ASTROPHYSICS AND SPACE SCIENCE LIBRARY 344

Jean Lilensten
Editor

Space Weather

Research towards Applications in Europe

esa

COST 724

ASSL

Springer

SPACE WEATHER

ASTROPHYSICS AND SPACE SCIENCE LIBRARY

VOLUME 344

SPACE WEATHER

RESEARCH TOWARDS APPLICATIONS IN EUROPE

Edited by

JEAN LILENSTEN
Laboratoire de Planétologie de Grenoble,
France

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-5445-9 (HB)
ISBN-10 1-4020-5446-7 (e-book)
ISBN-13 978-1-4020-5445-7 (HB)
ISBN-13 978-1-4020-5446-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

TABLE OF CONTENTS

Session 4: Radiation Environment of The Earth/Spacecraft and Aircraft Environment

Session 5: The Magnetic Environment GICs and Other Ground Effects

INTRODUCTION

J. LILENSTEN[1], A. GLOVER[2], A. HILGERS, A. BELEHAKI, Lj.R. CANDER, B. ZOLESI, M. RYCROFT, AND F. LEFEUVRE

[1]*Chair of ESWW2, editor*
[2]*co-Chair of ESWW2*

This book presents Space Weather review papers given as invited presentations at the Second European Space Weather week which was held at the European Space Agency's ESTEC site in Noordwijk, The Netherlands, from 14th to 19th November, 2005. As the meeting itself, it is divided into 5 chapters, each representing one of the "Science to Applications" sessions on a particular theme. These themes are:

- Session 1 : The solar weather/solar activity forecast and predictions (including propagation of transient features in the solar wind)
- Session 2 : Atmospheres (weather and climate, also including thermosphere and drag), global change
- Session 3 : Ionosphere/positioning and telecommunication
- Session 4 : Radiation environment of the Earth/spacecraft and aircraft environment
- Session 5 : Magnetic environment/GIC's and other ground effects

Each chapter is divided into an introduction written by the convener of the corresponding session, and the papers written by the invited speakers. Many of the poster contributions, including applications focussed papers, have been published elsewhere and the presentation material can be found via the ESWW2 website: http://www.esa-spaceweather.net/spweather/workshops/eswwII/esww2-proceedings.html.

The ESWW2 was the second in a series of annual workshops that are the result of much patient work that started about ten years ago involving many European entities. In 1996 ESA organised a round table on Space Weather to discuss possible

options for a European counterpart to the US National Space Weather Programme. The first ESA Space Weather workshop took place two years later. At this time, the community was developing and the perspective for a coordinated effort in the field of Space Weather was being investigated.

ESA has maintained a key role in structuring the field in Europe, primarily by organising a yearly workshop from 1998 to 2003 and by sponsoring a number of studies geared towards establishing the feasibility of a European Space Weather Programme. In parallel to these studies, an advisory committee was created, the Space Weather Working Team (SWWT), successively chaired by W. Riedler, R. Gendrin, F. Lefeuvre and nowadays by M. Hapgood. One of the early recommendations of the SWWT was to create a European COoperation in the field of Scientific and Technical Research action targeted at the science underpinning Space Weather. The COST proposal was accepted by the Brussels administration and was inaugurated in November 2003 under the number COST 724 Action on "Developing the Scientific Basis for Monitoring, Modelling and Predicting Space Weather". In 2005, it coordinates space weather theoretical related efforts of 26 countries. In addition, the COST 296 Action on "Mitigation of Ionospheric Effects on Radio Systems" is also related to the effects of space weather on the ionosphere and radio wave propagation.

In 2003 ESA embarked upon a Space Weather Applications Pilot Project. The aim of this project is to develop and extend the Space Weather user community through the development of targeted services, provided by a network of service providers, supported by a common infrastructure and using data from existing or easily adaptable assets. Through this service and community development a long-term view of the potential for space weather applications is currently being established.

The Space Weather Pilot Project consists of three main components:

1. a network of service development activities (SDA) involving service providers, users and data providers,
2. an infrastructure development, coordination and support activity, and
3. a quantitative evaluation of the costs and benefits of a European Space Weather service, based on, but not limited to, information received from the network of participants in the pilot project.

At the time of writing, the pilot projects incorporate 17 ESA co-funded Service Development Activities (SDAs) and a number of additional independently funded SDAs, each focusing on a wide range of space weather user domains. In addition to the individual service activities, a service support infrastructure was created. Together with the SDAs, this infrastructure, and consequently the pilot project as a whole, is named the "Space Weather European NETwork" (SWENET). This activity provides support to the SDA activities, and forms the centralised web based access point to a coordinated network of European Space Weather services. This service network activity also benefits from strong collaboration with the International Space Environment Service and the NOAA Space Environment Center.

In view of so many collaborative efforts, it was natural that the various initiatives combine their efforts to organise a joint European Space Weather Week, building on the existing series of ESA Space Weather Applications Workshops and with the aim of bringing together the diverse communities involved in Space Weather research and applications. The fact that the organising committee included participants of both scientific and applications focussed initiatives shows their degree of confidence and good will to collaborate.

This book forms one of the key outputs of the ESWW2 and constitutes a milestone in the building of the field of Space Weather in Europe. It shows how dynamic the research community within Europe is and also demonstrates knowledge and know-how at the highest international level. The contribution to this book from experts across Europe demonstrates the profound collaboration that already exists in this field and which it is an aim of this meeting to encourage. It is also an answer to the question of the profound status of the discipline: space weather relies both on science and application. In this book, the reader will find *theoretical* papers as well as papers dealing with application and service developments in various industrial domains.

The signatories of this introduction represent the Organising Committee of the ESWW2. In addition, the organisation of the meeting and publication of the enclosed material would not have been possible without the efforts of the session convenors, each of whom dedicated considerable time and effort towards organising the individual sessions and collecting the final papers. We also gratefully acknowledge the efforts of several referees who took time to review all of the scientific papers published here. We hope that the reader will find the enclosed material stimulating and that it will lead to further collaborative efforts in the field of Space Weather.

CHAPTER 1.0

THE SOLAR WEATHER/SOLAR ACTIVITY MONITORING AND FORECAST

RONALD VAN DER LINDEN
Royal Observatory of Belgium, Ringlaan 3, B-1180 Brussel (Belgium)
ronald.vanderlinden@oma.be

INTRODUCTION

The fundamental source of most of the perturbations studied in Space Weather research resides in the behavior of the Sun, our star. It can be found on the short time scale in the wide variety of solar activity, and, linked to that but on a longer timescale in the variability of the solar output. Solar activity spans a wide range of timescales, from the secular modulation of the well-known 11-year solar activity cycle, over the 27 days of solar rotation, down to sub-second timescales during eruptions. The most complete understanding of space weather and the largest possible leeway in the timescale of forecasting can thus be gained by studying these solar drivers of space weather.

Roughly speaking, there are three ways through which solar activity influences the earth and its environment: electromagnetic waves (radiation), high-energy particle fluxes and coherent plasma flows or clouds. Most of the energy we receive from the sun comes in the form of *electromagnetic waves* (radiation). The largest part of this is in the visible light, where mostly very little variation is seen due to solar activity (although we should remember that ancient and not-so-ancient chronicles exist that indicate that large solar flares can lead to significant short-term increases even in this range of the spectrum). Sunspots are seen in these wavelengths, but have little effect overall on the energy balance and certainly no consequences on the short term. Thus, although sunspots are one of the oldest testimonials of solar activity, their appearance and evolution is relevant only as a proxy to the associated, much larger variation in the high-energy wavelengths and to other manifestations of solar activity. Indeed, the strongest (and for human technology most problematic relative variations) are found in the EUV and X-ray bands, where the solar output varies by orders of magnitude due to outbursts in the solar atmosphere known as *solar*

J. Lilensten (ed.), Space Weather, 1–3.

flares. Such variations of high-energy radiation lead to changes in the degree of ionization of the upper layers of the earth's atmosphere (the ionosphere), which in turn modifies the transmissivity and reflectivity of the ionosphere for radio waves. Or in other words: radio communication is strongly affected by the changes induced by solar flares. This is but one important example of the consequences of this type of solar activity.

Important to realize is that there cannot be a prior warning for increased EUV and X-ray radiation due to solar flares from observations alone: since the emissions form part of the electromagnetic wave spectrum, they travel to the earth at the speed of light and arrive after about 8 minutes. Therefore, to be able to give advance warning of the EM radiation increases, one needs to be able to forecast the occurrence of solar flares. A lot of attention is devoted to this, from the operational, observational and theoretical sides. Operationally, one tries to recognize the evolution of complex sunspot groups and magnetic fields that lead up to eruption; observationally, one tries to determine precursor signals using image processing and other techniques, theoretically, one tries to model the evolution of the solar plasma and predict its points of criticality. In the paper presented by Peter Gallagher et al., a review is presented of some of the tools that exist or are under development for the forecasting of solar flares based on image recognition.

The second main source of Space Weather perturbations relevant for humans and technology are energetic particles. These constitute a real danger to the safety of humans in space and are considered a threat also for the health of crew and passengers on air transport. In addition, they endanger the functioning of spacecraft. Increased fluxes of energetic particles can be produced directly during eruptions in the solar atmosphere, or alternatively can result form acceleration of particles in interplanetary space at shock fronts produced by ejections of plasma from the solar surface (see below). Since these particles travel at very high speeds, they too reach the earth very fast (after 30 minutes to a few hours). Although one could consider that some advance warning comes from the observation of the production site, in practice this time is too short to be of much use operationally. Here, too, it is therefore important to develop a deeper understanding of the mechanisms at work in the production processes. A review thereof is presented in the paper by Vainio et al.

Finally, the earth is immersed in a constant flow of plasma from the solar surface known as the *solar wind*. We are fortunately protected from these flows by the earth's magnetic field, which forms a cavity know as the magnetosphere. However, intrinsic variations of the solar wind due to the appearance of 'closed' or 'open' magnetic field lines on the sun (the latter being referred to as *coronal holes*), leading to slow (typically 200–400 km/s) and fast (400–800 km/s) flows of plasma, can cause geomagnetic storms on earth. Even larger perturbations result from the ejection of large clouds of plasma from the solar surface into interplanetary space. Such ejections are called Coronal Mass Ejections (CMEs) and are frequently associated to flaring activity. When they impinge on the earth's magnetosphere, they can cause large geomagnetic storms that lead to lots of different adverse effects on

human technology, such as electric grid failure and loss of GNSS accuracy. Typical speeds of CMEs range from a few hundred to a few thousand km per second. The faster the plasma cloud travels, the larger the energy it carries and hence the more likely it is to cause geomagnetic storms. Also, the faster the CME, the less warning time can be given. Very fast CMEs have been known to reach the earth in less than 20 hours, though typically it takes up to a few days for the plasma to arrive.

As there is much more possibility for an advance warning based on observations (CMEs can be detected in several ways when leaving the sun), a significant effort is dedicated to estimating on an observational basis the properties of CMEs, the most relevant of which is its geo-effectiveness. A review of these techniques is presented in the paper by A. Zhukov. On the other hand, the observational characteristics do not tell the whole picture, and efforts are undertaken to understand the behavior of the CME plasma during its transit of interplanetary space by numerical modeling. As shown in the paper by S. Poedts et al., the numerical modeling can explain certain properties of CMEs, and there is a genuine promise of the ability to incorporate numerical modeling in the daily operations of forecast centres.

Last but not least, it is important to realize that mankind's awareness of the variations of the solar output and of the activity of the sun is really very recent. Although historical records exist of naked-eye sunspot sightings from thousands of years ago, the systematic recordings of sunspot numbers go back only a few hundred years. Its link with space weather effects was realized even more recently. It is important to also consider the relevance of solar activity in a far longer historical perspective and to look for tracers of such activity. Such an outlook is presented in the paper by Messerotti et al.

CHAPTER 1.1

USING CME OBSERVATIONS FOR GEOMAGNETIC STORM FORECASTING

ANDREI N. ZHUKOV

Royal Observatory of Belgium, Avenue Circulaire 3, B–1180 Brussels, Belgium
Skobeltsyn Institute of Nuclear Physics, Moscow State University, 119992 Moscow, Russia

CMEs, THEIR LOW CORONA AND INTERPLANETARY COUNTERPARTS

The phenomenon of coronal mass ejection (CME) plays a key role in solar-terrestrial relations. After the debate on the relative role of CMEs and solar flares it was established that non-recurrent geomagnetic storms (including all the strongest storms) are produced by CMEs (see e.g. Gosling et al. 1990; Kahler 1992; Gosling 1993).

A CME is defined as an observable change in coronal structure that occurs on a time scale between a few minutes and several hours and involves the appearance and outward motion of a new, discrete, bright, white-light feature in the coronagraph field of view (definition by Hundhausen et al. 1984 modified by R. Schwenn). Observational properties of CMEs are summarized in Hundhausen (1999), St Cyr et al. (2000), Yashiro et al. (2004). CMEs often exhibit classical 3-part structure: bright outer shell, dark cavity and eruptive prominence inside; a typical example observed by the Large-Angle Spectroscopic Coronagraph (LASCO, see Brueckner et al. 1995) onboard Solar and Heliospheric Observatory (SOHO) is shown in Fig. 1 (left panel). Most of CMEs, however, exhibit quite different morphologies, including halo CMEs (Fig. 1, right panel). A full halo CME has a shape of a bright irregular ring completely surrounding the coronagraph occulter (Howard et al. 1982), i.e. it is a CME with the angular width of 360°. Full halo CMEs are now interpreted as an end-on view of CMEs propagating approximately along the Sun–Earth line (see e.g. discussion by Plunkett et al. 1998). A partial halo (angular width larger than e.g. 120°) is an intermediate case between limb and full halo CMEs; it also may arrive to the Earth.

J. Lilensten (ed.), Space Weather, 5–13.

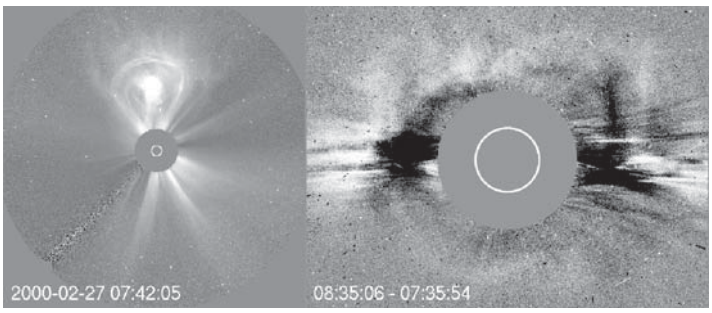


Figure 1. **Left panel:** A CME observed by SOHO/LASCO C3 coronagraph. **Right panel:** Running difference image (a previous image has been subtracted from the next image) showing a full halo CME observed by SOHO/LASCO C2 coronagraph. In both panels the inner white circle shows the solar disc, the larger gray circle represents the coronagraph occulter. All times in this paper are UT

The low coronal counterparts of frontside CMEs can be now routinely observed by Extreme-ultraviolet Imaging Telescope (EIT, see Delaboudinière et al. 1995) onboard SOHO. CME signatures observed by EIT are: coronal dimmings, EIT waves, erupting filaments (prominences), post-eruption arcades and a variety of limb signatures.

Dimmings (including transient coronal holes, TCHs, see Fig. 2, left panel) represent sudden local decreases in brightness (see e.g. Hudson and Cliver 2001 and references therein). Dimmings are the most frequent CME signature in the low corona and appear due to the evacuation of mass during the eruption (Harrison et al. 2003; Zhukov and Auchère 2004). When TCHs occur in pairs (Fig. 2, left panel),

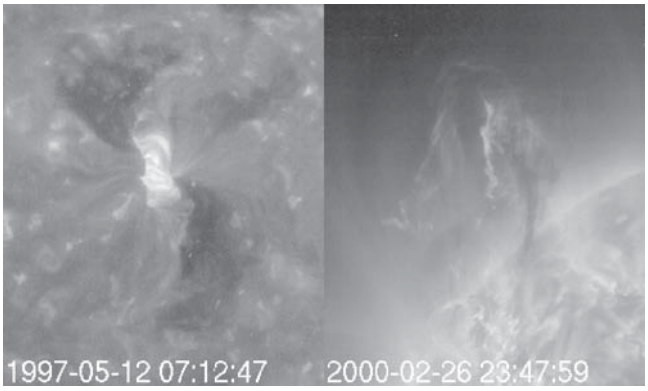


Figure 2. **Left panel:** Coronal dimmings of the transient coronal hole type (seen as two large dark areas) and a post-eruption arcade. **Right panel:** Eruptive prominence. Both images are taken by SOHO/EIT in the Fe XII bandpass (195 Å) showing coronal plasma at temperatures around 1.5 MK

they are usually interpreted as footpoints of the ejected interplanetary flux rope (Webb et al. 2000). Note that TCHs often represent only a part of the overall dimming: as much as 50% of the CME mass may be ejected from outside of them (Zhukov and Auchère 2004). EIT waves (Fig. 3) are bright fronts sometimes propagating from eruption sites (Thompson et al. 1998). They are produced by compression at the front of the fast magnetosonic wave or due to the opening of field lines during the CME lifting (see e.g. discussion by Zhukov and Auchère 2004). Post-eruption arcades (Fig. 2, left panel) naturally occur during the process of CME eruption from a bipolar region (see e.g. Hudson and Cliver 2001 and references therein). An erupting filament is seen as an erupting prominence when observed above the limb (Fig. 2, right panel) and represents the central part of an erupting flux rope. Diverse limb signatures are similar to morphologies of limb CMEs observed by LASCO and are rarely present in Earth-directed CMEs. Any of these features implies that a CME has occurred and often directly indicates the CME source region.

A number of signatures help us to discriminate interplanetary CME counterparts (ICMEs) using in situ observations of the solar wind. Elevated magnetic field, low proton and electron temperatures, low plasma beta, low magnetic field variance, counterstreaming electrons and various composition signatures are typical for ICMEs, see Cane and Richardson (2003) and references therein. Sometimes a smooth magnetic field rotation is observed; such ICMEs are called magnetic clouds (MCs). A fast ICME usually drives a fast forward shock. Between the shock and the ICME a shocked sheath region is situated, characterized by elevated temperatures and densities and rapidly varying magnetic field.

To be geoeffective (in the sense of producing a geomagnetic storm), a solar wind structure should: 1) arrive at the Earth and 2) contain suitable magnetic field orientation. The north–south interplanetary magnetic field (IMF) component Bz should be negative (southward), strong enough and long-lasting (Burton et al. 1975). ICMEs can often produce geomagnetic storms because their strong fields and low field variance may lead to geoeffective Bz configurations. We now discuss our capabilities for predicting CME arrival and geoeffectiveness.

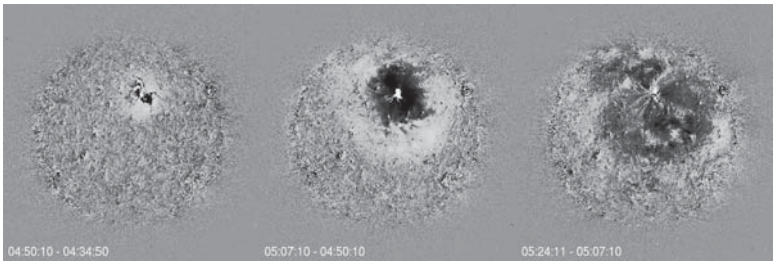


Figure 3. SOHO/EIT running difference images taken in the Fe XII bandpass (195 Å) showing the propagation of an EIT wave on May 12, 1997

PREDICTING ICME ARRIVAL TO THE EARTH

First of all, we should note that predicting a CME before it actually occurs is difficult. It is clear that active regions with a complicated magnetic field structure or high, mature filaments have a significant potential to produce a CME. We, however, still cannot predict the time of eruption well in advance (see e.g. discussion of sigmoidal active regions by Zhukov 2005). So, the first step is to detect a halo CME. Coronagraphic observations, however, cannot distinguish between frontside and backside halos as the occulting disc obscures a direct view of the CME initiation site. A forecaster thus has to look for low coronal CME signatures described above. Taking the halo CME on May 12, 1997 as an example (Fig. 1, right panel), one can determine that this is a frontside halo as it is associated with coronal dimmings, a post-eruption arcade (Fig. 2, left panel), an EIT wave (Fig. 3) and an eruptive prominence observed in Hα line (Thompson et al. 1998; Webb et al. 2000).

The next step is to determine the direction of eruption. It is safe to suggest that nearly symmetric frontside full halo CMEs (Fig. 1, right panel) are directed towards the Earth. It was, however, reported (Schwenn et al. 2005) that around 7% of all frontside full halos missed the Earth. It is clear that a forecaster also has to take into account the halo CME source region position on the solar disc. For the Earth-directed CMEs there is an obvious concentration of source regions near the disc center (Cane et al. 2000; Wang et al. 2002; Zhang et al. 2003; Manoharan et al. 2004; Srivastava and Venkatakrishnan 2004).

Source regions located farther from the disc center may be associated with partial halos. In most of such cases only an interplanetary shock is observed (e.g. Manoharan et al. 2004). The CME misses the Earth, but the angular extent of the shock is larger than that of the corresponding CME (Bothmer and Schwenn 1998). The storm can then be produced by strong southward fields in the sheath. The farther the source region is from the disc center, the larger the probability to encounter only a shock or no CME-associated structure at all. It was noted (Schwenn et al. 2005) that one in four frontside partial halo CMEs did not hit the Earth.

An asymmetry in the source region distribution has been reported by Wang et al. (2002) and Zhang et al. (2003): geoeffective CMEs have a slight preference to originate from the western hemisphere. This finding is still controversial (Cane et al. 2000; Srivastava and Venkatakrishnan 2004). An explanation of this asymmetry has been proposed by Wang et al. (2004): CMEs that are faster than the ambient solar wind are deflected to the east by the magnetic force of the ambient spiral IMF. A dynamic model of this interaction is still to be developed, and a statistical study including weaker events is needed to verify whether the longitudinal asymmetry indeed exists.

It was suggested by Zhang et al. (2003) that four major storms were produced by slow east-limb partial halo CMEs without any signatures on the solar disc. It is possible that EUV dimmings in slow CMEs are continuously replenished with plasma and thus are not pronounced. Alternative sources for these storms proposed by Zhukov (2005) involve an eruption observed by EIT close to the disc center, but without a corresponding CME detected by LASCO. This may be due to insufficient LASCO sensitivity. The Thomson scattering process is most efficient close to the

plane of the sky, so some Earth-directed events – which naturally have a lot of material out of the plane of the sky – may be missed by LASCO. If this interpretation is correct, east-limb partial halos found by Zhang et al. (2003) could be classified as backside CMEs.

It was reported that some near-Earth ICMEs either have no corresponding CMEs (Cane and Richardson 2003) or a non-halo CME is involved (Schwenn et al. 2005). Again, this may be due to insufficient LASCO sensitivity. However, in some of these cases CME signatures in the EIT data are easy to find. Then a more attentive inspection of the LASCO data may lead to the identification of the corresponding halo CME not detected earlier (Zhukov 2005). It seems that LASCO sensitivity allows us to detect even very weak CMEs. A careful statistical study is needed to verify whether ICMEs without corresponding LASCO CMEs indeed occurred.

Once an Earth-directed CME is observed, the time of its arrival at the Earth has to be estimated. Depending on the speed, CME travel times are typically in the range of 1–5 days. The shortest interval between the CME eruption and its arrival at the Earth is around 19 hours (halo CME on October 28, 2003). A difficulty is that the CME speed measured by LASCO is the plane of the sky projected speed. It might be different from the true propagation speed, especially for halo CMEs. So, empirical studies of the relation between the travel time and projected speed have been performed, see Gopalswamy et al. (2001), Cane and Richardson (2003) and

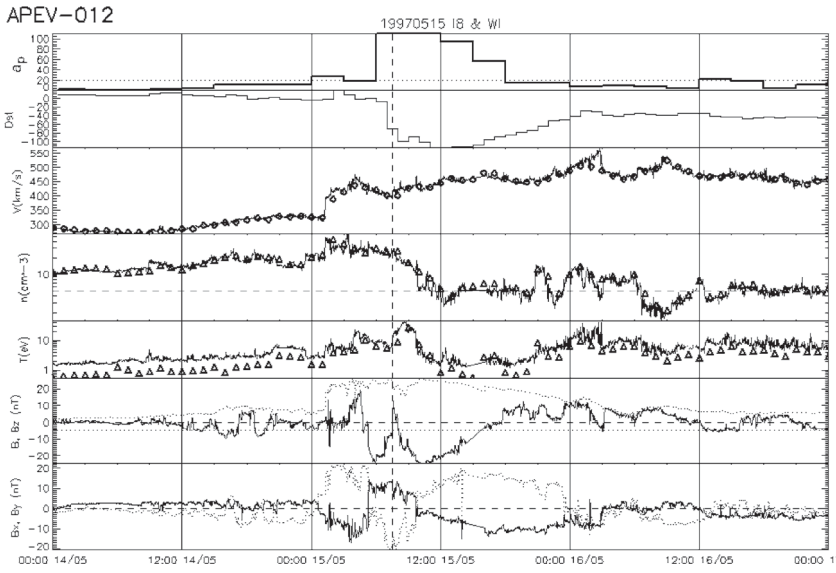


Figure 4. Solar wind (WIND spacecraft) and geomagnetic data for the storm on May 15, 1997. From top to bottom: a_p index; Dst index; solar wind speed; proton number density; proton temperature; IMF magnitude (dotted line) and its *Bz* component (solid line); IMF *By* (dotted line) and *Bx* (solid line) components. The plot is taken from the APEV database (http://dbserv.sinp.msu.ru/apev, courtesy A. V. Dmitriev)

references therein. A more advanced method proposed by Schwenn et al. (2005) takes into account the relation between the CME expansion speed (that can be measured for all CMEs) and the radial speed. Acceleration or deceleration of CMEs resulting from the interaction with ambient solar wind flow or other CMEs can hardly be taken into account now.

