human radiation deaths near sea level from extreme solar particle events with common terrestrial risks.

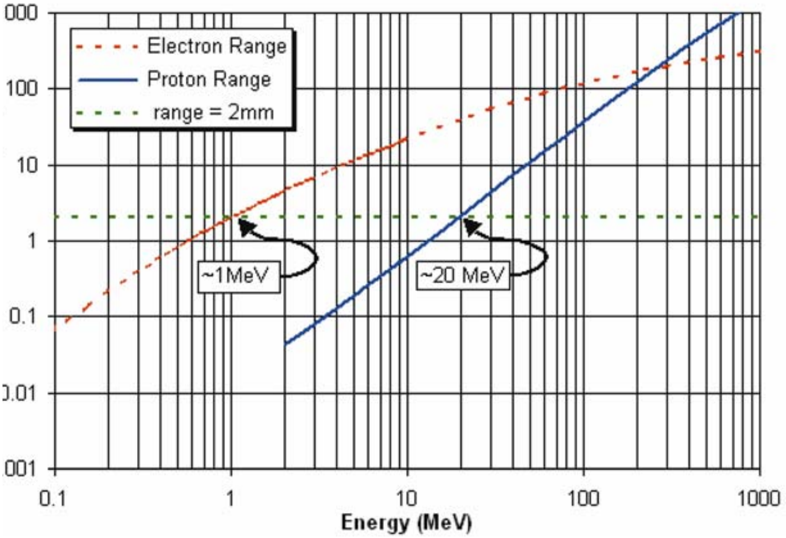


Figure 12.1. Ranges of electrons and protons in aluminium.

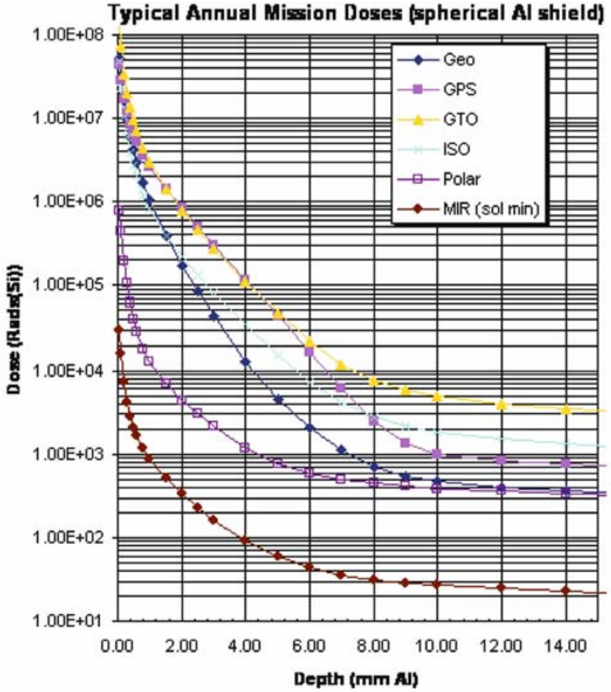


Figure 12.2. The dose expected behind a sphere of aluminium shielding for various satellite orbits, as functions of the sphere radius.

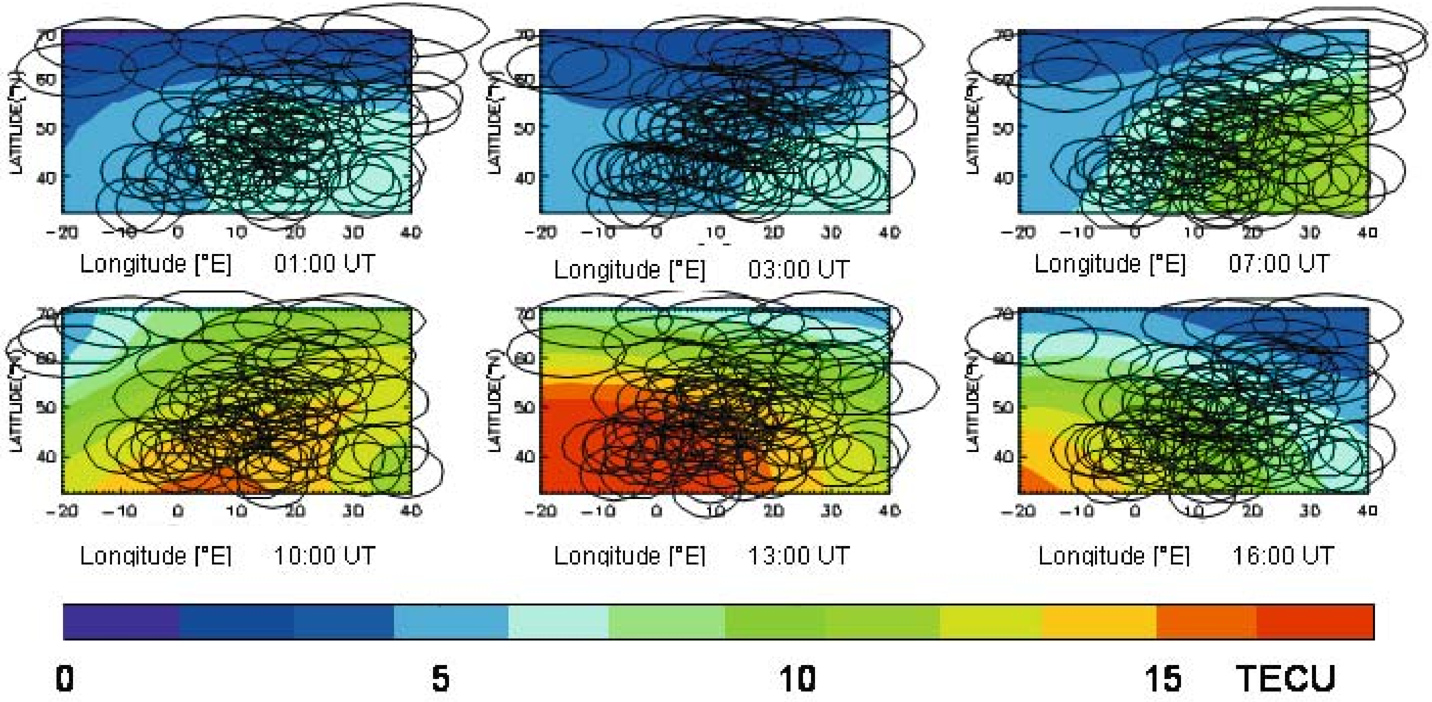


Figure 13.3. TEC maps over Europe on 10 January 1997. The ellipses mark the half width of the weighting functions constructed around the piercing points while assimilating the data into the background model (e.g., [11]).