For our example halo CME on May 12, 1997 the projected speed of 250 km/s was measured and the estimated true speed was around 600 km/s (Plunkett et al. 1998), suggesting the arrival on May 15. Indeed, a shock was registered by WIND spacecraft at 01:15 UT on May 15 followed by the magnetic cloud (Fig. 4; see also Webb et al. 2000). A geomagnetic storm with peak Dst $= -115$ nT was produced by the southward IMF in the leading part of the cloud.

PREDICTING THE IMF DIRECTION IN ICMEs

To assess the strength of a possible storm produced by an ICME, the IMF Bz component has to be predicted. There is no reliable prediction method yet, but some useful indications can be obtained from solar observations. If the photospheric magnetic field of the CME source region has a bipolar configuration (as is often the case), one can determine the orientation of the neutral line and – assuming the flux rope geometry – the direction of the flux rope azimuthal field. If the shear of the magnetic field can be determined, e.g. from the hemispheric chirality rule (see Bothmer and Schwenn 1998), the direction of the magnetic field in the erupting flux rope can be reasonably estimated. The orientation of the neutral line and the flux rope chirality were found to correspond respectively to the axis inclination and chirality of the resulting MC (Marubashi 1997; Bothmer and Schwenn 1998; Zhao and Hoeksema 1998; McAllister et al. 2001; Yurchyshyn et al. 2001). So, when the source region neutral line is oriented along the north–south direction and the axial field in the flux rope is northward (e.g. halo CME on February 17, 2000, Yurchyshyn et al. 2001), the corresponding MC will produce only a very weak geomagnetic disturbance. On the contrary, if a halo CME originates from an active region with the neutral line oriented along the east–west direction (e.g. July 14, 2000), southward Bz will be encountered either in the leading or in the trailing part of the flux rope and the MC will probably be geoeffective.

One can hardly determine which part of the complex three-dimensional flux rope loop structure (e.g. a leg or the apex) will be encountered by the Earth (McAllister et al. 2001). Additionally, it seems difficult to apply the method described above to ICMEs that cannot be fit by a simple flux rope model. The inclinations of flux rope axes close to the Sun and in the heliosphere do not always correspond to each other. The neutral line in our example case of May 12, 1997 was stretched in the north–south direction, see the post-eruption arcade configuration in Fig. 2, left panel, and the magnetogram in Fig. 3 of Webb et al. (2000). The erupting flux rope thus had to have the ENW orientation (see Bothmer and Schwenn 1998; Mulligan et al. 1998 for definition), i.e. to be not geoeffective. However, Fig. 4 shows that the interplanetary flux rope was of the SEN type (magnetic field was

rotating from south to north with axial field pointing towards the east) and produced a major geomagnetic storm. It was suggested (Cremades and Bothmer 2004) that during the years of low solar activity CMEs are deflected by the fast flows from polar coronal holes and thus have the tendency to have a small inclination with respect to the ecliptic plane; then one should observe mostly SN and NS clouds during low activity epoch. Statistical studies of the MC inclination give somewhat contradictory results (Mulligan et al. 1998; Huttunen et al. 2005), but it seems clear that highly inclined clouds may occur during the low activity years. Another explanation (Crooker 2000) suggests that interplanetary flux ropes result from the large-scale dipole field rather than from the local bipolar field of the source region. In this case, however, it is not easy to imagine how such a mechanism can produce a highly inclined MC during the low activity epoch.

CONCLUSIONS

In most of the cases, EIT and LASCO are capable of identifying the eruption of an Earth-directed CME, and provide us with a good estimate of the arrival time. Some indications of the resulting ICME's magnetic configuration can be obtained. However, predicting CMEs before they actually occur is still a challenge. The IMF *Bz* profile is difficult to predict. Additionally, false alarms (e.g. from partial halo CMEs) or missed events (when solar signatures are inconclusive) can occur. Forecasting may also be difficult if the geometry of a CME is not clear or in complicated cases of multiple (interacting) CMEs (Burlaga et al. 2002) – a problem not discussed in this paper. To monitor and forecast ICME propagation more precisely, CMEs should be tracked from a vantage point out of the Sun–Earth line; this will be provided by the STEREO mission (Solar–TErrestrial RElations Observatory).

ACKNOWLEDGEMENTS

LASCO and EIT data have been used courtesy of SOHO/LASCO and SOHO/EIT consortia. SOHO is a project of international cooperation between ESA and NASA. WIND SWE and MFI instrument teams are acknowledged for providing solar wind data, as well as the World Data Center for Geomagnetism (Kyoto, Japan) for the a_p and Dst indices. Author is grateful to G. Lawrence for the help in preparation of the manuscript and acknowledges support from the Belgian Federal Science Policy Office through the ESA-PRODEX programme.

REFERENCES

Bothmer, V., Schwenn, R.: The structure and origin of magnetic clouds in the solar wind. Ann Geophys, 16, 1–24 (1998)

Brueckner, G.E., Howard, R.A., Koomen, M.J., Korendyke, C.M., Michels, D.J., Moses, J.D., Socker, D.G., Dere, K.P., Lamy, P.L., Llebaria, A., Bout, M.V., Schwenn, R., Simnett, G.M., Bedford, D.K., Eyles, C.J.: The Large-Angle Spectroscopic Coronagraph (LASCO). Sol Phys, 162, 357–402 (1995)

Burlaga, L.F., Plunkett, S.P., St Cyr, O.C.: Successive CMEs and complex ejecta. J Geophys Res, 107 (A10), SSH 1–1–1–12 (2002)

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

Cane, H.V., Richardson, I.G.: Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. J Geophys Res, 108(A4), SSH 6–1–6–13 (2003)

Cane, H.V., Richardson, I.G., St Cyr, O.C.: Coronal mass ejections, interplanetary ejecta and geomagnetic storms. Geophys Res Lett, 27, 3591–3594 (2000)

Cremades, H., Bothmer, V.: On the three-dimensional configuration of coronal mass ejections. Astron Astrophys, 422, 307–322 (2004)

Crooker, N.U.: Solar and heliospheric geoeffective disturbances. J Atm Sol-Terr Phys, 62, 1071–1085 (2000)

Delaboudinière, J.-P., Artzner, G.E., Brunaud, J., Gabriel, A.H., Hochedez, J.F., Millier, F., Song, X.Y., Au, B., Dere, K.P., Howard, R.A., Kreplin, R., Michels, D.J., Moses, J.D., Defise, J.M., Jamar, C., Rochus, P., Chauvineau, J.P., Marioge, J.P., Catura, R.C., Lemen, J.R., Shing, L., Stern, R.A., Gurman, J.B., Neupert, W.M., Maucherat, A., Clette, F., Cugnon, P., van Dessel, E.L.: EIT: Extreme-ultraviolet Imaging Telescope for the SOHO mission. Sol Phys, 162, 291–312 (1995)

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

Gosling, J.T.: The solar flare myth. J Geophys Res, 98, 18937–18949 (1993)

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

Harrison, R.A., Bryans, P., Simnett, G.M., Lyons, M.: Coronal dimming and the coronal mass ejection onset. Astron Astrophys, 400, 1071–1083 (2003)

Howard, R.A., Michels, D.J., Sheeley, Jr., N.R., Koomen, M.J.: The observation of a coronal transient directed at Earth. Astrophys J, 263, L101–L104 (1982)

Hudson, H.S., Cliver, E.W.: Observing coronal mass ejections without coronagraphs. J Geophys Res, 106, 25199–25214 (2001)

Hundhausen, A.J., Sawyer, C.B., House, L., Illing, R.M.E., Wagner, W.J.: Coronal mass ejections observed during the Solar Maximum Mission: Latitude distribution and rate of occurrence. J Geophys Res, 89, 2639–2646 (1984)

Hundhausen, A.J.: Coronal Mass Ejections. In: Strong, K.T., Saba, J.L.R., Haisch, B.M., Schmelz, J.T. (eds) The many faces of the Sun: a summary of the results from NASA's Solar Maximum Mission. New York, Springer, pp 143–200 (1999)

Huttunen, K.E.J., Schwenn, R., Bothmer, V., Koskinen, H.E.J.: Properties and geoeffectiveness of magnetic clouds in the rising, maximum and early declining phases of solar cycle 23. Ann Geophys, 23, 625–641 (2005)

Kahler, S.W.: Solar flares and coronal mass ejections. Ann Rev Astron Astrophys, 30, 113–141 (1992)

Manoharan, P.K., Gopalswamy, N., Yashiro, S., Lara, A., Michalek, G., Howard, R.A.: Influence of coronal mass ejection interaction on propagation of interplanetary shocks. J Geophys Res, 109, A06109 (2004)

Marubashi, K.: Interplanetary magnetic flux ropes and solar filaments. In: Crooker, N., Joselyn, J.A., Feynman, J. (eds) Coronal Mass Ejections. AGU Geophys Monogr Ser, vol 99, pp 147–156 (1997)

McAllister, A.H., Martin, S.F., Crooker, N.U., Lepping, R.P., Fitzenreiter, R.J.: A test of real-time prediction of magnetic cloud topology and geomagnetic storm occurrence from solar signatures. J Geophys Res, 106, 29185–29194 (2001)

Mulligan, T., Russell, C.T., Luhmann, J.G.: Solar cycle evolution of the structure of magnetic clouds in the inner heliosphere. Geophys Res Lett, 25, 2959–2962 (1998)

Plunkett, S.P., Thompson, B.J., Howard, R.A., Michels, D.J., St Cyr, O.C., Tappin, S.J., Schwenn, R., Lamy, P.L.: LASCO observations of an Earth-directed coronal mass ejection on May 12, 1997. Geophys Res Lett, 25, 2477–2480 (1998)

Schwenn, R., Dal Lago, A., Huttunen, E., Gonzalez, W.D.: The association of coronal mass ejections with their effects near the Earth. Ann Geophys 23, 1033–1059 (2005)

Srivastava, N., Venkatakrishnan, P.: Solar and interplanetary sources of major geomagnetic storms during 1996–2002. J Geophys Res, 109, A10103 (2004)

St Cyr, O.C., Howard, R.A., Sheeley, N.R., Plunkett, S.P., Michels, D.J., Paswaters, S.E., Koomen, M.J., Simnett, G.M., Thompson, B.J., Gurman, J.B., Schwenn, R., Webb, D.F., Hildner, E., Lamy, P.L.: Properties of coronal mass ejections: SOHO LASCO observations from January 1996 to June 1998. J Geophys Res, 105, 18169–18185 (2000)

Thompson, B.J., Plunkett, S.P., Gurman, J.B., Newmark, J.S., St Cyr, O.C., Michels, D.J.: SOHO/EIT observations of an Earth-directed coronal mass ejection on May 12, 1997. Geophys Res Lett, 25, 2465–2468 (1998)

Wang, Y.M., Ye, P.Z., Wang, S., Zhou, G.P., Wang, J.X.: A statistical study on the geoeffectiveness of Earth-directed coronal mass ejections from March 1997 to December 2000. J Geophys Res, 107(A11), SSH 2–1 – SSH 2–9 (2002)

Wang, Y., Shen, C., Wang, S., Ye, P.: Deflection of coronal mass ejection in the interplanetary medium. Sol Phys, 222, 329–343 (2004)

Webb, D.F., Lepping, R.P., Burlaga, L.F., DeForest, C.E., Larson, D.E., Martin, S.F., Plunkett, S.P., Rust, D.M.: The origin and development of the May 1997 magnetic cloud. J Geophys Res, 105, 27251–27260 (2000)

Yashiro, S., Gopalswamy, N., Michalek, G., St Cyr, O.C., Plunkett, S.P., Rich, N.B., Howard, R.A.: A catalog of white light coronal mass ejections observed by the SOHO spacecraft. J Geophys Res, 109, A07105 (2004)

Yurchyshyn, V.B., Wang, H., Goode, P.R., Deng, Y.: Orientation of the magnetic fields in interplanetary flux ropes and solar filaments. Astrophys J, 563, 381–388 (2001)

Zhang, J., Dere, K.P., Howard, R.A., Bothmer, V.: IDENTIFICATION of Solar Sources of Major Geomagnetic Storms between 1996 and 2000. Astrophys J, 582, 520–533 (2003)

Zhao, X.P., Hoeksema, J.T.: Central axial field direction in magnetic clouds and its relation to southward interplanetary magnetic field events and dependence on disappearing solar filaments. J Geophys Res, 103, 2077–2083 (1998)

Zhukov, A.N.: Solar sources of geoeffective CMEs: a SOHO/EIT view. In: Dere, K., Wang, J., Yan, Y. (eds) Coronal and stellar mass ejections, Proceedings IAU Symposium No. 226, Cambridge, Cambridge University Press, pp 437–447 (2005)

Zhukov, A., Auchère, F.: On the nature of EIT waves, EUV dimmings and their link to CMEs. Astron Astrophys, 427, 705–716 (2004)

CHAPTER 1.2

SOLAR ACTIVITY MONITORING

PETER T. GALLAGHER[1,3], R. T. JAMES MCATEER[2], C. ALEX YOUNG[3], JACK IRELAND[3], RUSSELL J. HEWETT[3,4] AND PAUL CONLON[5,1,3]

[1] *School of Physics, Trinity College Dublin, Dublin 2, Ireland*
[2] *National Research Council, Laboratory for Solar and Space Physics, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[3] *L-3 Communications GSI, Laboratory for Solar and Space Physics, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[4] *Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[5] *School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland*

INTRODUCTION

Monitoring the ever-changing solar atmosphere is not only of interest to solar researchers, but has important practical benefits to the space weather community. The primary causes of adverse space weather conditions are solar flares (e.g., Gallagher et al. 2002) and Coronal Mass Ejections (CMEs; e.g., Gallagher et al. 2003), which for the most part result from energy released in the complex magnetic fields of sunspot groups or active regions. In addition, the Sun produces a continuous stream of plasma, known as the solar wind, which extends into the outer solar system. Solar wind speeds of greater than $\sim 700\,\mathrm{kms}^{-1}$ are of particular importance to space weather, and are known to emanate from solar features called coronal holes.

Systems such as Computer Aided CME Tracking (CACTus; Berghmans et al. 2002; Robbrecht and Berghmans 2004) and *SolarMonitor* (formerly known as the *Active Region Monitor*; Gallagher et al. 2002), represents initial steps towards providing space weather-relevant solar data in a timely fashion. The Virtual Solar Observatory (VSO; Hill et al. 2004) and the European Grid of Solar Observations (EGSO; Bentley et al. 2004), are two related initiatives. While VSO and EGSO address many of the problems of solar data storage and access, neither confront

J. Lilensten (ed.), Space Weather, 15–26.

issues relating to near-realtime processing of heterogenous data, nor do they provide operational data products of use to space weather forecasters.

Here we review the current status of autonomous solar monitoring techniques of relevance to space weather. The application of these techniques to up-coming missions in then discussed in the final section.

SOLAR ACTIVITY MONITORING

The techniques used to monitor solar activity depend on many factors, such as time- or size-scale, or region of interest. In the photosphere, sunspots are visible as well-defined, dark features, which in many cases can be identified and extracted using standard image processing techniques, such as intensity thresholding. CMEs, on the other hand, require more advanced methods, primarily due to their high speed, diffuse structure and complex morphology. The following sections therefore consider the methods employed to analyse and monitor the distinct structures visible in the photosphere, chromosphere and corona.

Photosphere

During 1917–2005, synoptic sunspot drawings were created and archived at the 150-foot Solar Tower at Mount Wilson, California. These included sunspot location, morphology and magnetic classification.

The National Oceanic and Atmospheric Administration (NOAA) in conjunction with the US Air Force (USAF) performs a similar task, by providing a daily update of sunspot locations, area, Zurich classification, longitudinal extent, total number of group sunspots and magnetic classification. These Solar Region Summary (SRS) data are updated at approximately 00:30 UT every day. Several observatories, such as Big Bear Solar Observatory, obtain regular white light or continuum images. These images show sunspot groups, but can be limited by atmospheric effects, duty cycle and are not always available online. The Michelson Doppler Imager (MDI; Scherrer et al. 1995) on SOHO, on the other hand, provides several "continuum" images per day. These data are indispensable for autonomous monitoring of sunspot group evolution.

On a half-hourly basis, *SolarMonitor* locates the most recent NOAA SRS and extracts the active region positions. The positions are then differentially rotated to the time of the most recent MDI images, and used to automatically extract zoomed views of each active region. These zoomed-in views are then posted on *SolarMonitor.org* as in Fig. 1.

Only recently has progress been made in combining real-time data with autonomous data-analysis techniques. Here we describe two complimentary techniques, namely multiscalar and multifractal methods, which show some promise for near-realtime solar monitoring applications. Both methods are based on the premise that image features cannot be adequately described by simple parameters, but require methods that capture their true multiscale structure. Furthermore, the

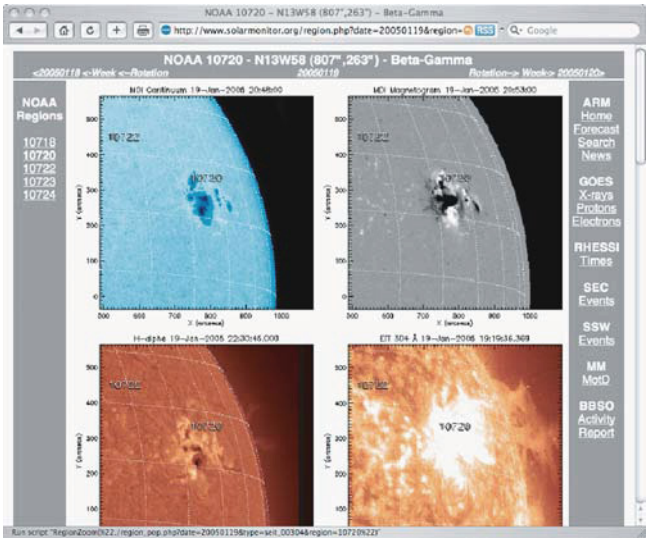


Figure 1. NOAA 10720 as seen at *SolarMonitor.org*. The top row were extracted from SOHO/MDI. The top-left shows a continuum image, while the top-right shows the longitudinal magnetic field. The bottom panels show corresponding images in H-alpha and SOHO/EIT (30.4 nm)

multiscalar and multifractal nature of the solar surface and atmosphere can be directly related to the mathematical predictions of turbulence theory (e.g., Lawrence et al. 1993).

Multifractal measures

Since Mandelbrot (1977) first introduced the fractal dimension, the idea of quantitatively describing the complexity of a system has been applied to many areas. Recently, there have been tantalizing glimpses in the literature concerning the diagnostic possibilities of fractal and multifractal analysis to solar images (e.g., Abramenko 2005), primarily due to the fact that the flows in the solar photosphere is highly turbulent, and is therefore expected to be fractal (Georgoulis 2005). This motivated McAteer et al. (2005) to address the long-standing problem of quantifying active region magnetic complexity using fractals. They found that larger, more complex regions with large fractal dimensions (see Fig. 2) tended to produce larger flares more often. They also showed that the fractal approach does not provide an unambiguous description of active region complexity. A more rigorous, multifractal approach may therefore be more appropriate.

The fractal dimension of any object can be thought of as the self-similarity of an image across all scale sizes, or the scaling index of any length to area measure, $A \sim l^{\alpha}$, where α is the Hölder exponent. However, a multifractal system will contain a spectrum of fractal indices of different powers, $A \sim l^{f(\alpha)}$, and takes account of the measure at each point in space. The three main multifractal indices commonly

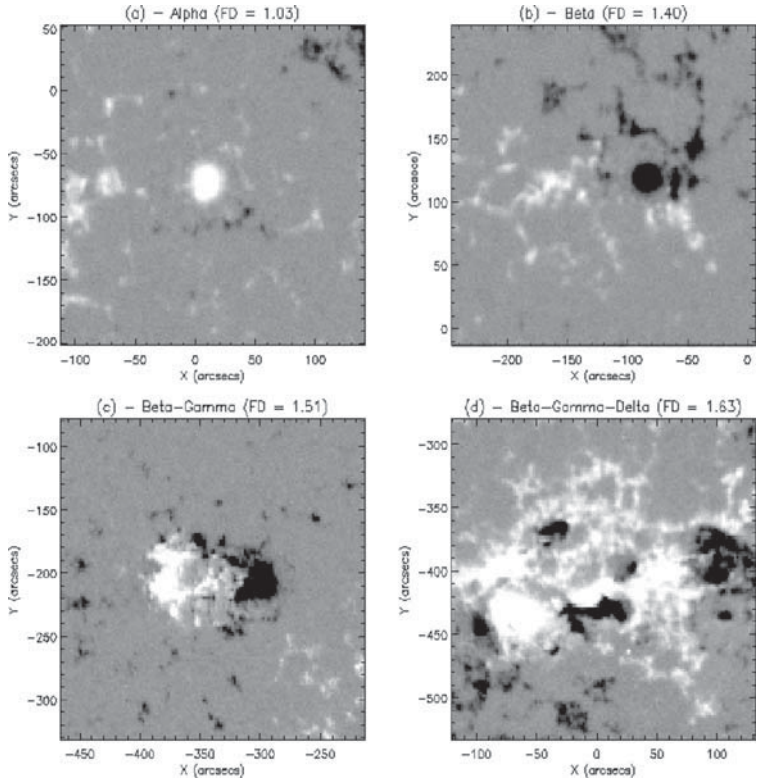


Figure 2. MDI magnetograms for four Mount Wilson classifications and their fractal dimensions. Panel (a) gives a simple α spot, panel (b) a bi-polar β, while panels (c) and (d) show the more complex $\beta\gamma$ and $\beta\gamma\delta$ classifications

used to represent a non-uniform measure are: Generalised correlation dimensions, $D_q = \tau/(q-1)$; Holder exponent, $\alpha = \mathrm{d}\tau/\mathrm{d}q$; Multifractal spectrum, $f(\alpha) = q\alpha - \tau$.

Figure 3 shows a comparison of the D_q and $f(\alpha)$ indices for a monofractal, a multifractal, and an active region. The monofractal exhibits a flat $D_q(\approx 1.89)$ spectrum and narrow $f(\alpha)$ spectrum whilst the multifractal exhibits a monotonically decreasing D_q spectrum (for increasing q) and wide $f(\alpha)$ spectrum. The magnetogram exhibits a similar multifractal behaviour. The degree of multifractality of an image can be measured as the drop of D_q or width of the $f(\alpha)$ spectrum.

Multiscale measures

The range of scales in fully developed turbulence was predicted by Kolmogorov (1941) to scale as $E(k) \sim k^{-5/3}$. While Fourier methods have been used extensively in image processing and fluid flow analysis (Abramenko 2005), it does not provide a complete spatially localised diagnostic. This limitation is particularly important in 2D astrophysical flows, where small-scale power may be highly intermittent.

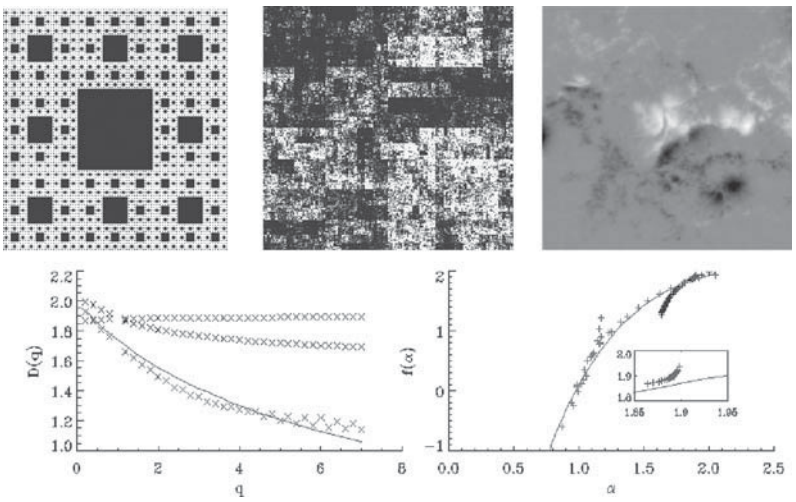


Figure 3. Top: A monofractal (top left), multifractal (top middle) and active region (top right). Bottom: Their corresponding D_q (bottom left) spectra and $f(\alpha)$ (bottom right) spectra

As photospheric flows are in a state of fully developed turbulence, multiscale methods are well suited to measuring and understanding the complex structures observed on the solar surface. This is particularly true for conditions in active regions. The wavelet transform is localised in space and hence allows the detection of local image features. This is essential as small regions of flux emergence/submergence are vital in detailing the evolution of the surface sources of space weather. Further examples of the importance of multiscale image processing include neutral line identification, and automated extraction of features from Hα and EUV images. Figure 4 shows the wavelet analysis decomposition of the active region in Fig. 3, along with a comparison of the global wavelet analysis and traditional Fourier approach. Wavelet analysis retains the localised spatial information, providing vital information on the turbulent flow. As such, wavelets provide a complement to multifractal analysis.

Chromosphere

The chromosphere has traditionally been imaged in the Hα or Ca II H and K absorption lines at ground-based observatories. The Global High Resolution Hα Network (GHN; *http://www.bbso.njit.edu/Research/Halpha/*) was established to provide near-continuous observations of the solar chromosphere, by coordinating facilities at Big Bear Solar Observatory (California), Kanzelhöhe Solar Observatory (Austria), Catania Astrophysical Observatory (Italy), Meudon Observatory (France) and Huairou Solar Observing Station and Yunnan Astronomical Observatory (China). Each station has a $1\,\mathrm{K} \times 1\,\mathrm{K}$ or $2\,\mathrm{K} \times 2\,\mathrm{K}$ CCD, with a spatial

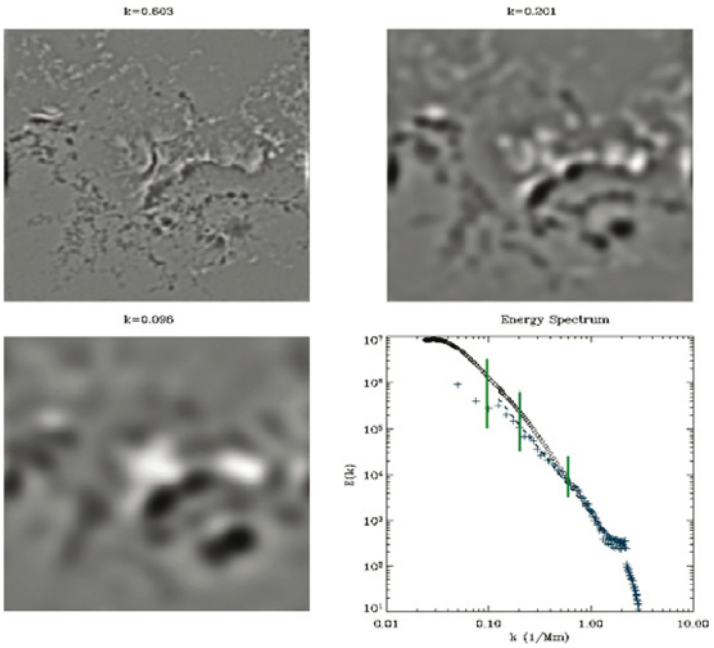


Figure 4. The wavelet decomposition of the magnetogram in Fig. 3 at three different scales. The energy spectrum (lower right) of the wavelet decomposition (diamonds) agrees well with the Fourier decomposition (crosses). The vertical lines show the wavenumber of each scale displayed

resolution of 1 arcsec per pixel. 1-minute cadence observations are obtained at each station, with higher cadence available for periods of increased activity. This enables the network to monitor flares, filament lift-offs, Morton waves, and a variety of other chromospheric phenomena.

Filaments

For the purposes of space weather, filaments are the most important feature to monitor in the chromosphere. Although filament eruptions can result in CMEs, their location and properties are not currently monitored in a near-realtime or systematic manner. That said, several groups have developed techniques to identify and extract filaments from Hα images.

Schuck et al. (2004) examined the pixel intensity statistics of full-disk Hα images, and demonstrated that dynamical changes in filaments are detectable prior to the occurrence of a solar flare (see Fig. 5). The method, which is based on the intensity distribution of each image, is fast, simple, and based on statistical measures that require no a priori knowledge of the location of the filaments. Shih and Kowalski (2003) also used well-established techniques to identify and extract filaments. Their technique employed two alternative preprocessing techniques to convert grayscale

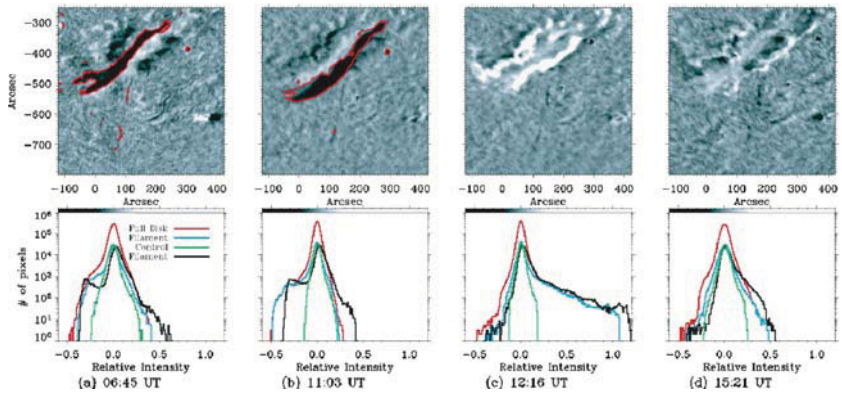


Figure 5. Top: Sequence of time-averaged H-alpha exposures of a filament region. *Bottom*: Corresponding intensity histograms (Schuck et al. 2004)

images into binary images: local thresholding based on median values and global thresholding with brightness and area normalization.

They then used morphological closing operations with multi-directional linear structuring elements to extract elongated shapes (i.e., filaments) in the image. They achieved excellent results for large filaments and moderate success for smaller filaments. Zharkova et al. (2005) and Fuller et al. (2005) reported a number of automated pattern recognition techniques for the EGSO Solar Feature Catalogues. Here, Hα images were first enhanced using a sharpening filter a filaments "seeds" then identified in order to initiate the region growing process. The next step involved growing the seeds according to the statistics of the local pixels. Once the images are segmented, the filament length, center and centre of curvature were determined from the pruned skeleton, which was obtained using morphological operators These methods could be of benefit to monitoring filament activity and predicting filament eruptions.

Corona

Active regions

Solar active regions were identified in *Skylab* and early rocket data as concentrations of hot (>2MK) EUV and soft X-ray emitting loops that extend to altitudes of up to $\sim$ 20 Mm above the surface (e.g., Gallagher et al. 2001). Since *Yohkoh* and SOHO in particular, the term "active region" is now widely used to refer to the entire volume of plasma and magnetic fields that extends from below sunspot concentrations in the photosphere, to the high corona. Here, we focus on the coronal component of active regions, visible by EUV/X-ray imaging instruments such as the Extreme ultraviolet Imaging Telescope (EIT; Delaboudiniere et al. 1995) on SOHO and Solar X-ray Imager (SXI; Hill et al. 2005; Pizzo et al. 2005).

The evolution and activity of active regions are monitored by eye by forecasters at NOAA/SEC, mainly using data from SXI. The sole autonomous system to monitor

active regions is the *SolarMonitor*, which is hosted at NASA Goddard Space Flight Center. *SolarMonitor* reads in the most recent EUV images from SOHO/EIT (30.4, 17.1, 19.5 and 28.4 nm) and locates and extracts active regions based on coordinated supplied by NOAA/SEC. The system also reads in SXI images and again extracts each region. These data-products are posted at *SolarMonitor.org*.

Coronal holes

X-ray images from *Skylab* showed clear evidence of X-ray coronal holes. These regions of low emission can persist for months, appearing for approximately 2 weeks during each solar rotation. Coronal holes, particularly during solar activity minimum, are associated with high-speed solar wind streams and recurring geomagnetic disturbances, making them extremely useful for predicting recurrent geomagnetic disturbances. For many years, forecasters have relied on images taken in the infrared He 1083 nm line observed at the Kitt Peak Observatory. Coronal holes are not as striking in these images, and a coronal hole "expert" determines the location of the holes and sends this information to forecasters at NOAA/SEC. In the 1990's NOAA/SEC began using images from *Yohkoh*/SXT, SOHO/EIT and most recently from SXI. Coronal holes are straightforward to identify from instruments sensitive to high coronal temperatures. This prompted the SXI team to develop coronal hole identification algorithms to identify and track coronal holes in a near-realtime basis.

Schrijver (2003) developed a potential-field source-surface model that enables coronal holes (amongst other things) to be identified and monitored in near-realtime (*http://www.lmsal.com/forecast/*). The model gives the trajectories of magnetic field-lines through a model solar corona that spans the spherical volume between the solar photosphere and 2.5 R_{sun}. The coronal field model used for these models is the potential field source surface (PFSS) model, which has the characteristics that the model field is potential. When eruptive events occur and particles are measured by satellites such as ACE, an idea of the photospheric source regions of these particles can be obtained using such maps.

Flares

The most complete listing of solar flares is maintained by NOAA/SEC. This list, which is updated every 30-minutes, contains information on the event start, peak and end, the type of event, the X-ray class (A–X class), and the region number associated with the event. They currently report fifteen different types of events, including bright surge on the limb, filament disappearances, eruptive prominence on the limb, optical flares observed in H-alpha, loop prominence system, sprays, X-ray flare from GOES Solar X-ray Imager (SXI) and X-ray events from GOES 8–12. For SXI flares, an SEC algorithm finds the brightest area in the SXI image and assigns the region number of the closest active solar region.

In terms of autonomous flare monitoring, the *SolarSoft* Latest Events Archive at *http://www.lmsal.com/solarsoft/last_events/* is probably the most reliable source for flare locations. Flares are identified via a straightforward differencing technique using EIT or SXI images. The flare location is then associated with the nearest

NOAA active region, and its position and magnitude archived. *SolarMonitor* also ingests the *SolarSoft* flare lists and presents them in tabular form.

Coronal Mass Ejections

CMEs have only been monitored in near-realtime since the launch of SOHO in 1995. During this period, CME monitoring has been heavily reliant on the visual inspection of LASCO images (e.g., LASCO CME List at *http://lasco-www.nrl.navy.mil/cmelist.html*). Seiji Yashiro and Nat Gopalswamy of NASA Goddard Space Flight Centre improved on this, by performing a detailed post-event analysis of a massive sample of CMEs observed since January 1996 (*http://cdaw.gsfc.nasa.gov/CME_list/)*. Unfortunately this is a labour-intensive task, and so there is a significant delay between data being received and analyzed. This methodology is therefore unsuitable for realtime applications, even though many of its data-products would certainly be of benefit.

CMEs and CME leading edges are difficult objects to detect and characterise using standard image processing techniques, due to their diffuse structure, ill-defined edges, and high velocities (up to $\sim 3000\,\mathrm{km\ s^{-1}}$). Stenborg and Cobelli (2003) were the first to apply a wavelet-based technique to study the multi-scale nature of coronal structures LASCO images. Their method employed as multi-level decomposition scheme via the isotropic a trous wavelet transform. Unfortunately, this implementation computes the transform at every integer scale (1, 2, 3,..), and is therefore computationally slow. This is a particular draw-back for real-time applications. Furthermore, this method is only a feature enhancement method not a feature detection method. Recently, Young and Gallagher (2006) used wavelet based techniques to analyse the structure and properties of CMEs in the low corona. Figure 6 shows a series of coronal images taken by the LASCO (white-light coronagraph with a field-of-view of 3–7 R_{sun}) onboard SoHO. Although the dynamic range of the instrument is large, there is a large, varying background that makes faint objects difficult to identify. Applying the multiscale methods of Young and

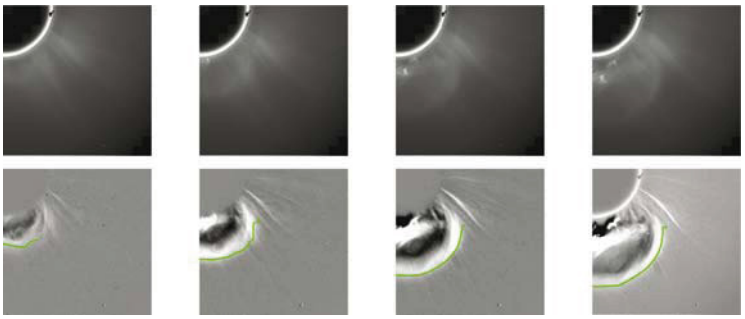


Figure 6. Top row: A sequence of unprocessed LASCO images with time moving from left to right. Bottom row: The previous sequence with the multiscale edges identified at scale 16

Gallagher (2006), the faint eruption can be detected and its edges identified without the need for background subtraction. The top row of images contains four sequential LASCO images in which a faint eruption can be seen propagating away from the Sun. The bottom row shows the same set of difference images overlaid with the scale 16 multiscale edges. These edges define edges of maximum intensity along the front of the eruption.

CACTus (*http://sidc.oma.be/cactus/*), developed at the Royal Observatory of Belgium, autonomously detects CMEs in LASCO images. The system provides principle angle, angular width and velocity estimation for each CME detected. CMEs can be seen in height-time plots as inclined ridges with the inclination angle corresponding to the propagation speed. The ridges are then detected with the Hough transform. By combining the ridges in height-time plots from all directions, the CME front can be reconstructed as it propagates outwards. The main drawback of this method is that the Hough transform imposes a linear height-time evolution, therefore forcing constant velocity profiles for each bright feature.

FUTURE PROSPECTS

Many of the recent advances in solar activity monitoring have relied on the development of new and innovative instruments and data-analysis methods. SOHO in particular has revolutionised our understanding of solar activity. Its suite of 12 instruments have enabled scientists to monitor solar activity from below the surface, through the solar corona, to the solar wind.

These capabilities will be improved upon with soon to be launched satellites, such as the Solar Dynamics Observatory (SDO) and the Solar Terrestrial Relations Observatory (STEREO). STEREO, to be launched in late 2006, will employ two nearly identical space-based observatories – one ahead of Earth in its orbit, the other trailing behind – to provide stereoscopic measurements of solar features, such as active regions and CMEs. SDO, to be launched in 2008, will contain a suite of three instruments to provide line-of-sight and vector magnetograms, EUV images, and EUV irradiance measurements. Advances such as these must be matched by the development of novel data-analysis and modelling techniques, which can be implemented in near-realtime. Community efforts, such as the Solar Image Processing Workshop series, are therefore essential to achieving accurate and reliable solar monitoring capabilities.

ACKNOWLEDGEMENTS

The authors would like to thank Sam Freeland and Marc DeRosa of Lockheed Martin's Space and Astrophysics Laboratory for providing information on *SolarSoft* Latest Events. This work is supported by a grant from NASA's Living with a Star TR&T program.

REFERENCES

Abramenko, V.: Multifractal analysis of solar magnetograms. Sol Phys, **228**(1–2), 29–42 (2005)

Bentley, R.D., Csillaghy, A., Scholl, I.: The European Grid of Solar Observatories. In: Oschmann, J.M. (ed.) Ground-based telescopes. Proceedings of the SPIE 5493, pp 170–177 (2004)

Berghmans, D., Foing, B.H., Fleck, B.: Automated detection of CMEs in LASCO data. In: Wilson, A. (ed.) From Solar Min to Max: Half a Solar Cycle with SOHO, Proceedings of the SOHO 11 Symposium, ESA SP–508, pp 437–440 (2002)

Delaboudiniere, J.P., Artzner, G.E., Brunaud, J., Gabriel, A.H., Hochedez, J.F., Millier, F., Song, X.Y., Au, B., Dere, K.P., Howard, R.A., 18 coauthors: EIT: Extreme-ultraviolet Imaging Telescope for the SOHO mission. Sol Phys, 162:291–312 (1995)

Fuller, N., Aboudarham, J., Bentley, B.: Filament recognition and image cleaning on Meudon Halpha Spectroheliograms. Sol Phys, **227**(1), 61–73 (2005)

Gallagher, P.T., Moon, Y.-J., Wang, H.: Active-region monitoring and flare forecasting – I. Data processing and first results. Sol Phys, **209**(1), 171–183 (2002)

Gallagher, P.T., Phillips, K.J.H., Lee, J., Keenan, F.P., Pinfield, D.J.: The extreme-ultraviolet structure and properties of a newly emerged active region. Astrophys J, 558, 411–422 (2001)

Gallagher, P.T., Dennis, B.R., Krucker, S., Schwartz, R.A., Tolbert, A.T.: RHESSI and TRACE observations of the 21 April 2002 X1.5 flare. Sol Phys, **210**(1), 341–356 (2002)

Gallagher, P.T., Lawrence, G.R., Dennis, B.R.: Rapid acceleration of a coronal mass ejection in the low corona and implications for propagation. Astrophys J, **588**(1), L53–L56 (2003)

Georgoulis, M.K.: Turbulence in the solar atmosphere: Manifestations an diagnostics via solar image processing. Sol Phys, **228**(1–2), 5–27 (2005)

Hill, F., Bogart, R.S., Davey, A., Dimitoglou, G., Gurman, J.B., Hourcle, J.A., Martens, P.C., Suarez-Sola, I., Tian, K., Wampler, S., Yoshimura, K.: The Virtual Solar Observatory: Status and initial operational experience. In: Oschmann, J.M. (ed.) Ground-based telescopes. Proceedings of the SPIE 5493, pp 163–169 (2004)

Hill, S.M., Pizzo, V.J., Balch, C.C., Biesecker, D.A., Bornmann, P., Hildner, E., Lewis, L.D., Grubb, R.N., Husler, M.P., Prendergast, K., 26 coauthors.: The NOAA Goes–12 Solar X-Ray Imager (SXI) 1. Instrument, Operations, and Data. Sol Phys, 226, 255–281 (2005)

Kolmogorov, A.N.: Dissipation of energy in locally isotropic turbulence. Dokl Akad Nauk SSSR 32:16 (Translated in American Mathematical Society Translations 1958, 8(2), 87) (1941)

Lawrence, J.K., Ruzmaikin, A.A., Cadavid, A.C.: Multifractal measure of the solar magnetic field. Astrophys J, 417, 805–811 (1993)

Mandelbrot, B.: The fractal geometry of nature. WH Freeman and Company, New York (1977)

McAteer, R.T.J., Gallagher, P.T., Ireland, J.: Statistics of active region complexity: A large-scale fractal dimension survey. Astrophys J, **631**(1), 628–635 (2005)

Pizzo, V.J., Hill, S.M., Balch, C.C., Biesecker, D.A., Bornmann, P., Hildner, E., Grubb, R.N., Chipman, E.G., Davis, J.M., Wallace, K.S., 4 coauthors: The NOAA Goes–12 Solar X-ray Imager (SXI). 2. Performance, 226, 283–315 (2005)

Robbrecht, E., Berghmans, D.: Automated recognition of coronal mass ejections (CMEs) in near-real-time. Astron Astrophys. 425, 1097–1106 (2004)

Scherrer, P.H., Bogart, R.S., Bush, R.I., Hoeksema, J.T., Kosovichev, A.G., Schou, J., Rosenberg, W., Springer, L., Tarbell, T.D., Title, A., 3 coauthors: The Solar Oscillations Investigation – Michelson Doppler Imager. Sol Phys, 162, 129–188 (1995)

Schuck, P.W., Chen, J., Schwartz, I.B., Yurchyshyn, V.: On the temporal relationship between Halpha filament eruptions and soft x-ray emissions. Astrophys J, 610, L133–L136 (2004)

Shih, F.Y., Kowalski, A.J.: Automatic extraction of filaments in H-alpha solar images. Sol Phys, 218, 99–122 (2003)

Schrijver, C.J., DeRosa, M.L.: Photospheric and heliospheric magnetic fields. Sol Phys, **212**(1), 165–200 (2003)

Stenborg, G.A., Cobelli, P.J.: A wavelet packets equalization technique to reveal the multiple spatial-scale nature of coronal structures. Astron Astrophys, 398, 1185–1193 (2003)

Young, C.A., Gallagher, P.T.: Interacting CMEs in the low corona. Astrophys J, submitted (2006)

Zharkova, V.V., Aboudarham, J., Zharkov, S., Ipson, S.S., Benkhalil, A.K., Fuller, N.: Solar feature catalogues in EGSO. Sol Phys, 228, 361–375 (2005)

CHAPTER 1.3

MODELING OF SOLAR ENERGETIC PARTICLES IN INTERPLANETARY SPACE

RAMI VAINIO[1], NEUS AGUEDA[2], ANGELS ARAN[2] AND DAVID LARIO[3]
[1]*Department of Physical Sciences, University of Helsinki, Finland*
[2]*Departament d'Astronomia i Meteorologia, Universitat de Barcelona, Spain*
[3]*The Johns Hopkins University, Applied Physics Laboratory, USA*

Abstract: Solar energetic particles (SEPs) in the interplanetary (IP) medium are transported under the influence of electromagnetic fields of the solar wind. These fields consists of the smooth background fields, which can be modeled by the MHD equations governing the expansion of the solar wind, and of the small-scale fluctuations (waves or turbulence) that scatter the particles in pitch angle and act as agents enabling their acceleration at IP shock waves. We review theoretical models of SEP transport and acceleration in the IP medium. We start from the simple analytical approaches (diffusion models), which assume quasi-isotropic particle distributions, and then continue to the more accurate numerical approaches based on the focused transport equation, not making this simplifying assumption. A careful analysis of two SEP events, an impulsive and a gradual one, is presented and the spatial scaling of their peak intensities, differential fluences and time-integrated net fluxes is discussed. We conclude that rather simple scaling laws for these quantities can be obtained for impulsive events but no simple scaling laws can be expected to govern the gradual SEP events

INTRODUCTION

Solar energetic particle (SEP) events are one of the main components of the solar driven space weather: producing most of the energetic particle fluence between 1 and 100 MeV in the interplanetary (IP) medium, they introduce an important radiation risk for space missions in the IP space. In addition, their effects include elevated radiation dose rates and high frequency (HF) radio blackouts at polar airline routes.

Since the 1980's SEP events have been divided in two classes, impulsive and gradual (Cane et al. 1986). The classification is based on bimodal distributions

J. Lilensten (ed.), Space Weather, 27–37.

observed in many variables characterizing the events: impulsive SEP events are related to impulsive X-ray flares, they are typically of short duration (from hours to days) and low intensity, they are electron rich, and their ion abundance ratios show enhancements in ^{3}He and heavies relative to the coronal abundances (e.g., Reames 1999). Gradual SEP events are related to coronal mass ejections (CMEs) and gradual X-ray flares, their duration is longer (from days to a week), they are proton rich and their ion abundance ratios agree with those of the coronal plasma (Reames 1999). It is rather commonly accepted that impulsive events are accelerated in solar flares and gradual SEP events at coronal and IP shocks related to CMEs (Reames 1999). As the sensitivity of the SEP measurements improved as a result of the ACE, Wind, and SOHO missions in the 23rd solar cycle, it was found that the division between the two classes is not as clear as previously believed: a third class of SEP events, i.e., hybrid or mixed events, was introduced (Kocharov and Torsti 2002 and references therein) to include events that look like gradual events from the point of view of their electromagnetic associations, their duration and magnitude, but show properties of impulsive events, e.g., in their ion abundance ratios implying either a direct flare-accelerated component (Cane et al. 2003) or the shock acceleration of supra-thermal remnants in the corona from previous impulsive flares (Tylka et al. 2001).

In this paper we will consider the SEP transport in the IP space, governed by the large-scale heliospheric electromagnetic fields and the wave-particle interactions between the SEPs and the low-frequency magnetic fluctuations of the solar wind plasma. We will start by describing the relevant particle transport equations and then consider numerical modeling of SEP events paying special attention to the spatial development of space-weather relevant quantities in the inner heliosphere.

TRANSPORT EQUATIONS

Diffusion Models

The earliest modern modeling efforts to describe the propagation of particles in the IP medium were based on diffusion–advection model of Parker (1965). This approach has still important applications in the transport of galactic and anomalous cosmic rays in the heliosphere. For SEPs, however, anisotropies and time dependence are very important factors, so only under strongly turbulent conditions can the approach yield accurate results for modeling SEP events. For more general considerations, like for modeling the spatial dependence of SEP event peak fluxes or event fluences, this approach may still give a reasonable starting point.

In the crudest approximation, one can neglect the effects of the solar wind expansion and just consider particle transport in the turbulent IP medium as diffusion. Assuming that the particles diffuse only along the IP magnetic field with a spatial diffusion coefficient $D = v\lambda/3$ we can model the transport using only one spatial coordinate, the radial heliocentric distance r. Here, v is the particle speed and λ is the mean free path of the particles parallel to the magnetic field related to

the amount of fluctuations in the magnetic field, but normally regarded as a free parameter of the transport model. Thus, the SEP transport equation can be written as

$$\frac{\partial n_p}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r} r^2 D_{rr} \frac{\partial n_p}{\partial r}, \tag{1}$$

where $n_p(r, t) = \mathrm{d}^4N/(\mathrm{d}^3r\mathrm{d}p)$ is the particle density per unit momentum, $D_{rr} = D\cos^2\psi$ is the radial diffusion coefficient and ψ is the angle between the radial direction and the local magnetic field.

Although one can justify the application of Eq. (1) to only a small fraction of SEP events, it has the attractive feature that it can be analytically solved for an impulsive injection of particles from the Sun. The result is (Wibberenz et al. 1989)

$$n_p = \frac{\mathrm{d}N}{\mathrm{d}p}\frac{1}{r^3}\frac{2-b}{\Gamma\{3/(2-b)\}}\left(\frac{r^2}{(2-b)^2 D_{rr} t}\right)^{3/(2-b)} \exp\left(\frac{-r^2}{(2-b)^2 D_{rr} t}\right) \tag{2}$$

where $\mathrm{d}N/\mathrm{d}p$ is the momentum spectrum of particles injected to the IP medium per steradian at the solar surface and $D_{rr} \propto r^b$ with $b < 2$ has been assumed.