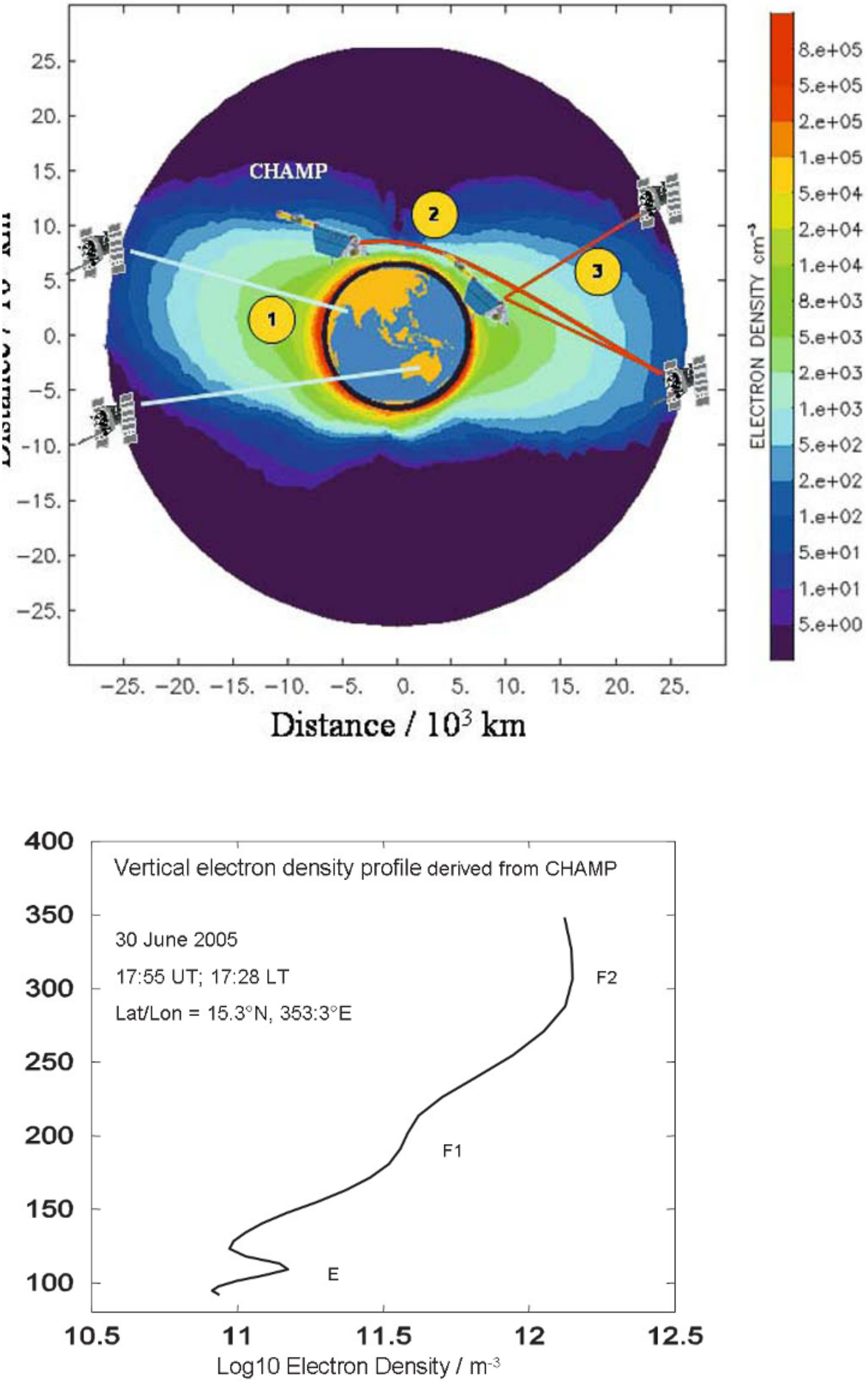


Figure 13.4. Illustration of GNSS-based ionospheric monitoring techniques: (1) ground-based TEC, (2) radio occultation, and (3) topside TEC which is used for reconstructing the electron density from CHAMP data as seen in this figure for 21 December 2001 (top panel, [14]) and an electron density profile retrieved on 30 June 2005 from CHAMP data (bottom panel, [15]).

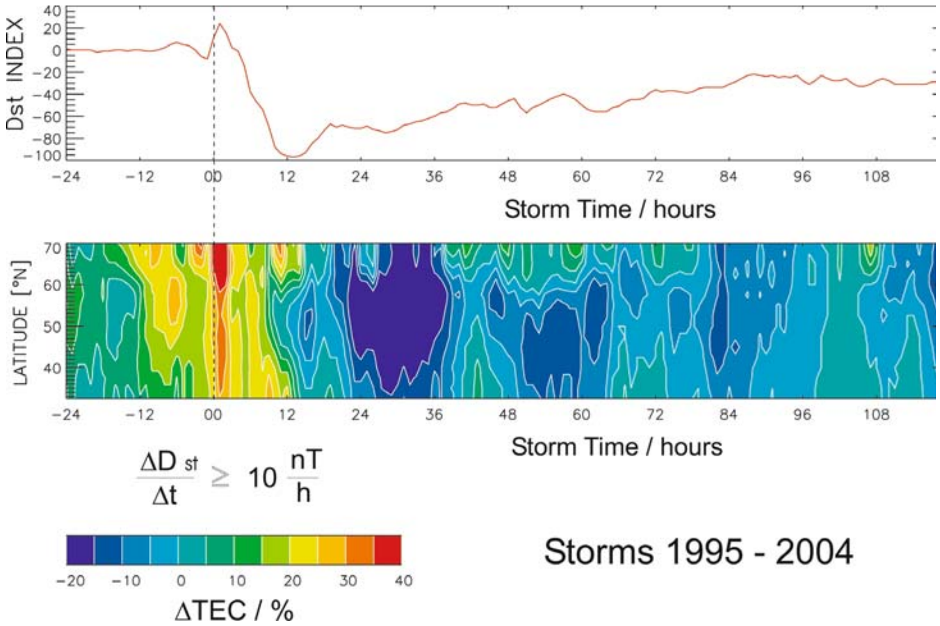


Figure 13.8. Latitudinal dependence of the average storm time pattern of TEC over Europe (15°E meridian). The upper curve depicts the disturbance storm time (*Dst*) index in nT and the lower graph shows the averaged percentage deviation of TEC for the period starting 24 hours before and ending 120 hours after the onset of the storm.

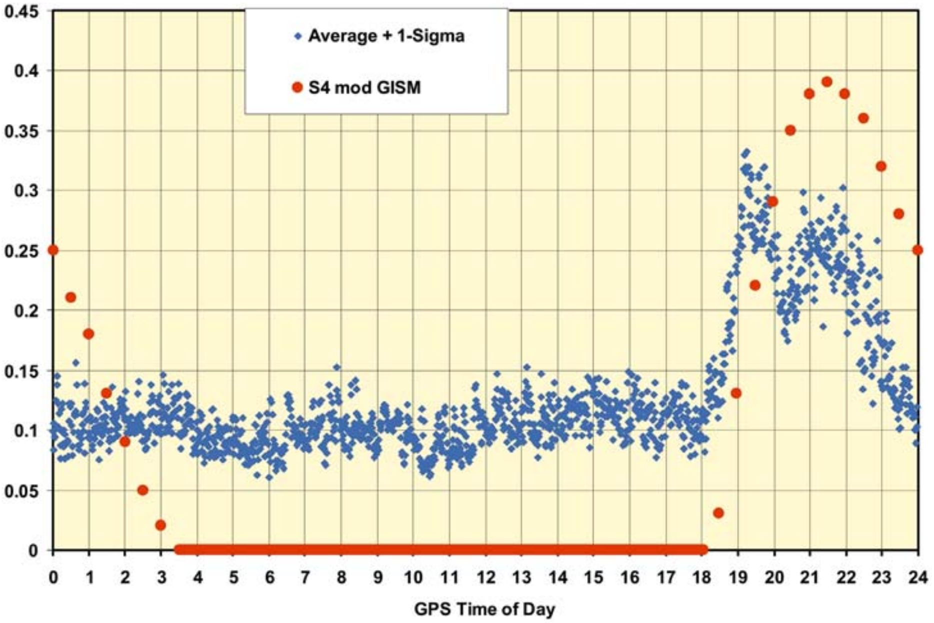


Figure 13.10. Amplitude scintillation index S_4 measured in Douala (blue dots) and predicted S_4 using the Global Ionospheric Scintillation Model (GISM, red dots) [31].

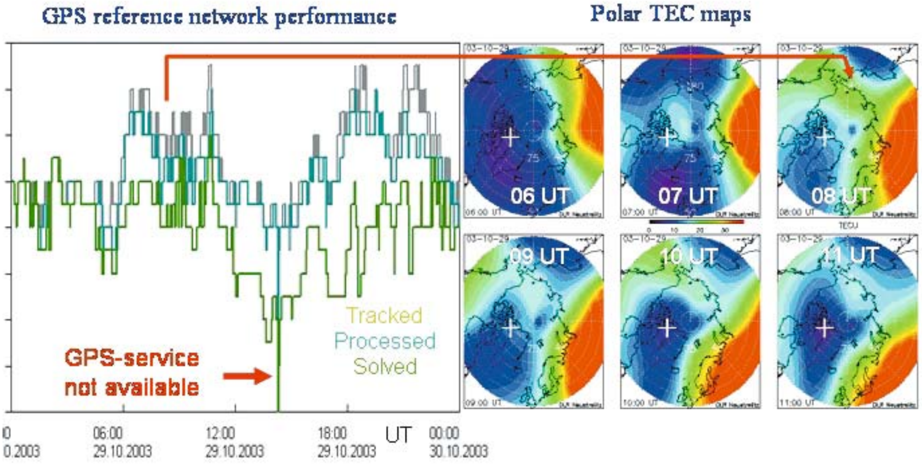


Figure 13.11. Ionospheric storm effects on the number of tracked, processed, and solved GPS/GLONASS satellites within the ascos network in Germany (source: Allsat GmbH , Hannover) in comparison with hourly polar TEC maps.