If the injection is extended in time, the solution of the transport equation can be obtained by using Eq. (2) as a Green's function, i.e., convolving the function with an extended injection profile $Q(E, t)$. Usually, the task of the modeler is to find out the time profile of the SEP injection as well as the IP mean free path. Since the time scales of the coronal and IP acceleration processes may be extended, a mere fit of the modeled omni-directional differential particle intensity, $I = (1/4\pi)n_p$, (hereafter, intensity) to observed intensity–time profile does not provide a unique solution to the problem. To reduce the ambiguity between the contributions of the injection and the transport model to the result, one needs to model the anisotropies of the particle distribution as well. In the simplest approach, one considers the net flux of particles per unit momentum across a spherical surface, which in a diffusion model is given by Fick's law

$$S_r = -D_{rr}\frac{\partial n_p}{\partial r}. \tag{3}$$

This quantity, with a simple relation to the first-order anisotropy, has sensitivity to the value of the diffusion coefficient and, thus, helps to separate the effects of the time-extended particle injection from the effects of particle transport in the solar wind.

We can now obtain a few spatial scaling laws from the diffusion equation and its analytical solution for short-duration solar injections:

(i) the time-integrated radial net flux scales like

$$\int S_r \mathrm{d}t \propto r^{-2};$$

(ii) the time of maximum intensity for an impulsive injection scales like

$$I_{\max}(r) \propto r^{-3};$$

(iii) the time-integrated intensity (or fluence) scales like

$$\int I(r,t)\,\mathrm{d}t \propto 1/(rD_{rr}).$$

Thus, only the time-integrated radial net flux scales like $1/r^2$, although this scaling law is often used for scaling the fluences and the peak intensities as well. We must note, however, that the scaling laws obtained from the diffusion equation are not always valid, but at least the following conditions have to be met: (a) the mean free path of the SEPs is much smaller than the heliocentric distance of the observer, i.e., $\lambda_{rr} \ll r$; (b) the time of maximum intensity of the solution (2) is much smaller than the adiabatic cooling time, i.e., $r/\lambda_{rr} \ll v/(2V)$, where V is the solar wind speed; (c) the duration of the injection at the Sun is shorter than the time of maximum intensity; and (d) the site of the injection, r_0, is close to the Sun, i.e., $r_0 \ll r$. For the validity of scaling-law (i), only the conditions (b) and (d) are necessary, the remaining scaling laws require all the conditions. For cases not meeting these (rather strict) conditions, we need to resort to numerical methods to investigate the scaling.

Focused Transport Model

Because of the focusing effect due to the outwards decreasing magnetic field magnitude in the inner heliosphere, the anisotropies become too large for the diffusion model to be applicable if the mean free path in the IP medium is comparable to the radial distance from the Sun. We, thus, need a transport equation describing the evolution of the particle distribution function $f(s, p, \mu, t)$, which gives the number of particles per unit volume of the six-dimensional phase space $(\mathbf{r}, \mathbf{p})$. In the focused transport approximation, the distribution function is governed by streaming along the magnetic field lines, by scattering off the fluctuations of the magnetic field and by magnetic focusing (i.e., mirroring) in the outwards decreasing magnetic field. The distribution function is a function of the coordinate measured along the mean magnetic field, s, the particle momentum p, the cosine of pitch-angle μ, i.e., the angle between the magnetic field and the velocity vector of the particle, and time t. Its relations to the particle density and streaming per unit momentum are

$$\begin{aligned} n_p(s,t) &= 2\pi p^2 \int f(s,p,\mu,t)\mathrm{d}\mu \\ S_p(s,t) &= 2\pi p^2 \int v\mu f(s,p,\mu,t)\mathrm{d}\mu \end{aligned} \tag{4}$$

where now the streaming has been defined with respect to a unit area perpendicular to the magnetic field instead of a spherical surface like in Eq. (3). The radial streaming can be obtained from this quantity by multiplying with $\cos\psi$.

The equation governing the evolution of the particle distribution is the focused transport equation (Roelof 1969)

$$\frac{\partial f}{\partial t} + v\mu\frac{\partial f}{\partial s} + \frac{1-\mu^2}{2L}v\frac{\partial f}{\partial \mu} = \frac{\partial}{\partial \mu}D_{\mu\mu}\frac{\partial f}{\partial \mu} \tag{5}$$

where the second term describes streaming of particles along the magnetic field lines, the third term describes the focusing of particles because of the mirror force and the last term accounts for the effect of magnetic fluctuations, which is modeled by pitch-angle diffusion. This simple form of the equation still neglects adiabatic deceleration (see, Ruffolo 1995, for the full equation), but at the energies of interest for space weather, this is typically a small effect. The pitch-angle diffusion coefficient has the form

$$D_{\mu\mu} = \frac{1}{2}(1-\mu^2)\varphi(\mu), \tag{6}$$

where, following the quasi-linear theory the (Jokipii 1966; Jeakel and Schlickeiser 1992), scattering frequency is usually modeled as $\varphi(\mu) = \varphi_0|\mu|^{q-1}$, with $q \in [1, 2]$ and related to the scattering mean free path as

$$\lambda = \frac{3v}{4}\int \frac{1-\mu^2}{\varphi(\mu)}\mathrm{d}\mu. \tag{7}$$

MODELING OF IP PARTICLE TRANSPORT AND INJECTION IN SEP EVENTS

In the following, we will consider two examples of modeling SEP events in the IP medium. Similar approaches have been used to model dozens of particle events by a number of authors (e.g., Heras et al. 1992; Torsti et al. 1996; Lario et al. 1998; Aran et al. 2004). We fit the injection and transport parameters to SEP observations at 1 AU using the focused transport equation, and then study the modeled event at different distances from the Sun, paying special attention to the inner parts of the heliosphere. This region will be accessed by several spacecraft in the next decade, including ESA's BepiColombo and Solar Orbiter missions.

Impulsive SEP Events

It is reasonable to consider the source of the SEPs in impulsive events to be close to the Sun. The focused transport equation is first solved for an impulsive injection from the corona to the IP magnetic field to obtain a Green's function of IP transport. The SEP injection at the root of an IP flux tube is then parameterized using a convenient mathematical function, e.g., the diffusive Reid–Axford profile (Reid 1964) with rise time and decay time constants. This is convolved with the simulated Green's function and the result is compared with observations of intensity and anisotropy. The parameters of the injection and IP transport are then varied until the best fit is found.

We have applied this method to the electron intensities measured by the EPAM instrument (Gold et al. 1998) onboard ACE during the impulsive event of May 1, 2000 (Fig. 1). The event is related to an impulsive M1 class X-ray flare from N20°W54° peaking at 10:27 UT. Although the event is associated with a narrow,

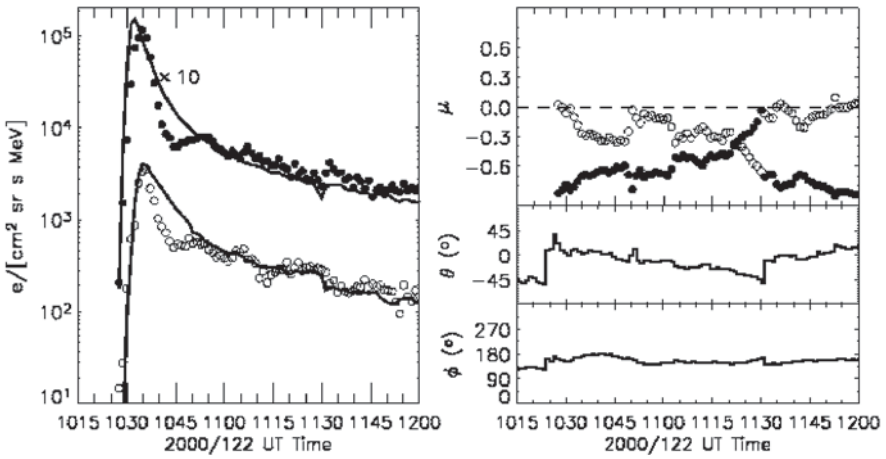


Figure 1. Impulsive SEP event of May 1, 2000 as observed by ACE/EPAM. Electron intensities at 175–290 keV in two sectors of the instrument (left) and their zenith pitch-angle cosines (right: μ) with respect to the magnetic field direction (right: elevation θ, azimuth ϕ) as measured by ACE/MAG (Smith et al. 1998). On the left, the curves give the modeled intensities and the circles give the data

fast CME as well, its characteristics are typical to an impulsive event (Kahler et al. 2001; Ho et al. 2003; Mason et al. 2004). We have fitted the four electron energy channels of the EPAM/LEFS-60 telescope (at 45–312 keV) using sectored intensities sensitive to anisotropies as well. Using a model of the full directional response of each sector, we calculate the sectored intensity of the electrons obtained from a convolution of a Monte Carlo simulated Green's of IP transport and the Reid–Axford profile. The best fit is found varying the parameters trying to minimize χ^2. Details of the modeling will be published elsewhere, but the simulation is very similar to those previously used in the studies of IP transport (e.g., Torsti et al. 1996; Kocharov et al. 1998).

The model fits the data satisfactorily at 1 AU. We only show one energy channel (175–312 keV) of four and two sectors of eight, but the other sectors and energy channels are taken into account in our fitting procedure as well and the quality of the fit is similar in all of them. The best-fit time scales of the rise and decay of the injection are 4.2 minutes and 1.2 minutes, respectively, and the radial mean free path, assumed to be independent of energy, is 0.6 AU.

We have investigated the time-intensity profiles of the modeled event at radial distances of 0.2 AU, 0.3 AU, 0.7 AU and 1 AU. We determined the time-integrated net-flux, the peak intensity, and the differential fluence of the event from the simulations (Fig. 2). The scaling of peak intensities is less steep than predicted by diffusion, which can be understood, because scatter-free transport corresponds to the scaling law $\propto \sec \psi / r^2$ and the event has a mean free path too long to be well described by diffusion. The scaling law of fluence, on the other hand, is steeper than the prediction of the diffusion-law (r^{-1} for a spatially constant radial mean

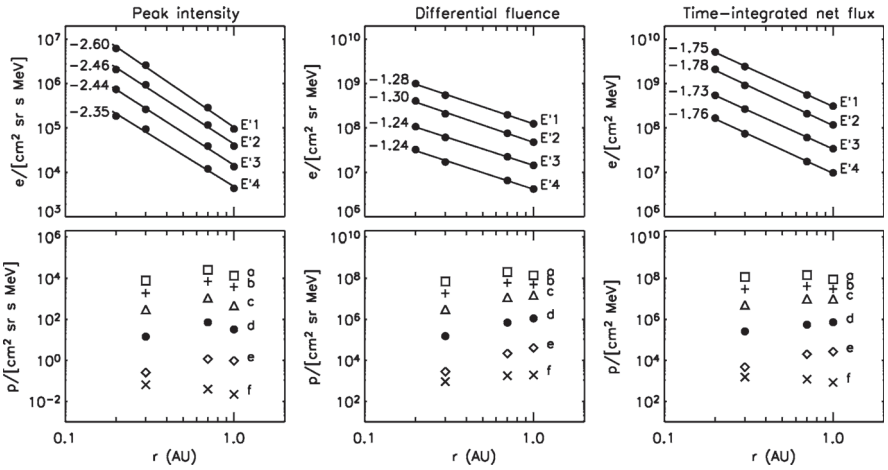


Figure 2. Peak intensities, differential fluxes and time-integrated net flux of an impulsive electron event (upper panels) and a gradual proton event (lower panels) as a function of radial distance in several energy channels. The electron energy channels are E1′: 45–62 keV, E2′: 62–102 keV, E3′: 102–175 keV, and E4′: 175–312 keV. See Fig. 3 for a description of proton energy channels a–f

free path), which can be understood as well, because the result for scatter-free transport would again be $\propto \sec \psi / r^2$ for a solar source. The time-integrated net flux behaves like the diffusion theory predicts, as expected, because the same scaling law can be obtained from the focused transport equation (5).

Gradual SEP Events

Gradual events in the IP space are more difficult to model than the impulsive ones for several reasons: (1) The source of the particles is the moving shock front driven by the CME through the IP medium; (2) the large spatial extent of the CME system and the long duration of the event means that transport conditions can vary during the event; and (3) the large intensities of the SEPs lead to a non-linear coupling between the accelerated particles and the plasma waves responsible for their scattering (e.g., Ng et al. 2003). Despite of these complications, dozens of gradual SEP events have been successfully modeled over the past two decades using an assumption of a particle source at the position of the IP shock and tracing the transport of the particles in the surrounding IP medium (Heras et al. 1992; Torsti et al. 1996; Lario et al. 1998; Aran et al. 2004, 2005).

As an example of gradual SEP event modeling, we consider the SEP event on June 6–8, 2000, observed by the ACE/EPAM at the L1 point, and by the CPME/IMP–8 (Sarris et al. 1976) orbiting the Earth (Fig. 3). The event was associated with a CME-driven shock that arrived at L1 at 08:41 UT of 8 June (DOY = 160.362). The CME was first observed by SOHO/LASCO C2 coronagraph on 6 June at 15:54 UT

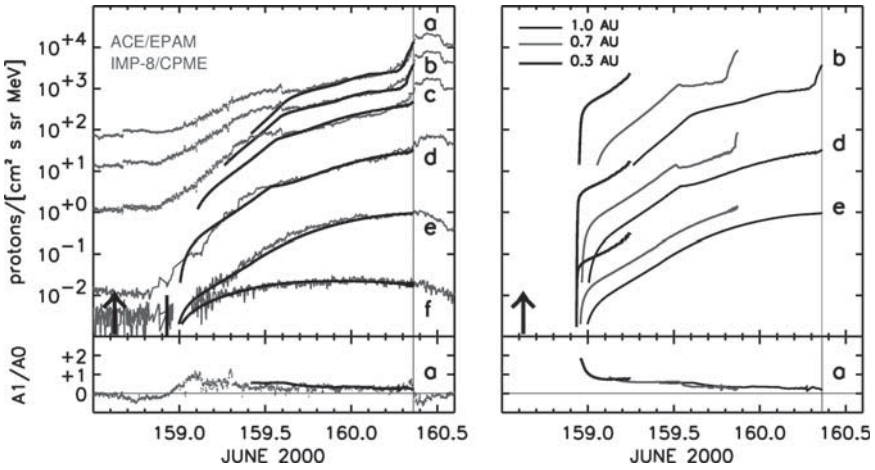


Figure 3. Gradual SEP event of June 2000 as observed by ACE/EPAM and IMP-8/CPME. The observed and fitted intensities and anisotropies at 1 AU (left panel) are given in addition to the modeled ones at 0.3 AU and 0.7 AU (right panel). The energy channels are denoted with labels a–f: a: 0.58–1.06 MeV, b: 1.06–1.90 MeV, c: 1.90–4.80 MeV (from ACE/EPAM); d: 4.6–15.0 MeV, e: 15.0–25.0 MeV, f: 25.0–48.0 MeV (from IMP–8/CPME)

with an estimated speed of 1119 km s^{-1}. The associated X3.2/3B N20°E18° flare started at 14:58 UT June 6 (DOY = 158.624), marked by the arrow in Fig. 3.

For eastern SEP events, modeling of the CME-driven shock evolution is essential to get an idea, when the observer obtains magnetic connection to the shock and, therefore, is able to observe the particles accelerated by the shock wave. We have simulated the propagation of the shock with the 2½D MHD code by Wu et al. (1983). Using the same functional form of the input pulse assumed by Smith and Dryer (1990), we inject a pulse centered at E18° with speed $V_s = 1138\,\mathrm{km\,s^{-1}}$, angular width $\omega = 140°$ and duration $\tau = 1\,\mathrm{h}$. According to the simulation, ACE obtains the connection to the shock near the end of June 6, marked by the vertical line in Fig. 3.

SEP transport ahead of the shock is modeled by solving the focused transport equation using a finite difference method (Lario et al. 1998; Aran et al. 2005). The injection of particles at the shock is described by a semi-empirical relation between the shock strength and injection rate. The mean free path in the upstream medium is taken to be $\lambda = \lambda_0(p/p_0)^{1/2}$, with the best fit values of $\lambda_0 = 0.1\,\mathrm{AU}$ at $p_0 = (2m_pE_0)^{1/2}$ and $E_0 = 0.789\,\mathrm{MeV}$. A region just upstream the shock with a mean free path $\lambda = \lambda_0(p/p_0)^{-4/5}$ is assumed where $\lambda_0 = 0.01\,\mathrm{AU}$. This region starts to act at 21:00 UT and has a width that varies with the energy: 0.04 AU for $2\,\mathrm{MeV} < E < 15\,\mathrm{MeV}$, 0.06 AU for $1\,\mathrm{MeV} < E < 2\,\mathrm{MeV}$ and 0.07 AU for $E < 1\,\mathrm{MeV}$. This region reproduces the effects that turbulence generated by the accelerated particles has on the particle transport close to the shock and allows us to fit the low-energy observations at the time of shock arrival. The fitted intensities

and anisotropies are given in Fig. 3, along with the modeled intensities within the same flux tube at 0.3 and 0.7 AU.

The peak intensities, fluences, and time-integrated net fluxes of the event are given in Fig. 2. Note that the time integration of this event only covers the period before the shock passage. Clearly, the scaling obtained is very different from the impulsive event.

MODELING OF IP PARTICLE ACCELERATION

Particle acceleration in the IP medium occurs in shock waves formed in the compression regions between streams of different velocities. These include the forward and reverse shocks bounding the corotating interaction regions (CIRs) and the bow shocks of the fast CMEs. The CIR shocks are usually formed at radial distances 2–5 AU and are not capable of producing very energetic particle events in the inner heliosphere. Thus, the focus, from the point of view of space weather effects, is on the CME-driven shocks.

Physical modeling of particle acceleration is usually based on solving Parker's (1965) diffusion–convection transport equation with a source of low-energy particles placed at the shock. This model can be combined with the generation of MHD waves by the streaming accelerated particles self-consistently and solved in steady state and planar geometry (Bell 1978). Such an approach can be used to describe the local acceleration of low-energy (below 1 MeV) ions during times of IP shock passage (Lee 1983; Gordon et al. 1999), i.e., the so-called energetic storm particle (ESP) events. The same quasi-stationary particle acceleration model can be combined with a non-diffusive transport model to describe particle intensities at large distances upstream from the shock. This kind of approach has been used in both analytical (Lee 2005) and numerical (e.g., Rice et al. 2003) calculation of SEP event intensities in gradual events. While these physical models give good insights to the particle acceleration processes in the IP medium, we do not yet know enough details of the acceleration mechanism and of the shock itself to allow detailed comparisons of theory and observations in individual gradual SEP events. Thus, transport modeling using phenomenological SEP source functions still remains an important tool for space weather studies.

SUMMARY AND OUTLOOK

We have reviewed the transport models used to describe the evolution of intensities and anisotropies during SEP events. In addition to giving fresh examples of typical modeling of impulsive and gradual SEP events, we used the models to calculate the peak intensities, the fluences and the time-integrated net fluxes of the events as a function of the radial distance from the Sun in the inner heliosphere. These quantities are the most important ones concerning the development of the solar corpuscular radiation environment as a function of distance from the Sun. The results were compared to the expectation derived from a simple analytical diffusion model.

The comparison shows that impulsive events show qualitatively similar scaling to the diffusion model, although at high values of the scattering mean free path, the scaling laws move closer to the scaling of scatter-free transport, as expected. The scaling of the gradual events, on the other hand, showed no similarities to the simple modeling. In our simulation, all the quantities under investigation showed more or less constant values as a function of radius, as a result of the interplay between geometry and time dependence of the source. On the other hand, we did not accurately model the non-linear coupling of the particles to the magnetic fluctuations responsible for their scattering in the IP medium (Ng et al. 2003). This effect may lead to completely different scaling laws: theoretical estimates predict that particle trapping close to the source is more efficient when the shock is close to the Sun (Vainio 2003). Thus, at small heliocentric distances the ESP events may be larger than close to 1 AU.

Our study demonstrates a pressing need for conducting more extensive modeling studies as well as analysis of observations at different distances from the Sun, to obtain reliable extensions of the present engineering models for SEP events like the model SOLPENCO (Aran et al. 2004).

ACKNOWLEDGEMENTS

We wish to thank B. Sanahuja for valuable discussions. We thank the ACE EPAM/SWEPAM/MAG teams for providing the ACE data used in this paper. We acknowledge the use of 330-s averaged IMP–8 CPME data available at the JHU/APL web site. RV acknowledges financial support of COST–724. DL was partially supported by NASA grant NAG5–13487. NA and AA acknowledge the financial support of the Ministerio de Ciencia y Tecnología (Spain) under the Project AYA2004–03022 and partial computational support by the Centre de Supercomputació de Catalunya (CESCA).

REFERENCES

Aran, A., Sanahuja, B., Lario, D.: An engineering model for solar energetic particles in interplanetary space. Final Report of ESA/ESTEC Contract 14098/99/NL/MM. Available from http://www.am.ub.es/∼blai (2004)

Aran, A., Sanahuja, B., Lario, D.: A first step towards proton flux forecasting. Adv Space Res, 36, 2333–2338 (2005)

Bell, A.R.: The acceleration of cosmic rays in shock fronts. I. Mon Not Roy Astron Soc, 182, 147–156 (1978)

Cane, H.V., McGuire, R.E., von Rosenvinge, T.T.: Two classes of solar energetic particle events associated with impulsive and long-duration soft X-ray flares. Astrophys J, 301, 448–459 (1986)

Cane, H.V., von Rosenvinge, T.T., Cohen, C.M.S., Mewaldt, R.A.: Two components in major solar particle events. Geophys Res Lett, 30(12), 8017, DOI 10.1029/2002GL016580 (2003)

Gold, R.E., Krimigis, S.M., Hawkins, S.E., III, et al.: Electron, Proton, and Alpha Monitor on the Advanced Composition Explorer spacecraft. Space Sci Rev, 86, 541–562 (1998)

Gordon, B.E., Lee, M.A., Möbius, E., Trattner, K.J.: Coupled hydromagnetic wave excitation and ion acceleration at interplanetary traveling shocks and Earth's bow shock revisited. J Geophys Res, **104**(A12), 28263 (1999)

Heras, A.M., Sanahuja, B., Smith, Z.K., Detman, T., Dryer, M.: The influence of the large-scale interplanetary shock structure on a low-energy particle event. Astrophys J, 391, 359–369 (1992)
Ho, G.C., Roelof, E.C., Mason, G.M., et al.: Onset study of impulsive solar energetic particle events. Adv Space Res, 32, 2679–2684 (2003)
Kahler, S.W., Reames, D.V., Sheeley, N.R., Jr.: Coronal mass ejections associated with impulsive solar energetic particle events. Astrophys J, 562, 558–565 (2001)
Kocharov, L., Torsti, J.: Hybrid solar energetic particle events observed on board SOHO. Solar Phys, 207, 149–157 (2002)
Kocharov, L., Vainio, R., Kovaltsov, G.A., Torsti, J.: Adiabatic deceleration of solar energetic particles as deduced from Monte Carlo simulations of interplanetary transport. Solar Phys, 182, 195–215 (1998)
Lario, D., Sanahuja, B., Heras, A.M.: Energetic particle events: efficiency of interplanetary shocks as 50 keV $< E <$ 100 MeV proton accelerators. Astrophys J, 509, 415–434 (1998)
Lee, M.A.: Coupled hydromagnetic wave excitation and ion acceleration at interplanetary traveling shocks. J Geophys Res, 88, 6109–6119 (1983)
Lee, M.A.: Coupled hydromagnetic wave excitation and ion acceleration at an evolving coronal/interplanetary shock. Astrophys J Suppl Ser, 158, 38–67 (2005)
Mason, G.M., Wiedenbeck, M.E., Miller, J.A., et al.: Spectral properties of He and heavy ions in ^{3}He-rich solar flares. Astrophys J, 574, 1039–1058 (2004)
Ng, C.K., Reames, D.V., Tylka, A.J.: Modeling shock-accelerated solar energetic particles coupled to interplanetary Alfvén waves. Astrophys J, 591, 461–485 (2003)
Jokipii, J.R.: Cosmic-ray propagation. I. Charged particles in a random magnetic field. Astrophys J, 146, 480 (1966)
Jeakel, U., Schlickeiser, R.: The Fokker-Planck coefficients of cosmic ray transport in random electromagnetic fields. J Phys G, 18, 1089–1118 (1992)
Parker, E.N.: The passage of energetic charged particles through interplanetary space. Planet Space Sci, 13, 9 (1965)
Reames, D.V.: Particle acceleration at the Sun and in the heliosphere. Space Sci Rev, 90, 413–491 (1999)
Reid, G.C.: A Diffusive Model for the Initial Phase of a Solar Proton Event. J Geophys Res, 69, 2659 (1964)
Rice, W.K.M., Zank, G.P., Li, G.: Particle acceleration and coronal mass ejection driven shocks: Shocks of arbitrary strength. J Geophys Res, **108**(A10), 1369, DOI 10.1029/2002JA009756 (2003)
Roelof, E.C.: Propagation of solar cosmic rays in the interplanetary magnetic field. In: Ögelman, H., Wayland, J.R. (eds), Lectures in high-energy astrophysics. NASA, Washington DC, p 111 (1969)
Ruffolo, D.: Effect of adiabatic deceleration on the focused transport of solar cosmic rays. Astrophys J, 442, 861–874 (1995)
Sarris, E.T., Krimigis, S.M., Armstrong, T.P.: Observations of magnetospheric bursts of high energy protons and electrons at 35 R_E with IMP–7. J Geophys Res, 81, 2341–2355 (1976)
Smith, C.W., L'Heureux, J., Ness, N.F. et al.: The ACE magnetic fields experiment. Space Sci Rev, 86, 613-632 (1998)
Smith, Z.K., Dryer, M.: MHD study of temporal and spatial evolution of simulated interplanetary shocks in the ecliptic plane within 1 AU. Solar Phys, 129, 387–405 (1990)
Torsti, J., Kocharov, L.G., Vainio, R., Anttila, A., Kovaltsov, G.A.: The 1990 May 24 solar cosmic-ray event. Solar Phys, 166, 135 (1996)
Tylka, A., Cohen, C.M.S., Dietrich, W.F., et al.: Evidence for remnant flare suprathermals in the source population of solar energetic particles in the 2000 Bastille day event. Astrophys J, 558, L59–L63 (2001)
Vainio, R.: On the generation of Alfvén waves by solar energetic particles, Astron Astrophys, 406, 735–740 (2003)
Wibberenz, G., Kunow, H., Iwers, B., Kecskemety, K., Somogyi, A.: Coronal and interplanetary transport of solar energetic protons and electrons. Solar Phys, 124, 353–392 (1989)
Wu, S.T., Dryer, M., Han, S.M.: Non-planar MHD model for solar flare-generated disturbances in the heliospheric equatorial plane. Solar Phys, 84, 395–418 (1983)

CHAPTER 1.4

SIMULATING CME INITIATION AND EVOLUTION: STATE-OF-THE-ART

S. POEDTS, B. VAN DER HOLST, C. JACOBS, E. CHANÉ, G. DUBEY AND D. KIMPE

CPA, K.U.Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

Abstract: A review is given of some recent results on CME initiation and evolution simulations obtained at the Centre for Plasma Astrophysics (CPA, K.U.Leuven) on the background of the international developments in this very dynamic field

INTRODUCTION AND MOTIVATION

The effects of space weather are defined by components on the sun as well as on Earth. The most important solar components are solar flares, halo Coronal Mass Ejections (CMEs) and the solar wind. Halo CMEs, which are often associated with a solar flare, can affect the earth by the release of a magnetic cloud of energetic particles. The speeds at which they travel through space range from 200 to more than 2000 km/s, the average mass of CME plasma sent towards the Earth is of the order of 10^{12}–10^{13} kg and the average energy of a CME event lies around 10^{24}–10^{25} J. The very fast CMEs create strong shock waves in which particles are accelerated giving rise to so-called gradual Solar Energetic Particle (SEP) events. Clearly, the background solar wind is important too, also because the locations of the so-called coronal holes, which produce the fastest solar wind, can enhance the effect that CMEs have on the earth's magnetosphere. We here focus on the modeling of the initiation and the interplanetary (IP) evolution of CMEs because of their crucial role in space weather.