Figure 14.1. Space weather forecast centres issue forecasts on all space weather environments. Here the NOAA/SEC forecast centre is shown. The photograph was taken in 2004.

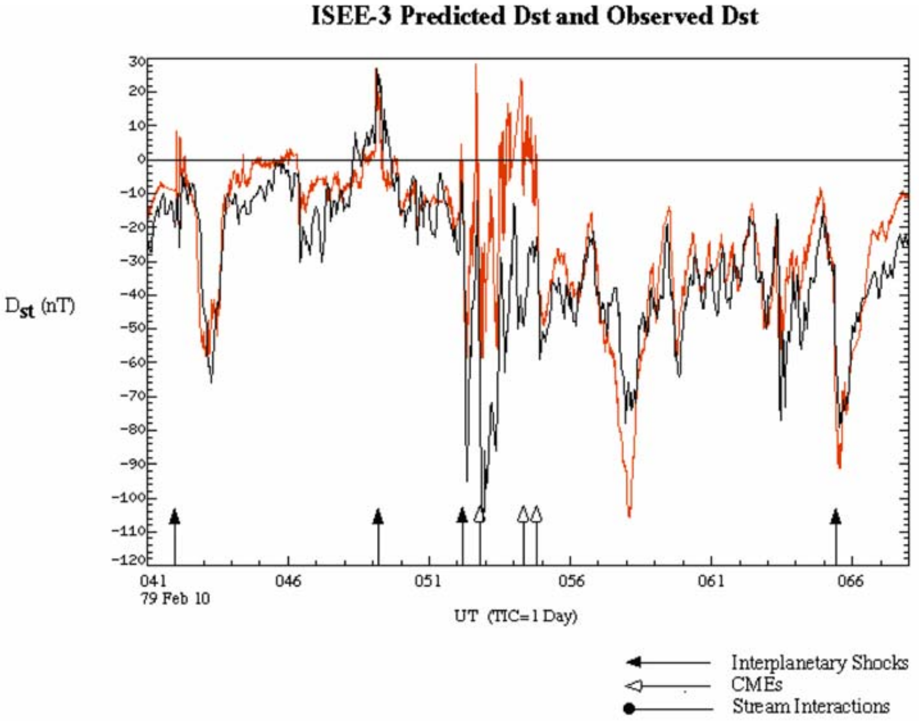


Figure 14.2. D_{st} index prediction from upstream solar wind measured spacecraft ISEE-3 (after Lindsay *et al.* (1999), who used the Burton *et al.* (1975) model, summarized in equation (14.6) in the text).

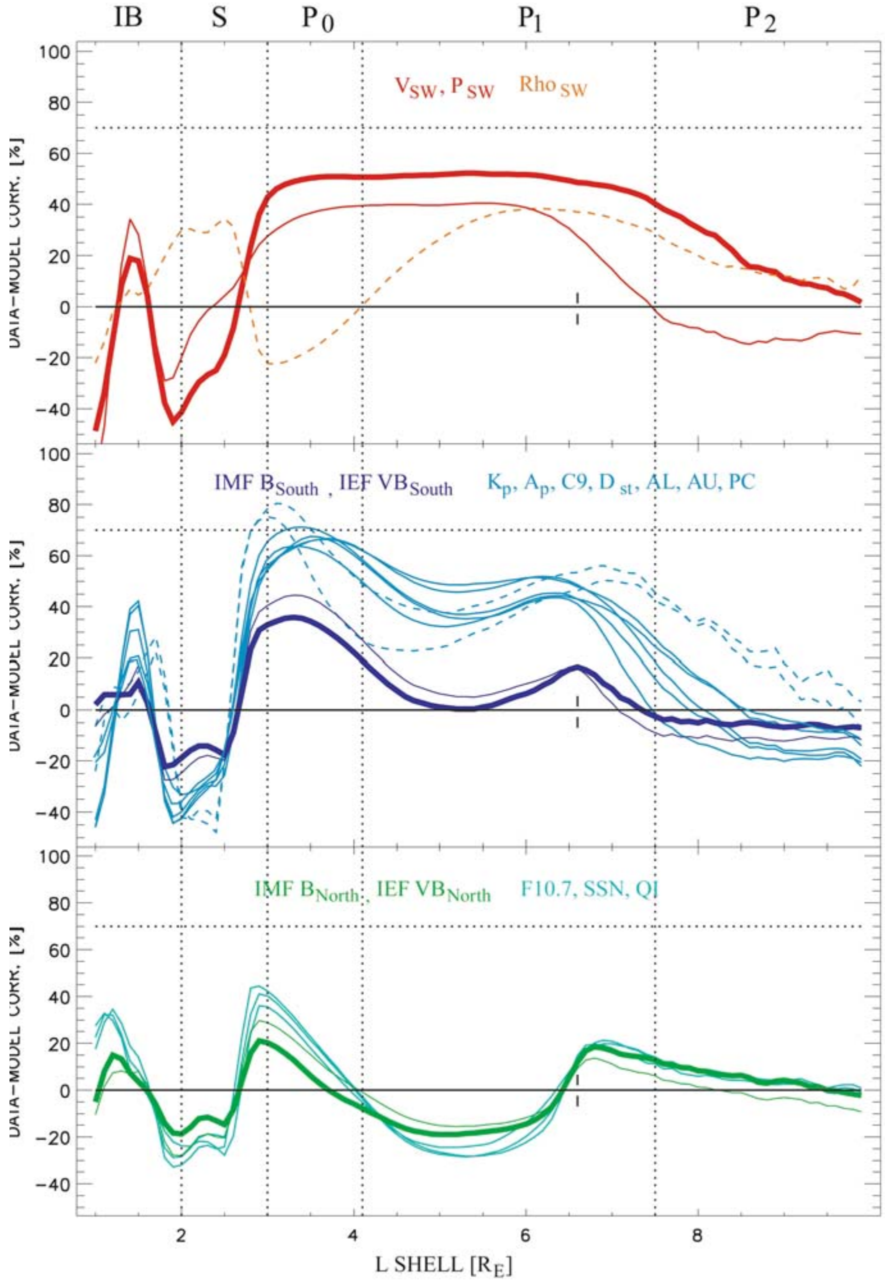


Figure 14.3. Solar, interplanetary and magnetospheric variable classification of their impact on electron flux using an empirical model and the correlation function. An FIR model driven with each variable produces a forecast electron flux. The correlation function (14.7) for observed and predicted electron flux at shell L is plotted as a function of L. Correlation curves are grouped in three different categories, each plotted in a different panel.