SOLAR WIND MODELING: RECENT DEVELOPMENTS

There were a lot of recent developments in the domain of solar wind modeling. Due to the availability of ever more CPU power and computer memory, advanced wind models now incorporate observational data as boundary conditions. This yields

J. Lilensten (ed.), Space Weather, 39–48.

realistic simulations enabling specific event studies and a detailed comparison of the simulation results with the observations. The Center for Space Environment Modeling (CSEM) at the University of Michigan has developed the first coupled model of the inner heliosphere extending from the low solar corona to the well-beyond-Earth orbit. At the heart of the coupled model is the Space Weather Modeling Framework (SWMF), a high-performance flexible computational tool that enables coupling state-of-the art models of the solar corona, the solar wind and solar energetic particles (see Gombosi et al., 2004, Tóth et al., 2005). Lionello et al. (2003) developed a three-dimensional magnetohydrodynamic (MHD) model of the solar corona and of the solar wind incorporating thermal conduction along the magnetic field, radiation losses, and heating into the energy equation. Lee et al. (2004) extended the CISM (Center for Integrated Space Weather Modeling) inner heliospheric model CORHEL to 10 AU to investigate how well this solar magnetogram-based 3 D MHD model describes the solar wind influence on Saturn's magnetosphere. Odstrcil et al. (2004) also developed a coupled numerical wind model deriving the ambient solar wind from coronal models utilizing photospheric magnetic field observations while transient disturbances are derived from geometrical and kinematic fitting of coronagraph observations of coronal mass ejections (CMEs). Odstrcil et al. (2005) applied the model to study the propagation of an interplanetary CME in evolving wind structures. Amari et al. (2005) considered a three-dimensional bipolar magnetic field driven into evolution by the slow turbulent diffusion of its normal component on the boundary. At the CPA, Jacobs et al. (2005) made a first attempt to quantify the effect of the background solar wind model on the evolution of IP CMEs by superposing the same simple CME model on three different 2.5 D (spherical, axi-symmetric) wind models reproduced with the same numerical code, the same numerical scheme, the same boundary conditions and the same numerical grid.

SIMULATION MODELS FOR CME INITIATION

Heliospheric models of CME propagation and evolution provide an important insight into the dynamics of CMEs and are a valuable tool for interpreting interplanetary in situ observations. Moreover, they represent a virtual laboratory for exploring conditions and regions of space that are not conveniently or currently accessible by spacecraft (Riley et al., 2005). Serious efforts have been undertaken to study coronal initiation and solar wind propagation together. Roussev et al. (2003a, b) developed the capability to use observed synoptic magnetograms to drive the coupled corona-solar-wind model in order to simulate the solar wind and to superpose on this wind a 3 D flux-rope model for a CME based on a loss of equilibrium, i.e. not on an initially unsatisfied force balance as in many earlier (and current) simulations. Manchester et al. (2004a, c) also modeled erupting flux ropes and the resulting CMEs in full 3 D MHD. Sokolov et al. (2004) then included a field line advection model to obtain a coupled corona-solar-wind-SEP model. Manchester et al. (2004b) used the SWMF tool to model specific Space Weather events from

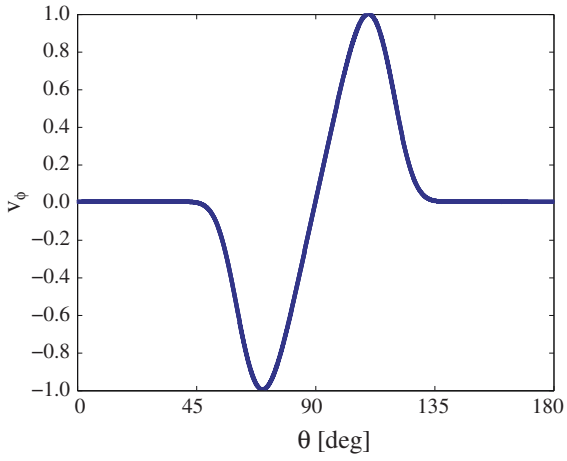
the Sun to the Earth, including the initiation of the CME and its evolution during its interplanetary propagation. Manchester et al. (2005) focused on the CME shock and sheet structures relevant for particle acceleration while Lugaz et al. (2005) concentrated on the evolution of the density structure of the CMEs. Chané et al. (2005, 2006) then studied the effect of the CME initiation parameters on the CME evolution, in particular, these authors focused on the effect of the polarity of the initial magnetic flux rope on the IP CME evolution path.

Foot Point Driven CMEs

Recently, at the CPA we created CMEs from each of the three steady 2.5 D wind models mentioned before by shearing the magnetic foot points by adding an extra longitudinal velocity at the solar surface in accordance to Mikic (1994). The added longitudinal velocity is given by

$$V_{0\varphi} = V_0(t)\Theta(\theta)\exp[(1-\Theta^4)/4],$$

where $\Theta = (\theta - 90)/\Delta\theta_m$, θ is the co-latitude in degrees, and $v_0(t)$ a function that specifies the time profile. The θ-dependence is illustrated in the figure on the right. The simulation stops when a time $t_{max} = 180\,h$ is reached. The shearing reaches its maximum value $\Delta\theta_m$ degrees above and below the equator. In the simulations $\Delta\theta_m = 20°$ and the maximum shear velocity varied between values of 3, 6, or 9 km/s. The model for the background wind affects the time of formation of the flux rope as is clear from Fig. 1 showing the evolution of the total magnetic energy in time. The amount of magnetic energy is expressed with respect to the total amount of magnetic energy in the stationary background wind.



The faster the shearing, the earlier the flux rope is formed. Also the time-interval between the succeeding flux ropes scales with the shearing velocity. The shear

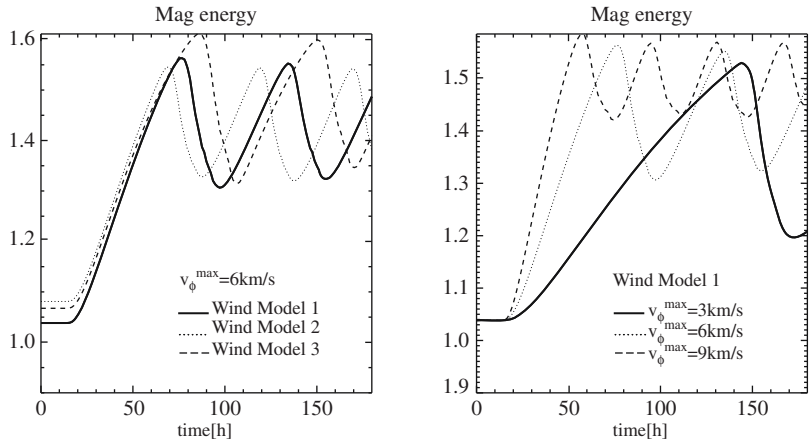


Figure 1. The evolution of the magnetic energy in time. Left: for three wind models and for maximum shear velocity, $v_\varphi{}^{\max} = 6\,\mathrm{km/s}$, of 6 km/s. Right: for a polytropic wind and for three different (maximal) shearing velocities

velocity also determines if a flux rope will be formed or not since shearing at a too low velocity does not lead to the formation of a flux rope. On the other hand, the faster the field lines are sheared, the faster the flux rope is moving upward. As clear from Fig. 1, the background wind also affects the energetics of the CME event and the velocity at which the flux rope is launched. For the given shearing velocities, which are rather high compared to observed velocity patterns in the solar photosphere, it turns out impossible to create fast CMEs with this initiation mechanism.

The helicity of the involved magnetic fields is generally believed to hold an important key to the onset of solar eruptions such as flares and CMEs. In the case of axial symmetry, the gauge invariant relative magnetic helicity can be shown to reduce to (Antiochos et al., 2002):

$$H_r = 2\int_V A_\varphi B_\varphi \mathrm{d}V,$$

with $\vec{A}$ the magnetic vector potential with a known longitudinal component (it is one of the dependent variables). A plot of the evolution in time of the total amount of relative helicity in the simulation volume is given in Fig. 2. Values for the total amount of relative helicity at the onset of the dramatic rise of the streamer and at the moment of flux rope formation are indicated with the $+$-signs and $\times$-signs in Fig. 2, respectively.

Magnetic Flux Emergence

CMEs can also be triggered by the emergence of additional magnetic flux of the same or the opposite polarity as the overlying magnetic field. Dubey et al. (2005)

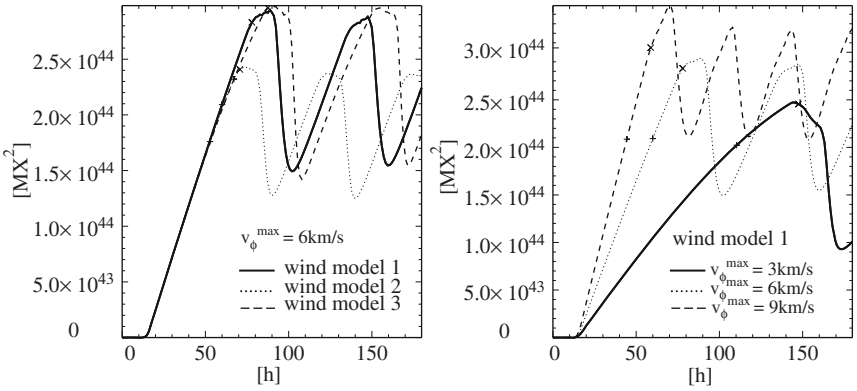


Figure 2. Evolution in time of the total relative helicity in the simulation volume (1–30 $R_{<001>}$). Left: three different wind models, $v_{\varphi}{}^{\max} = 6\,\mathrm{km/s}$. Right: three different shear velocities imposed on the streamer in the polytropic wind model

considered a flux rope in a dipole magnetic field that is initially kept stable by means of a line current (cf. Chen and Shibata, 2000). Allowing magnetic flux of the opposite polarity to emerge from the solar surface then breaks the force balance and causes the flux rope to be expelled. The evolution parameters, such as the velocity and the acceleration, depend on the flux emergence rate and on the total amount of flux that is emerged. We obtain velocities in the order of 350–400 km/s, which is precisely in the range of what is observed for the majority of the CMEs.

When the background dipole field is replaced by a genuine solar wind model, however, faster CMEs can be created with the same triggering mechanism. This is illustrated in Fig. 3 where the obtained velocities are plotted for the 'case B' of Dubey et al. (2006) with a background wind instead of a dipole magnetic field. The corresponding total amount of emerged magnetic flux then ranges from -2.2×10^{22} Mx (for $c_e = -1$) to -1.98×10^{23} Mx (for $c_e = -9$) in the Northern hemisphere, and the exact opposite in the Southern hemisphere. The corresponding flux emergence rates then varies from 1.22×10^{19} Mx($c_e = -1$) to 1.10×10^{20} Mx($c_e = -9$).

'Density Driven' CMEs

In order to study the propagation of *fast* coronal mass ejections and the related shock waves in the interplanetary space from the solar corona up to 1 AU, a very simple but frequently used CME model is adopted. A high density and high pressure, magnetized plasma blob is superposed on the background steady state solar wind model with an initial velocity, v_{cme}, in a prescribed radial direction, θ_{cme}. The velocity and density profiles in the initial disturbance are both of the form:

$$y = \frac{y_{cme}}{2}\left(1 - \cos\pi\frac{d_{cme} - \mathrm{d}}{d_{cme}}\right),$$

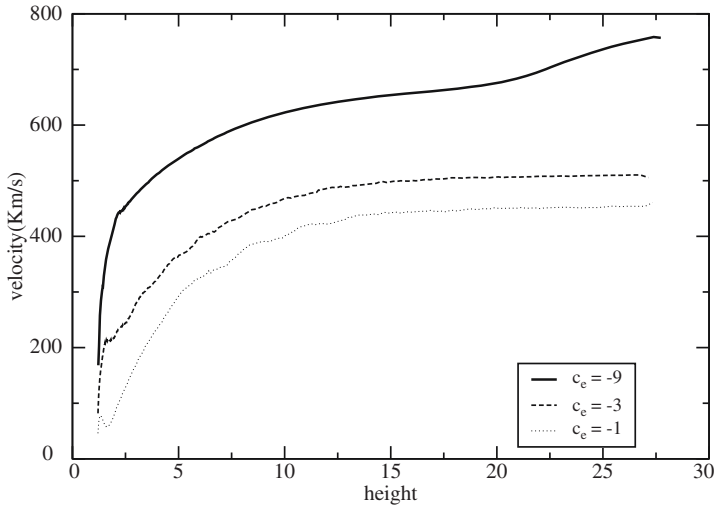


Figure 3. Velocity versus height (in solar radii) for three different amounts of magnetic flux emerged in a time span of 30 min (see the corresponding text for the details)

where y indicates the density or the radial velocity, y_{cme} is the maximum density or radial velocity in the plasma bubble, d_{cme} is the radius of the bubble and d the distance to the center of the bubble. The CMEs are further characterized by a given density, magnetic field strength and magnetic polarity. The initial CME magnetic field and the background wind magnetic field can have the same or the opposite polarity for an *inverse* and *normal* CME, respectively (see Chané et al., 2005). For the CMEs discussed in the present paper, $d_{cme} = 0.29\,\mathrm{R}_{\odot}$, $d = 1.5\,\mathrm{R}_{\odot}$, $v_{\mathrm{cme}} = 1000\,\mathrm{km/s}$ and the density is 5 times higher than the density on the surface of the Sun. In the initial plasma blob, the maximal magnetic field strength is chosen to be 0.344 mT (3.44 Gauss).

INTERPLANETARY CME EVOLUTION

Using their streamer and flux-rope MHD model, Wu et al. (2004) have numerically examined the Low and Zhang (2002) suggestion that the two types of CMEs (i.e. constant speed (fast) and accelerated (slow)) are caused by the initial magnetic topology due to the effect of magnetic reconnection processes. The numerical simulation shows in addition to the magnetic topology, that the solar surface condition also plays an important role to determine the two types of CMEs (see Wu et al., 2005a, 2005b). To study the CME propagation, Wang et al. (2005) have employed the LASCO and ACE observations of January 20, 2001 CME-CME interaction event together with MHD model to investigate the acceleration and deceleration and cannibalization of the CMEs. It was demonstrated that the acceleration and deceleration of the interacting CMEs are caused by the background solar

wind and the deflection of each other due to the interaction. The CME cannibalization is caused by the magnetic reconnection.

It turns out that the polarity of the flux rope magnetic field has a great influence on the evolution of the CME (see Chané et al., 2005). The polarity influences the mass distribution inside the CME, the spread angle, the evolution path, and also the velocity of the shock front. When launching a magnetized CME outside the equatorial plane the magnetic forces push the CME towards or away from the equator, depending if the polarity of the flux rope is *inverse* or *normal* (see Fig. 4). Also it is seen that the normal magnetized CME moves slightly faster than the magnetized inverse. This has consequences for predicting the time of arrival and geo-effectiveness of a CME. Clearly, it is closely related to the statistical studies of the origin of geo-effective CMEs mentioned by Zhukov (2006).

In spite of the rather simple and naive CME model used in these simulations, they yield some very good and realistic results. First of all, we were able to reproduce the magnetic topology predicted by Low and Zhang (2002). Behind the CME, in the equatorial plane, reconnection processes occur which lead to a gradually re-building of the helmet streamer and yield back flows along the helmet streamer, very much like seen in observations. Another interesting fact is that we can extend these CME simulations to 1 AU and that we can reproduce some of the cases in the ACE database, where the strength jump in magnetic field, the profile for the velocity, and the time of arrival are matching surprisingly well. As an example, we consider the full halo CME of April 4, 2000. At 16:32 UT, after a data gap of 90 minutes, the CME was observed for the first time in the C2 frame. At 15:24 UT, EIT observed a solar flare probably related to this CME event. According to the C3 measurements, the plane-of-sky speed of the CME was 984 km/s. We tried to

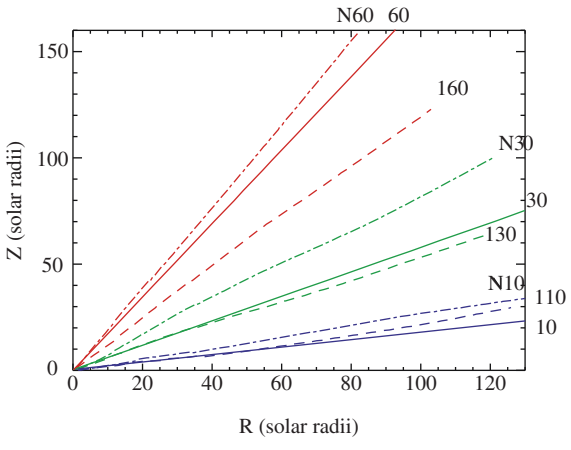


Figure 4. Evolution path of the center of relative mass for *inverse* (dashed lines) and *normal* (dashed-dot lines) CMEs. The CMEs were launched on 10°, 30°, and 60°, respectively. The solid lines show the initial launch angles

numerically simulate this CME up to 1 AU. We compared our data at 1 AU with the Advanced Composition Explorer (ACE) spacecraft data. We then tried to adjust our CME parameters (initial speed, initial magnetic strength, launch angle) in order to match the ACE data as good as possible. We thus actually use the signals close to the Earth in order to derive the initial characteristics of the CME.

Figure 5 shows the results of our final best fit. The velocity curve at 1 AU is well reproduced in this simulation. The peak in the density, indicating the flux rope passage, is too. However, this is an artifact of the 2.5 D simulation. For the z-component of the magnetic field, the simulated and measured profiles at 1 AU are similar but the magnetic cloud in our simulation seems to arrive a few hours too late. Nevertheless the main features of the CME are surprisingly well mimicked in spite of our simple CME model and in spite of the fact that we our simulations are only 2.5 D. Actually, to predict the intensity of a magnetic storm, the most important parameters are the z-component of the magnetic field and the radial velocity. Our model seems suitable for predicting these parameters. According to our latest simulation results, the CME had a maximum initial magnetic field strength of -2.5 G, an *inverse* magnetic field configuration and an average initial speed of 1772 km/s. This velocity is about 70% higher than the one measured. However, this can be due to the fact that the speed measured by LASCO is the plane-of-sky speed which should be lower than the real CME speed. The difference can amount to a factor of two and more.

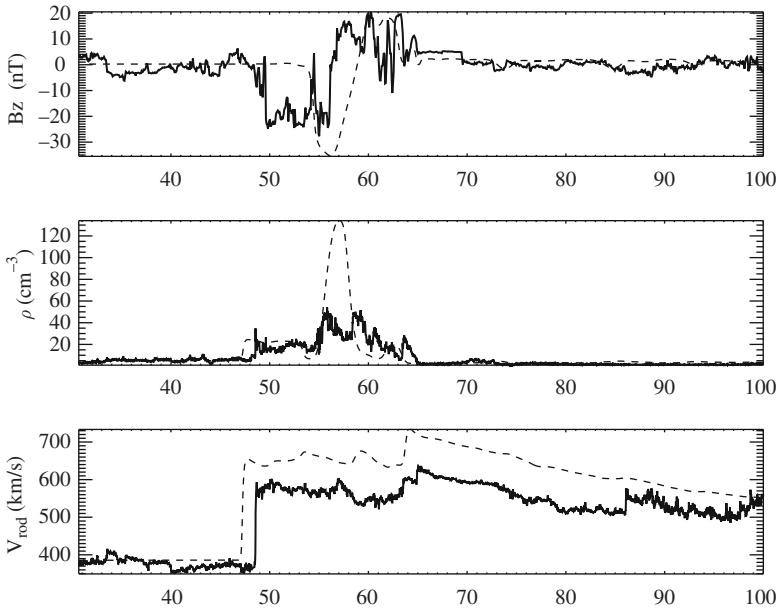


Figure 5. Comparison between ACE *in situ* data (solid curves) and our simulation (dashed curves). On the abcis, the time is given in hours

CONCLUSION

We discussed the state-of-the-art CME initiation and IP evolution models. First, we discussed CME initiation models based on magnetic foot point shearing and magnetic flux emergence. In spite of extensive parameter studies, however, the shearing models are not able to produce the fast CMEs that create shock waves in the IP space and are, therefore, important for space weather. When taking into account the drag of the background wind, flux emergence seems to yield CME velocities up to 800 km/s. For the CME evolution studies, however, we applied the much simpler, so-called 'density-driven', CME model. We then studied the influence of the initial magnetic polarity on the evolution of CMEs into the IP space up to 1 AU. We have shown that the evolution path of the CMEs was strongly related to the initial magnetic polarity. We have also shown that the time of arrival was influenced by the initial magnetic polarity: a *normal* CME propagates faster than an *inverse* CME and thus reached the Earth a few hours earlier. Last but not least, the *inverse* CMEs display a strong southward magnetic field (a well known source of magnetic storms) at 1 AU. The geo-effectiveness of a CME is thus strongly related to the initial magnetic polarity in its flux rope.

ACKNOWLEDGEMENT

These results were obtained in the framework of the projects GOA 2004/01 and OT 02/57 (K.U.Leuven), G.0451.05 (FWO-Vlaanderen) and C90203 (ESA Prodex 8). The results were obtained on the HPC cluster VIC of the K.U.Leuven.