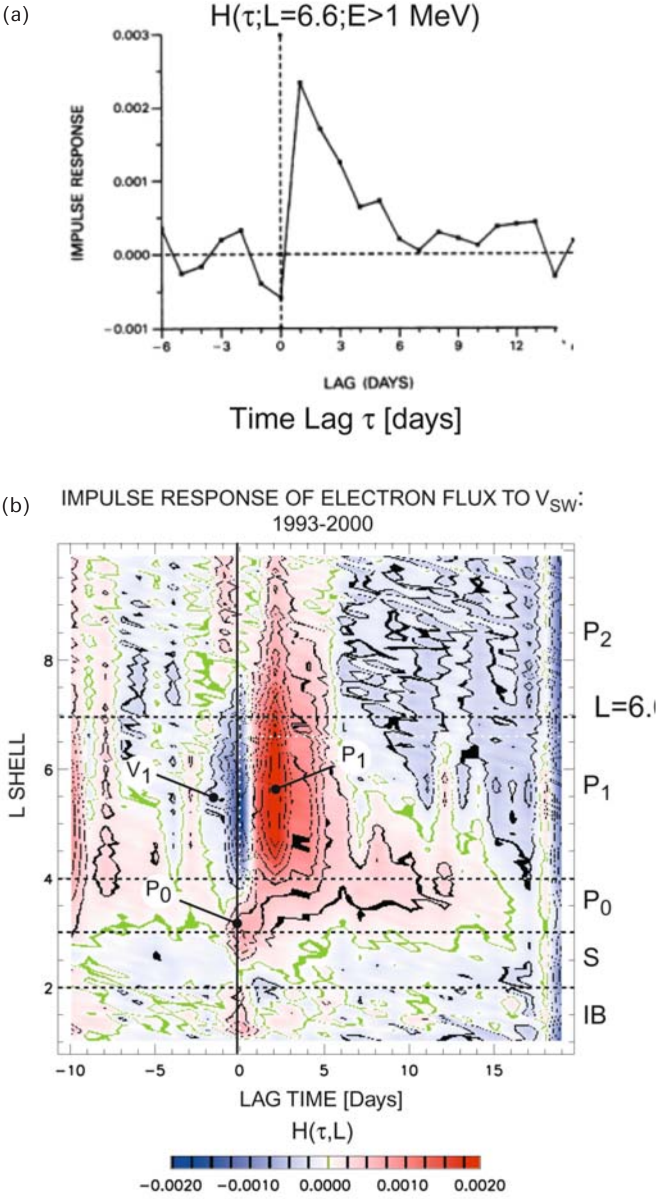


Figure 14.4. (a) Impulse response of energetic electron flux j_e, at geosynchronous orbit ($L = 6.6$) and energies larger than 1 MeV, as measured by LANL spacecraft, to the upstream solar wind plasma velocity V_{SW} (Baker *et al.*, 1990). (b) As above, but for all L in the range 1–10 and $E = 2$–6 MeV as measured by SAMPEX to V_{SW} (Vassiliadis *et al.*, 2002).

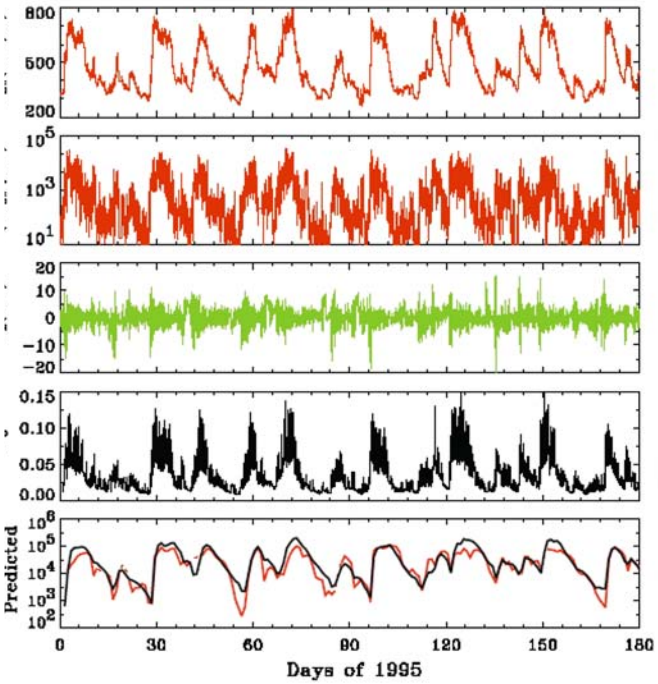


Figure 14.5. A forecast of the relativistic electron flux at geosynchronous orbit based on the diffusion model of Li *et al*. (2001). Solar wind parameters and the D_{st} index are displayed in the top panels.