BIBLIOGRAPHY

Antiochos, S., Karpen, J., DeVore, C.: ApJ, 575, 578 (2002)
Amari, T., et al.: ApJ, 595, 1231–1250 (2005)
Chané, E., et al.: AA, 432, 331–339 (2005)
Chané, E., et al.: AA, 447, 727–733 (2006)
Chen, P.F., Shibata, K.: ApJ, 545, 524–531 (2000)
Dubey, G., van der Holst, B., Poedts, S.: Proc. Solar Wind 11/Soho 16, Whistler, Canada, 11–17 June, 2005, ESA SP–592, 637–640 (2005)
Dubey, G., van der Holst, B., Poedts, S.: 2006, A&A, in press.
Gombosi, T.I., et al.: Comp. in Sc.& Eng., 6, No 2, 14–35 (2004)
Jacobs, C., et al.: AA, 430, 1099–1107 (2005)
Jacobs, C., Poedts, S., van der Holst, B.: 2006, AA, in press.
Lee, C.O., et al.: AGU Fall Meeting Abstracts, A1418 (2004)
Lionello, R., Linker, J.A., Mikic, Z.: 2003, AIP Conf. Proc. 679: Solar Wind Ten, 222–225 (2003)
Lionello, R., et al.: ApJ, 625, 463L (2005)
Low, B.C., Zhang, M.: ApJ, 564, L53–L56 (2002)
Lugaz, N., Manchester, W.B., Gombosi, T.I.: ApJ, 627, 1019–1030 (2005)
Manchester, W.B., et al.: ApJ, 610, 588–596 (2004a)
Manchester, W.B., et al.: JGR, 109, A02107 (2004b)
Manchester, W.B., et al.: JGR, 109, A01102 (2004c)
Manchester, W.B., et al.: ApJ, 622, 1225–1239 (2005)
Mikic, Z., Linker, J.A.: ApJ, 430, 898 (1994)

Odstrcil, D., Riley, P., Zhao, X.P.: JGR, 109, A02116 (2004)
Odstrcil, D., Pizzo, V.J., Arge, C.N.: JGR, 110, A02106 (2005)
Riley, P., et al.: Proc. IAU Symposium 226, CUP, 389–402 (2005)
Roussev, I.I., et al.: ApJ, 595, L57–L61 (2003a)
Roussev, I.I., et al.: ApJ, 588, L45–L48 (2003b)
Sokolov, I.V., et al.: ApJ, 616, L171–L174 (2004)
Tóth, G., et al.: in Multiscale Coupling of Sun–Earth Processes, A.T.Y. Lui, Y. Kamide, and G. Consolini, (eds), 383–397, Elsevier (2005)
Wang, A.H., Wu, S.T., Gopalswamy, N.: in Particle Acceleration in Astrophysical Plasmas, Geophysical Monograph Series 156, 185–195, D. Gallagher, J. Horwitz, J. Perez, R. Preece, and J. Quenby (eds) (2005)
Wu, C.C., Wu, S.T., Dryer, M.: SPh, 223, 259–282 (2004)
Wu, C.C., et al.: JGR, 110, doi:10.1029/2005JA011011 (2005a)
Wu, C.C., et al.: SPh, 225, 157–175 (2005b)
Zhukov, A.N.: this book (2006)

CHAPTER 1.5

SIGNATURES OF THE ANCIENT SUN CONSTRAINING THE EARLY EMERGENCE OF LIFE ON EARTH

M. MESSEROTTI[1,2] AND J. CHELA-FLORES[3,4]

[1] *INAF-Trieste Astronomical Observatory, Loc. Basovizza n. 302, 34012 Trieste, Italy*
[2] *Department of Physics, University of Trieste, Via A. Valerio 2, 34127 Trieste, Italy*
[3] *The Abdus Salam ICTP, Strada Costiera 11, 34014 Trieste, Italy*
[4] *Instituto de Estudios Avanzados, IDEA, Caracas 1015A, R.P.Venezuela*

Abstract/ Resume: A factor for understanding the origin and evolution of life on Earth is the evolution of the Sun itself, especially the evolution of space climate and weather. Many aspects of the Sun's history remain to be understood. We reconsider constraints that knowledge of our own star implies for the emergence of life on Earth. This provides further insights into what may happen in other solar systems. Fortunately, particles emitted by the Sun in the past have left a record in geologic samples, but on this bases we cannot exclude earlier dates for the onset of life on Earth. A very early origin of life has to take into account the imprints of solar energetic particles during the first billion years (Gyr) after the formation of the Sun, approximately from 4.6 till 3.6 Gyr before the present (BP). Our review includes the isotopic fractionation of the noble gases, the depletion of volatile elements on the Moon and constraints for the origin of life on Europa, the icy moon of Jupiter

ESTIMATES FOR THE AGE OF THE FIRST APPEARANCE OF LIFE ON EARTH

The rationalization of the lunar cratering record provides some guidance for estimating the possibility of life first arising on Earth. Distinct temporal possibilities for the earliest possible time for the first appearance of life are possible with additional inputs from closely related scientific areas. The lunar record may be supplemented with information retrieved from firstly, biogeochemistry (namely with data related to the fractionation of the stable isotopes of the biogenic elements). Secondly, associated with the earliest fossils of stromatolites our current understanding of micropaleontology leads to further possible constraints on the first

J. Lilensten (ed.), Space Weather, 49–59.

appearance of life on Earth. However, various theories of the evolution of the early Sun will further constrain the origin of the earliest life on Earth.

Lunar Record Constraints (4.4–4.2 Gyr BP)

Although the processes taking place during this period are not represented in the geological record, the current scenario of planetary origin gives us a means of inferring the activity that may have frustrated or encouraged emergent life. During the first 100 million years the flux of impactors would have set up the conditions for the separation of iron and silicate, giving rise to a metallic core. During this formation of the planetary embryo a major impact with another planet-size body would have given rise to the expulsion of a large amount of matter from the embryonic Earth and given rise to the Moon (Canup and Asphaug, 2001). The satellite cooled quickly, but did not form an atmosphere, possibly due to the smaller cross section than the Earth. Another significant effect of the Moon-forming impact was to blow away the original atmosphere that the embryonic Earth had captured from the solar nebula (Kasting and Catling, 2003). The planet was much more dynamic geologically and most of the records of large impacts were deleted, but the same geological activity was most likely responsible for partial out gassing of a secondary atmosphere, the exact nature of which can be inferred from the isotopic composition of the noble gases: It has been shown that comets are capable by themselves of providing noble gases in the correct proportions provided that the laboratory experiments duplicate the conditions for cometary formation (Owen and Bar-Nun, 1995). Besides the temperatures had descended to about 100 °C or below by about 4.4 Gyr BP (Schwartz and Chang, 2002). This scenario for planetary origin allows the possibility of an early origin and evolution of life on Earth. However, it should be remembered that the lunar record demonstrates that some difficulties may arise in this scenario since the Imbrium basin on the Moon was formed by a large impact as late as 3.8 Gyr BP. This implies the persistence of catastrophic impacts for life on Earth, since our planet has a larger effective cross section than our satellite (Sleep et al., 1989).

Biogeochemistry Constraints (3.8–3.9 Gyr BP)

The photosynthesis of prokaryotes includes the stromatolitic-forming cyanobacteria, formerly called blue-green algae. In this process a specific enzyme that leads in several steps to the synthesis of glucose captures carbon dioxide. But the carbon dioxide in the environment and nutrients contain the two stable isotopes of carbon ^{12}C and ^{13}C. The process of photosynthesis favours ^{12}C over ^{13}C. Geologic process partitions the stable isotopes in opposite ways; for instance limestone is depleted in ^{12}C and enriched in ^{13}C. The fossil records of organic matter that have been enriched in ^{12}C can be traced back in sedimentary rocks right back to some of the earliest samples such as the 3,800 Myr-old metamorphosed sedimentary rocks from Isua, West Greenland. These geochemical analyses of the ancient rocks militate in

favour of the presence of bacterial ecosystems in the period that we are discussing in this section, namely 3.8–3.9 Gyr BP (Schidlowski, 1988; Schidlowski, et al., 1983). The question of the metamorphism to which the Isua samples have been subjected has raised some controversy in the past (Hayes et al., 1983).

Stromatolitic Constraints (3.5–3.6 Gyr BP)

Stromatolites consist of laminated columns and domes, essentially layered rocks. Prokaryotic cells called cyanobacteria form them. In addition, they are users of chlorophyll-*a* to capture the light energy that will drive the photosynthetic process. These microorganisms are mat-building communities. At present they are ubiquitous, even in the Dry Valley lakes in Antarctica mat-building communities of cyanobacteria have been well documented (Parker et al., 1982). Right back into ancient times such mats covered some undermat formation of green sulphur and purple bacteria. Such underlying microorganisms are (and were) anaerobes that can actually use the light that impinges on the mat above them by using bacteriochlorophylls that absorb wavelengths of light that pass through the mat above them (Schopf, 1999). Not only has the cyanobacterium spread worldwide, but it has also extraordinary temporal characteristics. Stromatolites have persevered practically without changes for over three billion years.

The exact date for the earliest stromatolitic fossils is at present under discussion (Brasier et al., 2002; Schopf et al., 2002). They have been dated at around 3.5 Gyr BP (Schopf, 1993). Hence, the origin of life if the fossils are accepted, must be in the time interval discussed in this section, or even earlier considering that the cyanobacterium itself is already quite a complex cell.

ISOTOPIC FRACTIONATION OF THE NOBLE GASES ON EARTH

A signature of the early Sun is provided by isotopic fractionation of the five stable noble gas elements, namely, He, Ne, Ar, Kr, and Xe. The early atmosphere arose from collisions during the accretion period, the so-called heavy bombardment of the surface of the Earth. Planetesimal impacts increase the surface temperature affecting the formation of either a proto-atmosphere or a proto-hydrosphere by degassing of volatiles (Matsui and Abe, 1986). This generated a ‘steam atmosphere’. One of its consequences was a rapid hydrodynamic outflow of hydrogen, including some of its compounds such as methane, carrying along heavier gases in its trail (Hunten, 1993). The mechanism postulated is that of aerodynamic drag. The upward drag of noble gas atoms of similar dimension competes with an opposite force due to gravity. Hence, since the various isotopes of these gases have different masses the net result is the occurrence of a mass-dependent fractionation of the various noble gas isotopes. For even heavier atoms, the gravity effect can be stronger than the aerodynamic drag and such atoms would not show the remarkable fractionation typical of the noble gases.

By looking at other main-sequence stars at equivalent early periods of their evolution, we became aware of an associated larger output of solar EUV radiation. With the early Sun such an ultraviolet excess radiation is a possible factor that can trigger the phenomenon of mass fractionation in the noble gases. The case of the $^{22}Ne/^{20}Ne$ ratio is an example, since its value is larger than in the Earth's mantle, or in the solar wind. The observed fractionation of the noble gases can be taken as a signature of two aspects of the early Sun: firstly, the presence of the postulated escape flux, and secondly (more relevant for the main topic of this paper), as evidence for the solar energy source that drives the outward flux of gases. The emergence of appropriate conditions for life on Earth has to wait until the decrease of solar radiation that characterizes the terrestrial accretion period. The beginning of such a favourable period begins once accretion has ended. The surface heat flux diminishes, leading to the steam atmosphere raining into a global ocean (Kasting, 1993). This splitting of a primitive atmosphere into a hydrosphere and a secondary atmosphere leaves behind carbon and nitrogen compounds that will be ingredients for subsequent steps of chemical evolution and, eventually, the dawn of life.

DEPLETION OF VOLATILE ELEMENTS ON THE MOON

The Moon is depleted of volatile elements such as hydrogen, carbon, nitrogen and the noble gases, possibly due to the fact that the most widely accepted theory of its formation is the impact of the Earth by a Mars-sized body during the accretion period. Exceptionally though, volatiles are abundant in lunar soils. The lunar surface evolved during the heavy bombardment period, adding material with a different composition to the Sun, and not derived from the Sun. Ions from the solar wind are directly implanted into the lunar surface (Kerridge, 1975; Kerridge et al., 1991). This component was detected during the Apollo missions. The isotopic composition of the noble gases in lunar soils has been established as being subsequent to the formation of the Moon itself. But nitrogen has a special place in the research for the nature of the astrochemistry of the early solar system. Unlike some of the other biological elements (CHNOPS or carbon, hydrogen, nitrogen, oxygen, phosphorus and sulphur), in lunar soils it is estimated that between 1.5 and 3 Gyr there was an increment of some 50% in the ratio $^{15}N/^{14}N$. This result has been abundantly confirmed. By performing single grain analyses Wieler and co-workers have searched for evidence of a predominantly non-solar origin of nitrogen in the lunar regolith (Wieler et al., 1999). There have also been attempts to analyze trapped N in the lunar regolith (Hashizumi et al., 2000). These works suggest that, on average, some 90% of the N in the grains has *a non-solar source*, contrary to the view that essentially all N in the lunar regolith has been trapped from the solar wind, but this explanation has difficulties accounting for both the abundance of nitrogen and a variation of the order of 30% in the $^{15}N/^{14}N$ ratio. The origin of non-solar component is an open problem. Indeed, Ozima and co-workers propose that most of the N and some of the other volatile elements in lunar soils may actually have come from the Earth's atmosphere rather than the solar wind (Ozima et al., 2005). This

hypothesis is valid provided the escape of atmospheric gases, and implantation into lunar soil grains, occurred at a time when the Earth had essentially no geomagnetic field. This is a valuable approach since it could clearly be tested by examination of lunar far-side soils, which should lack the terrestrial component. This question is not just pertinent to the astrogeological aspects of the evolution of the Moon, but by giving us a solid grasp on the evolution of the early Earth atmosphere, those factors that influenced the conditions favourable to the onset of life on Earth will be clearer. Hopefully with the availability of new missions, such STEREO involving two spacecraft in heliocentric orbit to study coronal mass ejections (CMEs), further measurements of the isotopic N-abundances may contribute to sorting out the astrochemical signatures of the early solar system that are awaiting to be deciphered. Such knowledge of N, one of the most intriguing of the six CHNOPS elements, will be considerable progress in the study of the origin of life on Earth.

PREPARING THE SOLAR SYSTEM FOR THE EMERGENCE OF LIFE

Various processes may have contributed to an early onset of the phenomenon of life, solar activity being one of the most relevant. The more intense solar wind of the early Sun would have a dramatic effect on the possibilities of preparing the Solar System for the emergence of life. In fact, the shock wave of the encounter of the intense solar wind with the spreading accretion disk blows away the residual gas and fine dust still present in the disk. Some evidence for this assertion may be found in meteorites (Bertout et al., 1991). In spite of the fact that the processes taking place from that moment onwards are not represented in the terrestrial geologic record, the current scenario of planetary origin gives us a means of inferring the activity that may have frustrated, or encouraged, the emergence of life. During the first 100 million years the flux of impactors would have set up the conditions for the separation of iron and silicate, giving rise to a metallic core. During this formation of the planetary embryo a major impact with another planet-size body gave rise to the expulsion of a large amount of matter from the primitive Earth, giving rise to the Moon. Our satellite cooled quickly, but it did not form an atmosphere. This may have been due to the smaller lunar cross section compared to the Earth. The original atmosphere that the Earth had captured from the solar nebula must have been largely blown away by the intense solar wind of the T-Tauri phase of the solar evolution. The planet was much more dynamic geologically and most of the records of large impacts were deleted, but the same geological activity was most likely responsible for partial out gassing of a secondary atmosphere, the exact nature of which can be inferred from the isotopic composition of the noble gases. It has been shown that comets are capable by themselves of providing noble gases in the correct proportions. This remark has been confirmed by laboratory experiments duplicating the conditions for cometary formation (Owen and Bar-Nun, 1995). Temperatures had descended to about 100 °C after the end of accretion at 4.4 Gyr BP. This scenario for planetary origin allows, in principle, the possibility of an early origin and evolution of life on Earth, provided that the solar climate

and solar weather were sufficiently clement. However, it should be remembered that the lunar record demonstrates that some difficulties may arise in this scenario, since the Imbrium basin on the Moon, for instance, was formed by a large impact as late as 3.8–3.9 Gyr BP (Hartmann et al., 2000). This was a real cataclysmic spike in the cratering record. This event is known as the Late Heavy Bombardment (LHB). This implies the persistence of catastrophic impacts for the emergence of life on Earth, since our planet has a larger effective cross section than our satellite (Sleep et al., 1989). Recent discussions of the origin and intensity of the late heavy bombardment is further supported by more recent work (Gomes et al., 2005) that suggests that the LHB was triggered by the rapid migration of the giant planets. This phenomenon produced major changes in the space weather conditions. But even more, it triggered a massive delivery of planetesimals into the inner Solar System. Those conditions were an impediment for the emergence of life. Alternatively, if life had emerged before the LHB, it would most likely have been annihilated and started again after the major perturbations of the LHB had faded out. The analogous problem of bombardment of terrestrial-like planets in extra-solar systems is the subject of further recent attention (Levison et al., 2003).

EFFECTS OF RADIATION

As astrobiology studies the origin, evolution, distribution and destiny of life in the universe, in the present section we shall discuss in turn the four stages at which solar and extra-solar physics have a frontier in common with astrobiology.

Solar Radiation as a Factor in the Origin of Life

The incidence of non-ionizing UVR on the early surface of Earth and Mars to a large extent can be inferred from observations. Ionizing radiation, mainly due to nuclear and atomic reactions is relevant: X-rays are emitted spanning the whole spectrum of hard X-rays to soft X-rays (0.01–10 nm); gamma rays are present too. The primary components affecting space climate are: galactic cosmic rays and solar cosmic radiation. To these we should add events that contribute to solar weather, such as solar particle radiation consisting of the low-energy solar-wind particles, as well as more energetic solar burst events consisting of solar particles that arise from magnetically disturbed regions of the Sun. These events vary in frequency according to the 11-year cycle. However, scenarios for an early onset of life that have been proposed in the past have to deal with space weather that was radically different in the early Sun. Knowledge of the prehistory of solar particle radiation can be approached with a combined effort from observations of present-day emissions, together with studies of energetic solar particles recorded in extraterrestrial materials, notably the Moon material that became available with the Apollo missions, as well as with the study of meteorites. First of all we consider the magnitude of the ionizing radiation that may have been present at the time when life emerged on Earth, during the Archean (3.8–2.5 Gyr BP). According to

some theoretical arguments (Mojzsis et al., 1999), the origin of life may be traced back even earlier, during the Hadean (4.6–3.8 Gyr BP). Indeed, these authors argue that the simplest interpretation of carbon isotopic data may point to the presence of diverse photosynthesizing, methanogenic, and methylotrophic bacteria on Earth before 3.85 Gyr BP. Isotopic and geologic evidence suggest that in the Archean the atmosphere was anoxic (Walker, et al., 1983). As a result the abundance of ozone would not have acted as a UV defense mechanism for the potential emergence of life. UVB (280–315 nm) radiation as well as UVC (190–280 nm) radiation could have penetrated to the Earth's surface with their associated biological consequences (Margulis et al., 1976; Cockell, 1998).

Extra-solar Radiation in the Evolution of Life

Gamma ray bursts are powerful explosions that are known to originate in distant galaxies, and a large percentage likely arises from explosions of stars over 15 times more massive than our Sun. A burst creates two oppositely directed beams of gamma rays that race off into space. The Swift mission, launched in November 2004, contributes to determine recent burst rates. Such data allows the evaluation of life's robustness during the Ordovician (510–438 million years ago). During this geologic period there was a mass extinction of a large number of species (440–450 million years ago). This was the second most devastating extinction in Earth history. Present evidence has led to the conjecture that the extinction was triggered by a gamma ray burst (Thomas et al., 2005). There is no direct evidence that such a burst activated the ancient extinction. The conjecture is based on atmospheric modelling. The main conclusion to be derived from these calculations is that gamma ray radiation from a relatively nearby star explosion, hitting the Earth for only 10 seconds, could deplete up to half of the atmosphere's protective ozone layer. Recovery could take at least 5 years. With the ozone layer damaged, UVR from the Sun could kill much of the life on land and near the surface of oceans and lakes, and disrupt the food chain.

Solar Radiation in the Distribution of Life

To illustrate further the necessity for a comprehensive approach to influence of space weather on the origin of life, we should consider the underlying presence of variable, and to some extent, incompletely known output of solar radiation during its first Gyr. Experiments have been performed in the recent past at the ISS. We should keep in mind that during the early life of the Sun, the UV flux was much higher than it is today. The relevant wavelength regions are the XUV and soft X rays. These wavelengths are absorbed at the top of the atmosphere. Research at the ISS has been supplemented with laboratory tests. Several problems related with early biological evolution have been discussed in the past under the simulation of the early solar radiation environment (Lammer et al., 2002). Work on space weather influence on biological systems include the implications for the biosphere of magnetic field

reversals (Biernat et al., 2002); the influence on biological systems of solar flares (Belisheva et al., 2002), and some work on uracil dosimetry to estimate the possible preservation of the molecules of life (Bérces et al., 2002). If the distribution of life in the solar system took place by transfer of microorganisms, knowledge of solar weather is needed for the early stages of its evolution, to have some constraints on the possible transfer of microorganisms, as investigated extensively by Horneck and co-workers (Horneck and Cockell, 2001 for references). *Bacillus subtilis* is a Gram-positive harmless bacterium. It is capable of producing endospores resistant to adverse environmental conditions such as heat and desiccation and is widely used for the production of enzymes and specialty chemicals. The inactivation of *B. subtilis* spores has been studied in the Earth's orbit under different simulated ozone-column abundances to provide quantitative estimates of the potential photobiological effects of such an early ozone-free atmosphere (Horneck and Cockell, 2001). These authors find that the spectral sensitivity of DNA increases sharply toward shorter wavelengths from the UVB to UVC region. They conclude that this is the primary reason for the observed high lethality of extraterrestrial UV radiation that could provide a barrier to the distribution of life in the solar system. However, it should be kept in mind that the most radiation resistant organism known at present exhibits a remarkable capacity to resist the lethal effects of ionizing radiation. The specific microorganism is a non-spore forming extremophile found in a small family known as the Deinococcaceae. In fact, *Deinococcus radiodurans* (whose name comes from the Greek for "terrible berry that withstands radiation") is a Gram-positive, red-pigmented, non-motile bacterium. It is resistant to ionizing and UV radiation. Several authors have studied these (Battista, 1997, Daly et al., 2004 and Levin-Zaidman et al., 2003). Members of this Family can grow under chronic radiation [50 grays (Gy) per hour] or recover from acute doses of gamma radiation greater than 10,000 Gy without loss of viability. Survivors are often found in cultures exposed up to 20,000 Gy. Seven species make up this Family, but it is *D. radiodurans,* whose radio-resistance appears to be the result of an evolutionary process that selected for organisms that could tolerate massive DNA damage. For the sake of comparison, the bacterium *E. coli* is approximately 200 times less resistant to gamma radiation, whereas humans cannot tolerate radiation of up to 5 Gy. Independent of the various UV defense mechanisms discussed in Sec. 2, the surface of the Earth is largely protected from cosmic radiation by the atmosphere itself. The annual dose of cosmic radiation for Germany is 0.3 mGy/year at sea level and 25 mGy/year at an altitude of 15 Km (Baumstark-Khan and Facius, 2001). Besides, also for comparison, we know that the survival fraction for mammalian cells in radiotherapy becomes negligible for a dose of 500 Gy (Kassis and Adelstein, 2004). It appears that the capacity of extremophiles to withstand ionizing radiation is due to adaptation to desiccation, as both environmental challenges (lack of water and excessive radiation doses) lead to similar massive DNA repair mechanisms. In this context, cyanobacteria have extraordinary ability to withstand desiccation and then rapidly absorb water when it becomes available. For example, a cyanobacterial population in gypsum quickly regains its ability to photosynthesize after addition of water (Van Thielen and

Garbary, 1999). Possibly the radiation resistance ability of *D. radiodurans* may be due to its genome. (It assumes an unusual toroidal morphology that may contribute to its radio-resistance.)

Solar Radiation as a Factor in the Destiny of Life

The question of solar radiation also has a frontier with the fourth aspect of astrobiology. In about 4–5 billion years the brightness of the Sun will increase and its radius will increase (Sackmann, et al., 1993). The consequence of the Sun abandoning its present steady state will lead to a swelling of its outer atmosphere. At the same time while the radius is increasing helium atoms will be at such a temperature that fusion into beryllium and carbon will occur. This is known in nuclear physics as the triple alpha point. This process lasts a few seconds. The energy from this 'helium flash' will lead to a sequence of events that will largely increase the emission of solar wind, carrying away a large fraction of the solar mass. This stage is well known to us. Indeed, there are many known examples of the stellar mass that the increased solar wind will take away (this is the planetary nebula stage). These events will set definite constraints of the destiny of life in our own solar system. Our knowledge of other stars mapped on a Hertzsprung-Russell diagram gives us enough confidence with the later stages of the evolution of our own Sun as it leaves the Main Sequence. These phenomena set strong constraints to the destiny of life in the solar system. But some further work on the models of the sun is necessary before making definite predictions on the period following the departure from the Main Sequence.

DISCUSSION AND CONCLUDING REMARKS

The main thesis that we have maintained in this work is that solar activity, space weather and astrobiology should be brought within a unified framework. This approach naturally leads us to the suggestion of exploiting instrumentation from somewhat dissimilar sciences (astronomy and astrobiology) with a unified objective. We have attempted a preliminary comprehensive discussion of how research in the conditions of the early Sun combine with observations in several disciplines to give us insights into the factors that lead to the emergence of life in a given solar system (biogeochemistry, lunar science, micropaleontology and chemical evolution). These considerations are necessary to approach the conditions that will allow life to emerge in a given solar system anywhere in the universe.

ACKNOWLEDGEMENTS

This work has been carried out in the framework of COST Action 724 on the development of the scientific basis of Space Weather.

REFERENCES

Battista, J.R.: Against all odds: The survival strategies of *Deinococcus radiodurans.* Ann. Rev. Microbiol. **51**, 203–224 (1997)

Baumstark-Khan, C., Facius, R.: Life under conditions of ionizing radiation, in: Astrobiology the Quest for the Conditions of Life, G. Horneck and C. Baumstark-Khan (eds), Springer, Berlin, pp. 261–284 (2001)

Belisheva, N.K., Semenov, V.S., Tolstyh, Yu. V., Biernat, H.K.: Solar Flares, Generation of Solar Cosmic Rays, and Their Influence on Biological Systems, ESA SP 518, 429–430 (2002)

Bérces, A., Kovács, G., Kerekgyarto, Rontó, Gy., Lammer, H., Kargel, G. Kömle, N.I.: Uracil Dosimetry in Simulated Extraterrestrial Condition, ESA SP 518, 431–432 (2002)

Bertout, C., Basri, G., Cabrit, S.: The Classical T-Tauri Stars: Future Solar Systems? in: *The Sun in Time*, C.P. Sonett, M.S. Giampappa and M.S. Matthews (eds), The University of Arizona, Tucston, pp. 683–709 (1991)

Biernat, H.K., Erkaev, N.V., Penz, T., Lammer, H., Manrubia, S.C., Selsis, F., Vogl., Muhlbachler, S.: Magnetic Field Reversals on Earth: Possible Implications for the Biosphere, ESA SP 518, 433–434 (2002)

Brasier, M.D., Green, O.W., Jephcoat, A.P., Kleppe, A.K., Van Kranendonk, M.J., Lindsay, J.F., Steele, A. and Grassineau, N.V.: Questioning the evidence for Earth's oldest fossils, Nature, **416**, 76–81 (2002)

Canup, R.M., Asphaug, E.: Origin of the Moon in a giant impact near the end of the Earth's formation. Nature, **412**, 708–712 (2001)

Cockell, C.S.: Biological Effects of High Ultraviolet Radiation on Early Earth – A Theoretical Evaluation, J. Theor. Biol. **193**, 717–729 (1998)

Daly, M.J., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., Venkateswaran, A., Hess, M., Omelchenko, M.V., Kostandarithes, H.M., Makarova, K.S., Wackett, L.P., Fredrickso, J.K., Ghosal, D.: Accumulation of Mn (II) in *Deinococcus radiodurans* Facilitates Gamma-Radiation Resistance, Science, **306**, 1025–1028 (2004)

Gomes, R., Levison, H.F., Tsiganis, K. and Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets, Nature **435**, 466–469 (2005)

Hartmann, W.K., Ryder, G., Dones, L., Grinspoon, D.: The time-dependent intense bombardment of the primordial Earth/Moon system, in: Origin of Earth and Moon, R.M. Canup and K. Righter (eds), University of Arizona Press, Tucson, pp. 493–512 (2000)

Hashizumi, K., Chaussidon, M., Marty, B., Robert, F.: Solar Wind Record on the Moon: Deciphering Presolar from Planetary Nitrogen. Science, **290**, 1142–1145 (2000)

Hayes, J.M., Kaplan, I.R., Wadeking, K.W.: Precambrian organic geochemistry, preservation of the record, in: Schopf, J.W. (ed.), Earth Earliest Biosphere, Princeton University Press, Princeton, pp. 93–134 (1983)

Horneck, G., Cockell, C.S.: The History of the UV Radiation Climate of the Earth—Theoretical and Space-based Observations, Photochemistry and Photobiology **73**, 447–451 (2001)

Hunten, D.M.: Atmospheric Evolution of the Terrestrial Planets, Science, **259**, 915–920 (1993)

Kassis, A.I., Adelstein, S.J.: Radiobiologic Principles in Radionuclide Therapy, The Journal of Nuclear Medicine, **45**, 1–4 (2004)

Kasting, J.F.: Earth's Early Atmosphere, Science, **259**, 920–926 (1993)

Kasting, J.F., Catling, D.C.: Evolution of a habitable planet, Annual Reviews of Astronomy and Astrophysics, **41**, 429–463 (2003)

Kerridge, J.F.: Solar nitrogen: evidence for a secular increase in the ratio of nitrogen–15 to nitrogen–14, Science, 1975, 162–164 (1975)

Kerridge, J.F., Signer, P., Wieler, R., Becker, R.H., Pepin, R.O.: Long term changes in composition of solar particles implanted in extraterrestrial materials, in: The Sun in Time, C.P. Sonett, M.S. Giampappa and M.S. Matthews (eds), The University of Arizona, Tucston, pp. 389–412 (1991)

Lammer, H., Hickel, A., Tehrany, M.G., Hanslmeier, A., Ribas, I., Guinan, E.F.: Simulating the Early Solar Radiation Environment: X-Ray Radiation Damage Experiments, ESA SP 518, 469–470 (2002)

Levin-Zaidman, S., Englander, J., Shimoni, E., Sharma, A.K., Minton, K.W., Minsky, A.: Ringlike structure of the *Deinococcus radiodurans* genome: a key to radioresistance? Science, **299**, 254–256 (2003)

Levison, H.F., Agnor, C.: The Role of Giant Planets in Terrestrial Planet Formation, Ap. J., **125**, 2692–2713 (2003)

Margulis, L., Walker, J.C.G., Rambler, M.: Re-assessment of the roles of oxygen and ultraviolet light in Precambrian evolution, Nature, **264**, 620–624 (1976)

Matsui, T., Abe, Y.: Evolution of an impact-induced atmosphere and magma ocean on the accreting Earth. Nature, **319**, 303–305 (1986)

Mojzsis, S.J., Krishnamurthy, R., Arrhenius, G.: Before RNA and after — Geophysical and geochemical constraints on molecular evolution, in Gesteland, R., et al., eds., RNA world, 2nd ed.: Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, pp. 1–49 (1999)

Owen, T., Bar-Nun, A.: Comets, impacts and atmospheres, Icarus, **116**, 215–226 (1995)

Ozima, M., Seki, K., Terada, N., Miura, Y.N., Podosek, F.A., Shinagawa, H.: Terrestrial nitrogen and noble gases in lunar soils, Nature, **436**, 655–659 (2005)

Parker, B., Simmons, Jr., G., Wharton, Jr., R. Seaburg, K.G., Love, F.: Gordon Removal of organic and inorganic matter from Antarctic lakes by aerial escape of bluegreen algal mats, J. Phycol. 18, 72–78 (1982)

Sackmann, I.-J., Boothroyd, A.I., Kraemer, K.E.: Our Sun. III. Present and Future, Astrophysical Journal, **418**, 457–468 (1993)

Schidlowski, M.: A 3.800-million-year isotopic record of life from carbon in sedimentary rocks, Nature, 333, 313–318 (1988)

Schidlowski, M., Hayes, J.M., Kaplan, I. R.: Isotopic Inferences of Ancient Biochemistries: Carbon, Sulfur, Hydrogen, and Nitrogen, in Earth's Earliest Biosphere its Origin and Evolution, J. William Schopf (ed.), Princeton University Press, Princeton, New Jersey, pp. 149–186 (1983)

Schopf, J.W.: Microfossils of the Earth Archaean Apex Chert: New evidence of the antiquity of life, Science, **260**, 640–646 (1993)

Schopf, J.W.: Cradle of Life: The Discovery of Earth's Earliest Fossils, Princeton University Press, Princeton, New Jersey, pp. 186–190 (1999)

Schopf, J.W., Kudryavtsev, A.B., Agresti, D.G., Wdowiak, T.J., Czaja, A.D.: Laser-Raman imagery of Earth's earliest fossils, Nature, **416**, 73–76 (2002)

Schwartz, A.W. and Chang, S. From Big Bang to Primordial Planet-Setting the Stage for the Origin of Life, in Life's Origin, J.W. Schopf (ed.), University of California Press, Berkeley, pp. 78–112 (2002)

Sleep, N., Zahnle, K., Kasting, J.F. Morowitz, H.J.: Annihilation of ecosystems by large asteroid impacts on the early Earth, Nature, **342**, 139–142 (1989)

Thomas, C.H., Jackman, A.L., Melott, C.M., Laird, R.S., Stolarski, N., Gehrels, J.K., Cannizzo, Hogan, D. P.: Astrophysical Journal Letters. 622, L153 (2005)

Van Thielen, N., Garbary, D.J.: Life in the rocks-endolithic algae, in: Enigmatic microorganisms and life in extreme environmental habitats. J. Seckbach (ed.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 245–253 (1999)

Walker, J.C.G., Klein, C., Schidlowski, M., Schopf, J.W., Stevenson, D.J., Walter, M.R.: Environmental evolution of the Archean-Proterozoic Earth, in: Earth's Earliest Biosphere (J.W. Schopf, ed.), Princeton University Press, Princeton, pp. 160–190 (1983)

Wieler, R., Humbert, F., Marty, B.: Evidence for a predominantly non-solar origin of nitrogen in the lunar regolith revealed by single grain analyses, Earth and Planetary Science Letters, **167**, 47–60 (1999)

CHAPTER 2.0

THE SUN'S INTERACTION WITH THE EARTH'S THERMOSPHERE AND CLIMATE SYSTEM

A.S. RODGER
British Antarctic Survey, Madingley Road, Cambridge CB3 0ET, UK

There are many processes associated with the Sun, the thermosphere and the Earth's climate that are well understood, but the prediction of these systems is still far from accurate. There are several reasons for this. Some important physical, chemical or biological mechanisms may not be sufficiently well quantified, or indeed not yet determined. Complex interactions and feedback mechanisms operate over wide spatial and temporal scales, and make prediction of the emergent behaviour particularly challenging.

The mean radiative forcing of the climate system is an example where there are still some important uncertainties (Fig. 1, from Intergovernmental Panel on Climate Change, 2001). This shows that two of the largest uncertainties are terms associated with the Sun and with aerosols. Whilst much is known about climate forcing by visible radiation, Marsh and Haigh (both this volume) give succinct summaries of several alternative mechanisms by which variations in the electromagnetic and charged particle output of the Sun can affect the climate system.

There is $\sim 7\%$ variation in the flux of ultraviolet radiation at wavelengths $\sim$200 nm through a solar cycle, and this causes marked changes in the stratosphere, and especially of the ozone there. Quantitative modelling (Haigh, this volume) demonstrates that the changes induced above the tropopause influence tropospheric climate, primarily through changes in the propagation and dissipation of gravity waves and planetary waves. This is an important finding as it provides a robust mechanism whereby changes occurring at high altitude can affect the climate at the Earth's surface.

Figure 1 shows that the effects of aerosols are the largest uncertainty in radiative forcing. Marsh (this volume) describes how the role of cosmic rays can be important in a chain of reactions whereby aerosols are formed and cloud cover may be altered. Changing cloud cover can either have a positive or a negative effect on

J. Lilensten (ed.), Space Weather, 61–64.

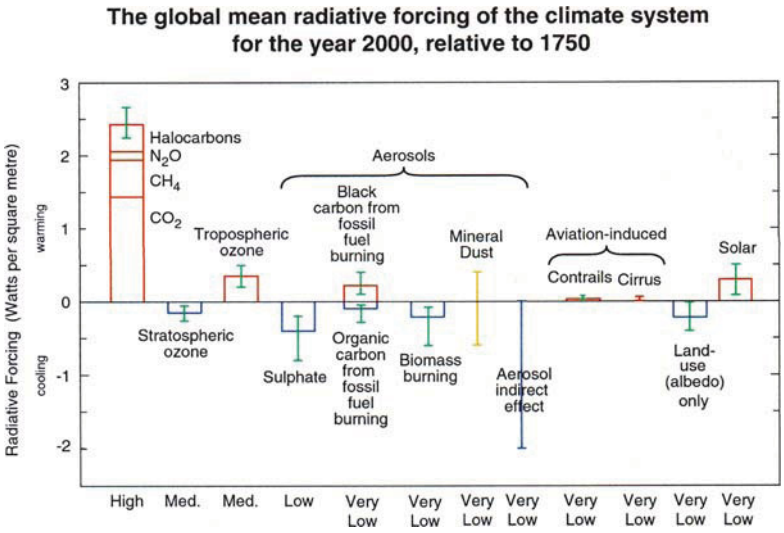


Figure 1. Many external factors force climate change. These radiative forcings arise from changes in the atmospheric composition, alteration of surface reflectance by land use, and variation in the output of the Sun. Except for solar variation, some form of human activity is linked to each. The rectangular blocks represent estimates of the contributions of these forcings – some of which yield warming, and some cooling (IPCC 2001). In many cases the uncertainties are appreciable

the surface air temperature depending upon the altitude at which this occurs. New laboratory–based experiments are now being undertaken to quantify these effects.

Even with a good understanding of most of the physical laws it may not be possible to produce accurate predictions. One reason is that the spatial and temporal variations in the energy inputs cannot be measured with sufficient resolution. Figure 2 provides a summary of the spatial and temporal domains that are important for understanding the structure and dynamics of the thermosphere. As an example, a modest rotation of the interplanetary magnetic field can alter the rate of energy transfer from the solar wind to the magnetosphere by two orders of magnitude in a few minutes, as reconnection at the magnetopause changes. However, the details of where reconnection occurs and what controls its rate remain elusive.

Most of the power transferred via reconnection eventually ends up in the thermosphere. During substorms, $\sim 10^{11}$ W is deposited in the nightside thermosphere causing winds to change both in speed and direction. During geomagnetically quiet times, the largest power input ($\sim 10^{10}$ W) is through the effects of tides propagating up from the lower atmosphere. Both increases and decreases in the flux of charged particles trapped in the van Allen radiation belts can exceed several orders of magnitude through the course of a storm. In each of these examples, exactly where and when energy is deposited is not yet predictable.

Whilst in some systems it is safe to ignore low amplitude, short duration events, it is not safe to do in the upper atmosphere. There is no doubt that the ionosphere-

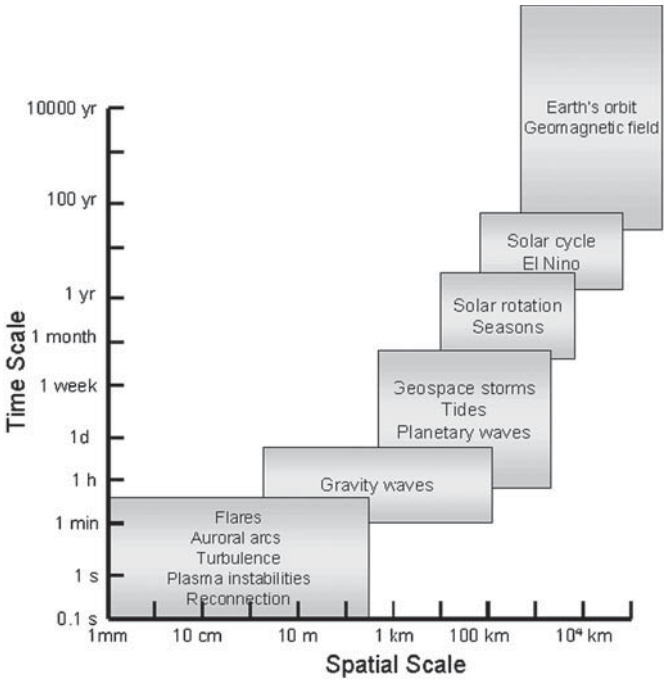


Figure 2. Internal and external processes that affect the chemistry and dynamics of the thermosphere expressed in spatial and temporal domains

thermosphere signature of substorms is associated with the equator-most auroral arc, a feature that is only $\sim$1 km wide. Small-scale features in the electric field can lead to significant localised Joule heating that is particularly prevalent in the thermospheric footprint of the magnetospheric cusp, and in the vicinity of auroral arcs where there is complex feedback between the electric field strength and ionospheric conductivity.

The processes summarised above change the temperature and winds of the thermosphere, and hence affect the drag acting on satellites in low Earth orbit. Also the neutral wind plays a critical role in determining where F–region ionospheric scintillations are observed, as the growth rate of the instability mechanism depends upon the plasma motion in the rest frame of the neutral particles. The neutral winds are often ignored when considering the dynamics of the upper atmosphere, partly because they are very difficult to measure but to do so may lead to significant errors as neutral winds can reach 1000 ms^{-1}. Thus there is considerable complexity in trying to predict the global thermosphere winds and temperatures, and hence determine satellite drag. Doornbos (this volume) shows that, by using measurements of the drag acting on the satellites themselves, much more accurate predictions are possible, compared with empirical approaches that have been used for over a decade.

As Fig. 2 shows, there are changes occurring on all timescales and hence attribution of the cause of change can be particularly difficult. This is further complicated if data series are broken, or indeed start and stop at different epochs of a cycle. Ulich et al. (this volume) provide a comprehensive review of the difficulties in determining the secular changes of the ionosphere-thermosphere system.

Modelling change on timescales from minutes to centuries is something that the meteorological community has been undertaking for a few decades, with models of ever–greater complexity. For weather forecasting, data assimilation is increasingly important. Keil (this volume) gives an overview and some practical examples of how the space weather community can benefit from these developments in meteorology. Other approaches such as neutral networks may be useful.

To predict the two rather loosely related topics of the thermospheric effects of space weather and the impacts of solar variations on the Earth's climate system, several fundamental scientific questions need to be addressed. For the Sun–climate area of science, the tentative mechanisms already identified need to be quantified, as does the way in which solar-induced changes in the thermosphere, mesosphere and stratosphere affect the climate of the tropopause. In the thermosphere, insufficient knowledge of the spatial and temporal distribution of energy deposition is the most limiting factor in accurate prediction for space weather purposes. Also both areas of science require long-term measurements of the incoming solar radiation as a function of wavelength.

REFERENCES

Doornbos, E.: Thermosphere density model calibration. (this volume)

Haigh, J.: Solar variability and climate. (this volume)

Intergovernmental Panel on Climate Change, Climate Change: The Scientific Basis, Cambridge University Press, UK (2001)

Keil, M.: Numerical space weather prediction: can meteorologists forecast the way ahead? (this volume)

Marsh, N.D.: Influence of solar activity cycles on Earth's climate (this volume)

Ulich, T., Clilverd, M.A., Jarvis, M.J., Rishbeth, H.: Unravelling signs of global change in the ionosphere (this volume)

CHAPTER 2.1

SOLAR VARIABILITY AND CLIMATE

JOANNA D. HAIGH
Blackett Laboratory, Imperial College London, UK

Abstract: Solar radiation is the fundamental energy source for the atmosphere and the global average equilibrium temperature of the Earth is determined by a balance between the energy acquired by the solar radiation absorbed and the energy lost to space by the emission of heat radiation. The interaction of this radiation with the climate system is complex but it is clear that any change in total solar irradiance (TSI) has the potential to influence climate. In the past, although many papers were written on relationships between sunspot numbers and the weather, the topic of solar influences on climate was often disregarded by meteorologists. This was due to a combination of factors of which the key was the lack of any robust measurements indicating that solar radiation did indeed vary. There was also mistrust of the statistical validity of the evidence and, importantly, no established scientific mechanisms whereby the apparent changes in the Sun might induce detectable signals near the Earth's surface. Another influence was a desire by the meteorological profession to distance itself from the Astrometeorology movement popular in the 19th century (Anderson 1999). Nowadays, with improved measurements of solar and climate parameters, evidence for an influence of solar variability on the climate of the lower atmosphere has emerged from the noise. This article provides a brief review of the observational evidence and an outline of the mechanisms whereby rather small changes in solar radiation may induce detectable signals near the Earth's surface[1]

SOLAR INFLUENCES ON THE EARTH'S LOWER ATMOSPHERE

Measurements and Reconstructions

Assessment of climate variability and climate change depends crucially on the existence and accuracy of records of meteorological parameters. Ideally records would consist of long time series of measurements made by well-calibrated instruments

[1]It is not possible to review here all potential mechanisms for solar-climate links. What is presented offers, necessarily, a personal perspective but, of the areas that are not covered, two may be pertinent: the effects of solar energetic particles on stratospheric composition (see e.g. Jackman et al. 2005) and the possible influence of galactic cosmic rays on clouds through ionisation processes (see Marsh, this volume).

J. Lilensten (ed.), Space Weather, 65–81.

located with high density across the globe. In practice, of course, this ideal cannot be met. Measurements with global coverage have only been made since the start of the satellite era about 25 years ago. Instrumental records have been kept over the past few centuries at a few locations in Europe. For longer periods, and in remote regions, records have to be reconstructed from indirect indicators of climate known as proxy data.

Proxy data provide information about weather conditions at a particular location through records of a physical, biological or chemical response to these conditions. Some proxy datasets provide information dating back hundreds of thousands of years which make them particularly suitable for analysing long term climate variations and their correlation with solar activity. One well established technique for providing proxy climate data is dendrochronology, or the study of climate changes by comparing the successive annual growth rings of trees (living or dead). Much longer records of temperature have been derived from analysis of oxygen isotopes in ice cores obtained from Greenland and Antarctica and evidence of very long term temperature variations can also be obtained from ocean sediments.

Figure 1 presents reconstructions of the Northern Hemisphere surface temperature record produced using a variety of proxy datasets. There are some large differences between them, especially in long-term variability, but there is general agreement that current temperatures are higher than they have been for at least the past 2 millennia. Other climate records suggesting that the climate has been changing over the past century include the retreat of mountain glaciers, sea level rise, thinner Arctic ice sheets and an increased frequency of extreme precipitation events. A key

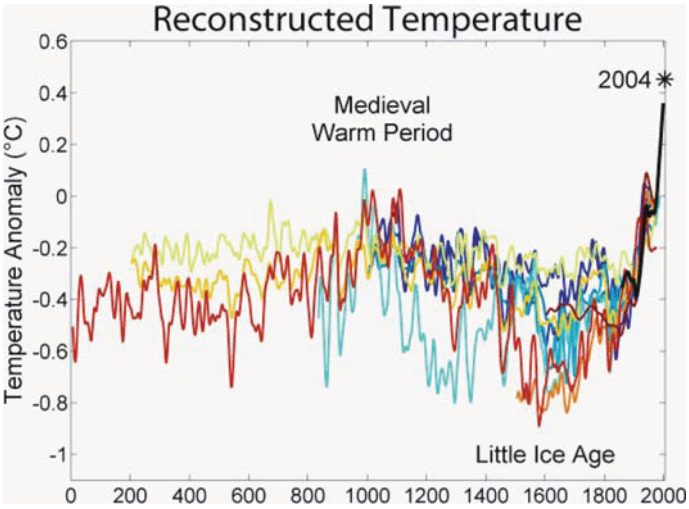


Figure 1. Northern Hemisphere surface temperature (5-year running-mean) over the past 2 millennia. The thick black curve (1856-present) is from measurements; the other curves are reconstructions by various authors based on proxy data (figure from http://www.globalwarmingart.com/wiki/Image:2000_Year_Temperature_Comparison.png)

concern of contemporary climate science is to attribute cause(s) to these changes, including the contribution of solar variability.

Solar Signals in Climate Records

Many different approaches have been adopted in the attempt to identify solar signals in climate records. Probably the simplest has been spectral analysis, in which cycles of 11 (or 22 or 90, etc.) years are assumed to be associated with the Sun. In another approach time series of observational data are correlated with time series of solar activity. This can be developed to extract the response in the measured parameter to a chosen solar activity forcing factor. A further sophistication allows a multiple regression, in which the responses to other factors are simultaneously extracted along with the solar influence. Each of these approaches gives more certainty than the previous one that the signal extracted is actually due to the Sun and not to some other factor, or to random fluctuations in the climate system, but it should be remembered that such detection is based only on statistics and not on any understanding of how the presumed solar influence takes place.

Millennial, and longer, timescales

Ocean sediments have been used to reveal a history of temperature in the North Atlantic by analysis of the minerals believed to have been deposited by drift ice (Bond et al. 2001). In colder climates the rafted ice propagates further south where it melts, depositing the minerals. These materials also preserve information on cosmic ray flux, and thus solar activity, in isotopes such as ^{10}Be and ^{14}C. Thus simultaneous records of climate and solar activity may be retrieved. An example is given in Fig. 2 which shows fluctuations on the 1,000 year timescale well correlated between the two records, suggesting a long-term solar influence on climate.

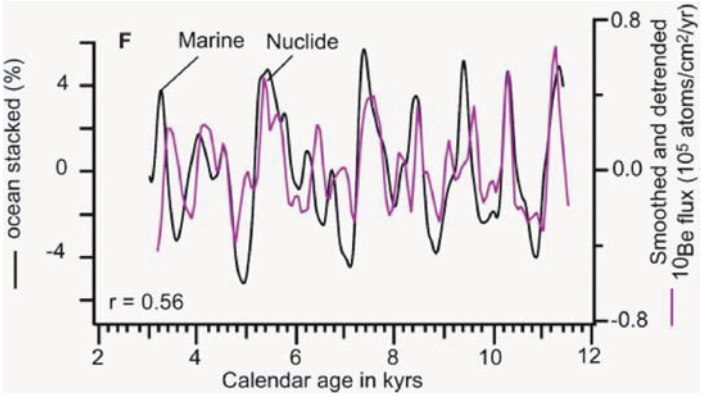


Figure 2. Records of ^{10}Be (lighter curve) and ice-rafted minerals (darker curve) extracted from ocean sediments in the North Atlantic. (Bond et al. 2001)

On even longer timescales the amount of solar radiation received by the Earth is modulated by variations in its orbit around the Sun. The distance between the two bodies varies during the year due to the ellipticity of the orbit which varies with periods of around 100,000 and 413,000 years due to the gravitational influence of the Moon and other planets. At any particular point on the Earth the amount of radiation striking the top of the atmosphere also depends on the tilt of the Earth's axis to the plane of its orbit, which varies cyclically with a period of about 41,000 years, and on the precession of the Earth's axis which varies with periods of about 19,000 and 23,000 years. Averaged over the globe the solar energy flux at the Earth depends only on the ellipticity but seasonal and geographical variations of the irradiance depend on the tilt and precession. These are important because the intensity of radiation received at high latitudes in summer determines whether the winter growth of the ice cap will recede or whether the climate will be precipitated into an ice age. Thus changes in seasonal irradiance can lead to much longer-term shifts in climatic regime. Cyclical variations in climate records with periods of around 192341100 and 413 kyr are generally referred to as Milankovitch cycles after the geophysicist who made the first detailed investigation of solar-climate links related to orbital variations.