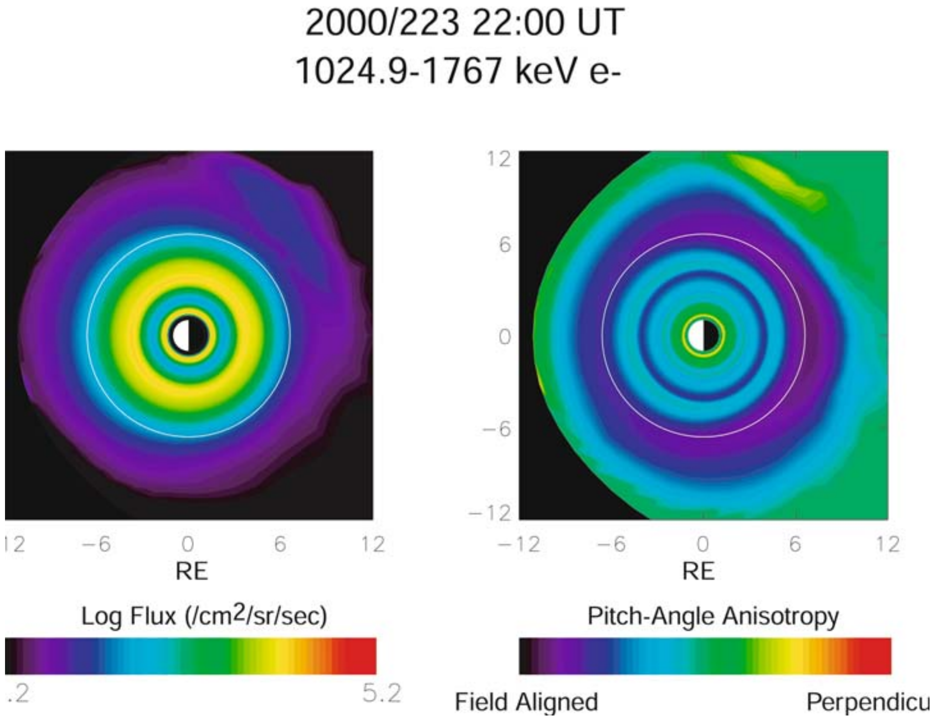


Figure 14.6. The equatorial cross-section of the energetic electron flux in the inner magnetosphere as predicted by the Fok *et al*. convection–diffusion model, implemented in collaboration with the University of Alaska/GI and Johns Hopkins University/APL.

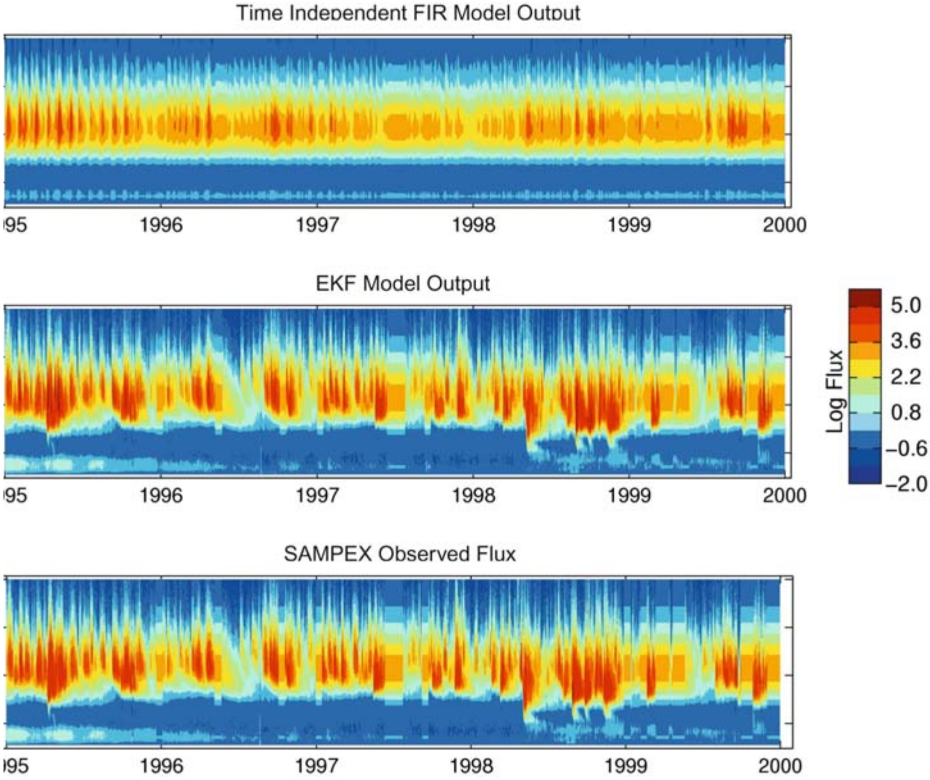


Figure 14.7. Assimilation of SAMPEX/PET 2–6 MeV daily electron flux via an extended Kalman filter. The flux is shown as a function of time and *L* shell. The first panel displays the original (uncorrected) FIR model output; the second, the Kalman-filter version; and the third, the SAMPEX observations. Note the improvement in next-day forecast accuracy as quantified by the prediction efficiency (after Rigler, 2004).

Printing: Mercedes-Druck, Berlin
Binding: Stein + Lehmann, Berlin