Century scale

On somewhat shorter timescales it has frequently been remarked that the Maunder Minimum in sunspot numbers in the second half of the 17th century coincided with what has become known as the "Little Ice Age" during which western Europe experienced significantly cooler temperatures. Fig. 3 shows this in terms of winter temperatures measured in London and Paris compared with the ^{14}C ratio found in tree rings over the same period. Similar cooling has not, however, been found

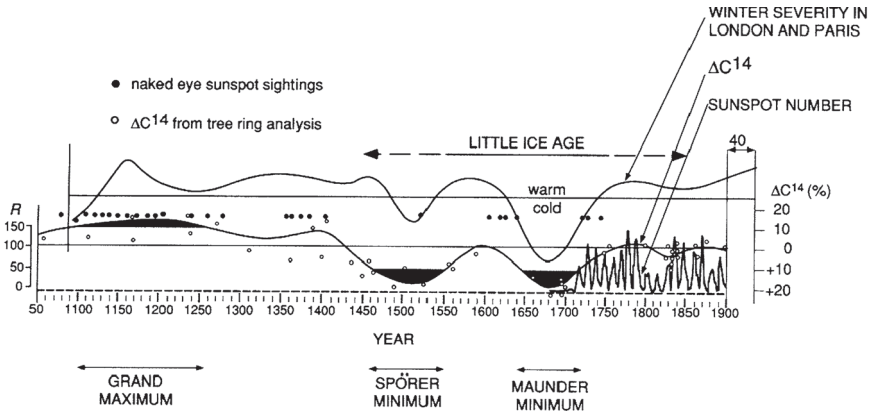


Figure 3. From a paper by Eddy (1976) suggesting that winter temperatures in NW Europe are correlated with solar activity. Note the coincidence of the "Little Ice Age" with the Maunder Minimum in sunspot number

in temperature records for the same period across the globe so attribution of the European anomalies to solar variability may be unwarranted.

A paper published by Friis-Christensen and Lassen (1991) caused considerable interest when it appeared to show that temperature variations over the observational period could largely be ascribed to solar variability. The measure of solar activity used was the length of the solar cycle (SCL) and, as can be seen in Figure 4 (top), this value appeared to coincide almost exactly with the Northern Hemisphere land surface temperature record. This result has been challenged, however, by Laut and Gundermann (2000) who show that the extrapolation used to complete recent cycle lengths was flawed. Their version (extended back to 1550) is shown in Fig. 4 (bottom) and the correspondence between the two records in the 20th century now appears much less marked.

Studies of the attribution of causes to recent climate change (see below) are now able to extract a solar signal in centennial scale climate records but uncertainties still remain in the absolute magnitude of the solar effect.

Solar cycle

Many studies have purported to show variations in meteorological parameters in phase with the "11-year" solar cycle. Some of these are statistically not robust and some show signals that appear over a certain interval of time only to disappear, or even reverse, over another interval. There is, however, considerable evidence that solar variability on decadal timescales does influence climate. An example is shown in Fig. 5 which presents the mean summer time temperature of the upper troposphere (between about 2.5 and 10 km) averaged over the whole northern hemisphere. This parameter varies in phase with the solar 10.7 cm index with an amplitude of 0.2–0.4 K.

The hemispheric average, however, hides the fact that the solar signal is not uniformly distributed. Solar signals detected in sea surface temperatures (SSTs) by White et al. (1997) show that SSTs do not increase uniformly in response to enhanced solar activity: indeed, the pattern shows latitudinal bands of warming and cooling. They also show that the amplitude of the change is larger than would be predicted by radiative considerations alone, given the known variations in TSI over the same period.

A similar pattern at the surface is shown for the solar signal in the results of a multiple regression analysis of NCEP/NCAR Reanalaysis zonal mean temperatures (Haigh 2003). In this work data for 1978–2002 were analysed simultaneously for ten signals: a linear trend, El Niño-Southern Oscillation (ENSO), North Atlantic Oscillation (NAO), solar activity, stratospheric aerosol from volcanic eruptions, Quasi-Biennial Oscillation (QBO) and the amplitude and phase of the annual and semi-annual cycles. The patterns of response for each signal are statistically significant and separable from the other patterns. The solar response shows largest warming in the stratosphere and bands of warming, of >0.4 K, throughout the troposphere in mid-latitudes.

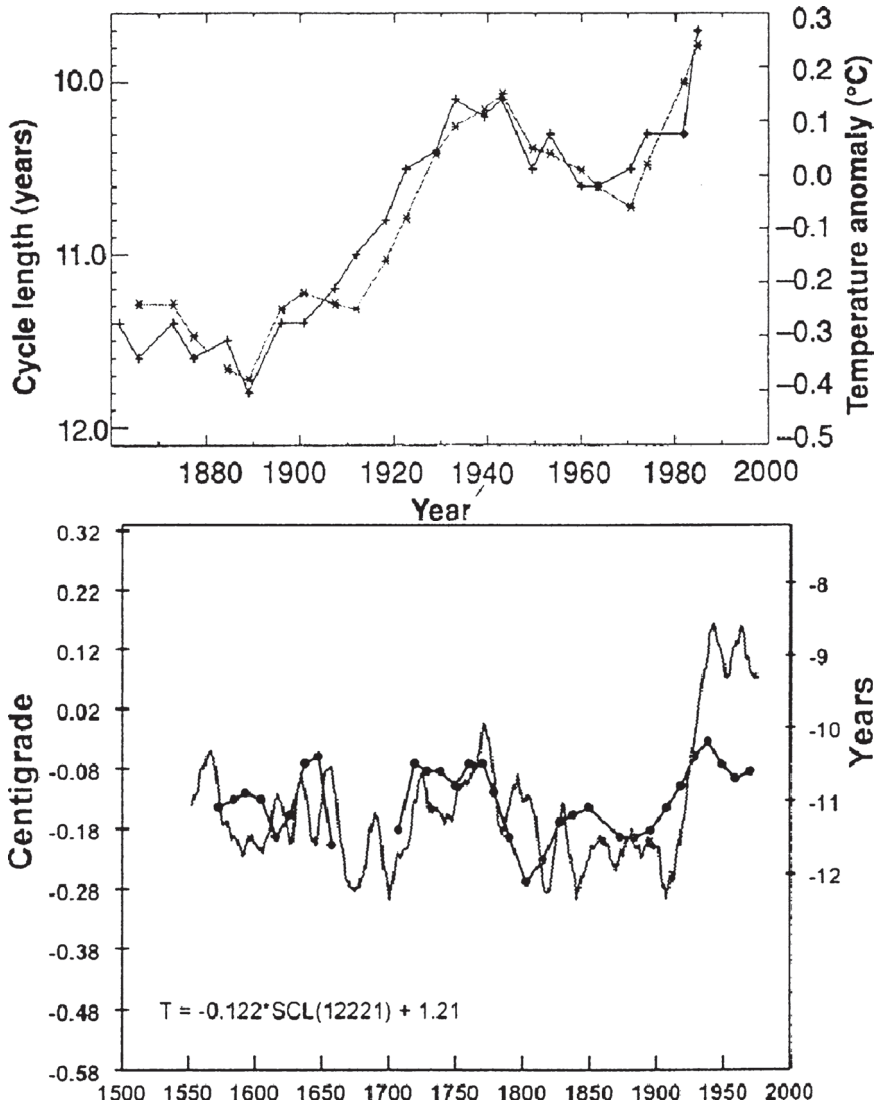


Figure 4. Top: Records of northern hemisphere land temperature (stars) and length of the solar cycle (inverted, pluses) (Friis-Christensen and Lassen 1991). Bottom: Time series of smoothed solar cycle lengths (darker curve) as fitted by Laut and Gundermann (2000) to the Mann et al. (1999) Northern Hemisphere temperature record (lighter curve)

Shown in Fig. 6 are some results from a similar multiple regression analysis of zonal mean zonal winds (Haigh et al. 2005). These show that the effect of increasing solar activity is to weaken the westerly jets and to move them slightly polewards.

Other evidence for the influence of solar cycle variability on climate, specifically surface temperatures and cloud, is discussed by Marsh (this volume).

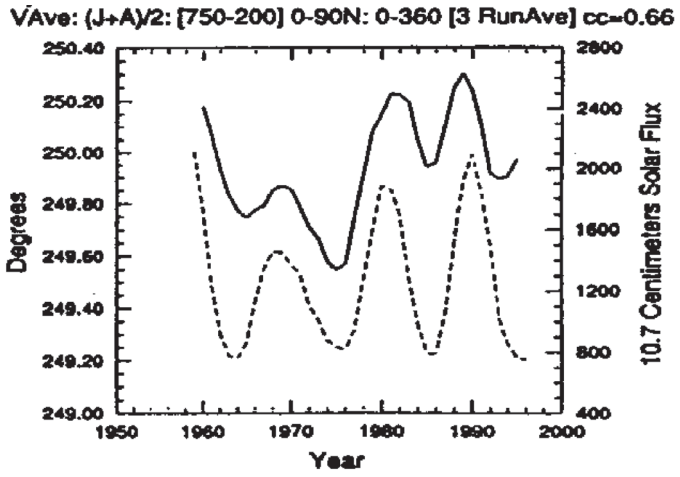


Figure 5. Time series of the mean temperature of the 750–200 hPa layer for the whole northern hemisphere in summer (solid line) and the solar 10.7 cm flux (dashed line). (van Loon and Shea 2000)

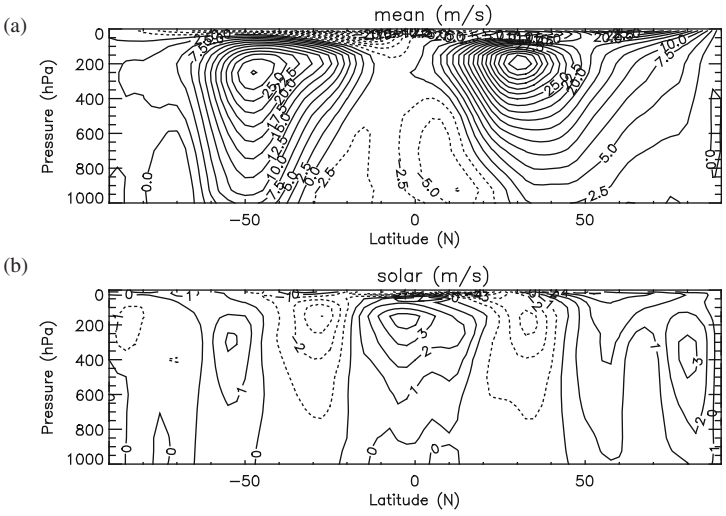


Figure 6. Zonal mean zonal wind (positive values indicate westerlies – i.e. winds from the west) **(a)** December, January, February (DJF) mean of NCEP reanalysis data 1979–2002; **(b)** Solar signal in DJF from multiple regression analysis of NCEP data. (Haigh et al. 2005)

Shorter timescales

Various studies have suggested that the atmosphere responds to solar activity effects on timescales much shorter than the solar cycle. A response to the solar 27-day rotation has been clearly observed in middle atmosphere composition and

temperature (e.g. Hood and Zhou 1999). On even shorter timescales correlations have been observed between decreases in GCRs associated with solar coronal mass ejections and the areas of cyclonic storms (Tinsley 2000).

VARIATIONS IN SOLAR IRRADIANCE AND MECHANISMS FOR INFLUENCE ON CLIMATE

Total Solar Irradiance and Radiative Forcing of Climate Change

Direct measurements of TSI made outside the Earth's atmosphere began with the launch of satellite instruments in 1978. Previous surface-based measurements did not provide sufficient accuracy, as they were subject to uncertainties and fluctuations in atmospheric absorption that may have swamped the small solar variability signal.

Figure 7(a) presents all existing satellite measurements of TSI and it is clear that significant uncertainties remain related to the calibration of the instruments and their degradation over time. For example, data from the newest instrument, the Total Irradiance Monitor (TIM) on the SORCE satellite, is giving values approximately $5\,\mathrm{Wm}^{-2}$ lower than other contemporaneous instruments which disagree among themselves by a few Wm^{-2}. This uncertainty is a serious problem underlying current solar-climate research. The *variation* in TSI over the past two 11-year cycles is known to greater accuracy showing approximately 0.08% ($\sim 1.1\,\mathrm{Wm}^{-2}$) variation.

There is a related uncertainty, however, in the existence of any underlying trend in TSI over the past 2 cycles. Figure 7(b) presents one attempt to composite the measurements into a best estimate. It shows essentially no difference in TSI values between the cycle minima occurring in 1986 and 1996. The results of Willson (2003), however, show an increase in irradiance of 0.045% between these dates. The discrepancy hinges on assumptions made concerning the degradations of the Nimbus7/ERB and ERBS/ERBE instruments, data from which fills the interval, from July 1989 to October 1991, between observations made by the ACRIM I and II instruments. If such a trend were maintained, it would imply an increase in radiative forcing of about $0.1\,\mathrm{Wm}^{-2}$ per decade. Compared in terms of climate forcing, this is appreciable, being about one-third that due to the increase in concentrations of greenhouse gases averaged over the past 50 years. These discrepancies are important because all the available TSI reconstructions discussed below either use directly, or are scaled to fit, the recent satellite measurements, using one or other of the TSI composites.

The time interval of satellite observations contains information only on the short term components of solar variability but, in order to assess the potential influence of the Sun on centennial-scale climate change, it is necessary to know TSI further back into the past. In reconstructing past changes in TSI, proxy indicators of solar variability, for which longer periods of observation are available, are used to produce an estimate of its temporal variation over the past centuries. There are several different approaches taken to "reconstructing" the TSI, all employing a substantial degree of empiricism, and some examples are given in Fig. 8. It is clear that the available estimates diverge as they go back in time.

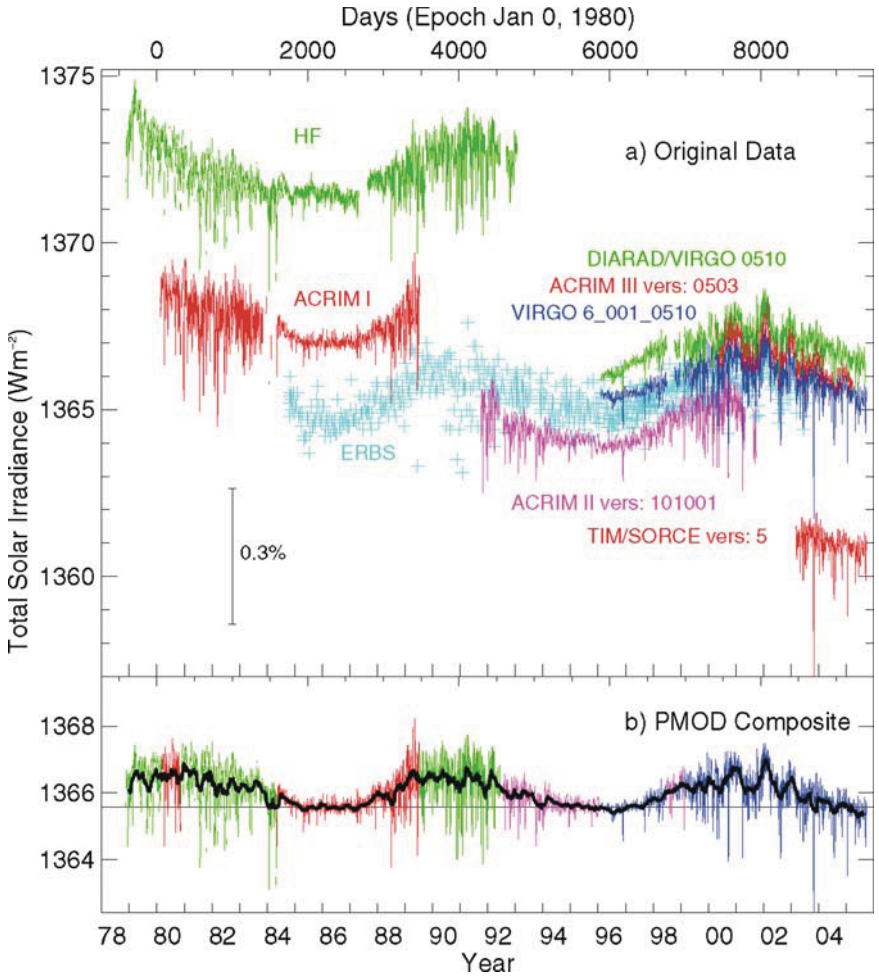


Figure 7. **(a)** Daily-averaged total solar irradiance: all measurements made from satellites **(b)** Composite of measurements to produce best estimate of TSI (figure courtesy of Claus Fröhlich, http://www.pmodwrc.ch)

Top-of-atmosphere (TOA) solar radiative forcing (RF) may be deduced from anomalies in total solar irradiance. It is, however, necessary to scale TSI by a factor of 0.18 to take account of global averaging and global albedo. Thus a 1.7 Wm^{-2} increase in TSI since 1750 translates into a TOA RF of 0.3 Wm^{-2}, as shown in the IPCC (2001) radiative forcing bar chart (see Fig. 1 of Rodger, this volume). We note in passing that the solar value included in this well-publicised figure might be considerably larger or smaller if a different year were assumed for the pre-industrial start-date.

Simulations with computer models of the global circulation of the atmosphere (GCMs) have been carried out to represent the climate from ∼1860 to 2000 with

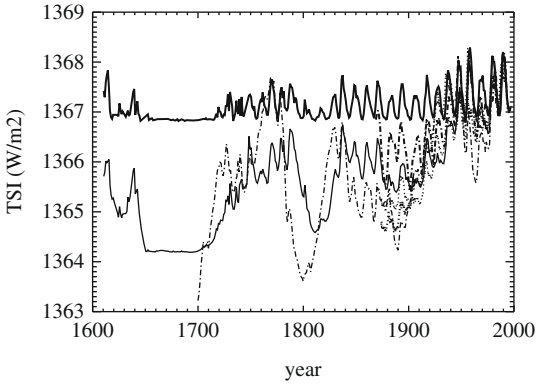


Figure 8. Sunspot numbers (dark solid curve at top, arbitrary scale) and TSI reconstructions (all other curves) by various authors

time-evolving natural (solar and volcanic) and anthropogenic (greenhouse gases, sulphate aerosol) forcings. The GCMs are generally able to reproduce, within the bounds of observational uncertainty and natural variability, the temporal variation of global average surface temperature over the 20th century with the best match to observations obtained when all the above forcings are included. Separation of the effects of natural and anthropogenic forcing suggests that the solar contribution is particularly significant to the observed warming over the period 1900–1940. However, uncertainties remain with the solar effect, particularly regarding the impact of the choice of solar reconstruction. The amplitude of the early 20th century warming depends on the choice of TSI reconstruction used in the GCM. Rind (2000) suggests, given some anthropogenic warming and large natural variability, that solar forcing is not necessarily involved but the analyses made by Stott et al. (2000) and Meehl et al. (2003) show that it is detected.

Going back over the past millennium the study of Crowley (2000), using a 1-D Energy Balance Model (EBM) shows that natural forcings can explain the amplitude of natural variability in northern hemisphere surface temperature over the pre-industrial period. In particular he points out that the higher levels of volcanism prevalent during the 17th century, as well as lower solar irradiance, may contribute to the cooler northern hemisphere experienced at that time. This period is sometimes referred to as "The Little Ice Age" although how global it was is contentious.

Historically many authors have suggested, based on analyses of observational data, that the solar influence on climate is larger than would be anticipated based on radiative forcing arguments alone. The problems with these studies (apart from any question concerning the statistical robustness of their conclusions) is that (i) they frequently apply only in certain locations, and (ii) they do not offer any advances in understanding of how the supposed amplification takes place. Recently some interesting developments have been made to address these

problems based on two different approaches: the first uses detection/attribution techniques to compare model simulations with observations. These studies (North and Wu 2001; Stott et al. 2003) show that the amplitude of the solar response, derived from multiple regression analysis of the data using model-derived signal patterns and noise estimates, is larger than predicted by the model simulations (by up to a factor 4). This technique does not offer any physical insight but suggests the existence of deficiencies in the models and also shows how the solar signal may be spatially distributed. The second approach proposes feedbacks in the climate system (that may already exist in GCMs) and uses these to explain features found both in model results and in observational data. The proposed mechanisms are generally concerned with radiative and thermodynamic processes, water vapour feedback and clouds. Mechanisms involving stratosphere-troposphere dynamical coupling are discussed below. Another suggestion (Meehl et al. 2003) is that the solar influence might be larger where there is less cloud (so that more radiation is absorbed at the surface) and that this would lead to changes in circulation associated with anomalies in horizontal temperature gradient. These GCM results, however, have yet to be confirmed by observations or other model simulations.

Solar Spectral Irradiance and Photochemical Effects in the Stratosphere

Response of stratospheric ozone to solar UV variability

Ozone is produced by short wavelength solar ultraviolet radiation and destroyed by radiation at somewhat longer wavelengths. Images of the Sun acquired in the visible and ultraviolet show that the amplitude of solar cycle variability is greater in the far ultraviolet (see Fig. 9). This means that ozone production is more strongly modulated by solar activity than its destruction and this leads to a higher net production of stratospheric ozone during periods of higher solar activity.

Both observational records and model calculations show approximately 2% higher values in ozone columns at 11-year solar cycle maximum relative to minimum. However, there are some discrepancies between satellite observations and model predictions in the vertical and latitudinal distributions of the response. Figure 10 (right) shows a typical model calculation with the largest changes in the middle stratosphere and less above and below, while the panel on the left shows the solar cycle modulation of ozone derived from a satellite dataset.

The discrepancies between the signals derived from observations and models is greatest in equatorial regions. The models show a maximum in the middle stratosphere 35–40 km) while the observational datasets suggest something rather different: a larger response near the stratopause and possibly a second maximum in the lower stratosphere. This remains a key area of uncertainty in solar effects on the atmosphere. There may be some factors missing in the models and their poor simulation of the lower stratospheric response suggests that this might be related to ozone transport. However, full 3D GCMs produce very similar profiles to those of

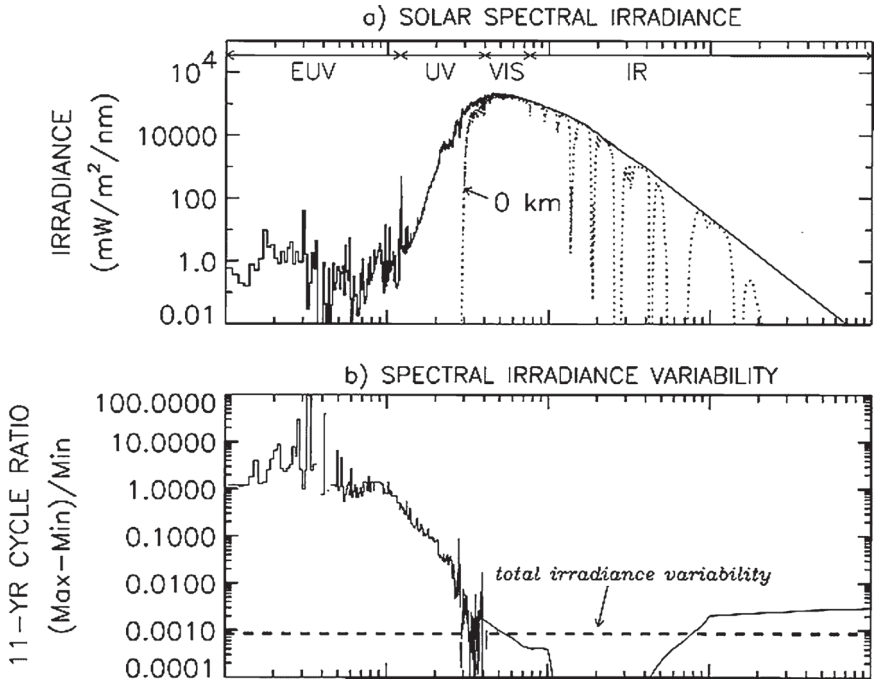


Figure 9. Top: solar spectrum. Bottom: Fractional difference in solar spectral irradiance between maximum and minimum of the 11-year cycle. Lean (1998)

the 2D models with less complete dynamics so the mechanisms involved are not clear. It should also be borne in mind that the observational data are only available over less than two solar cycles so there remains some doubt about the statistical robustness of the signals derived from them.

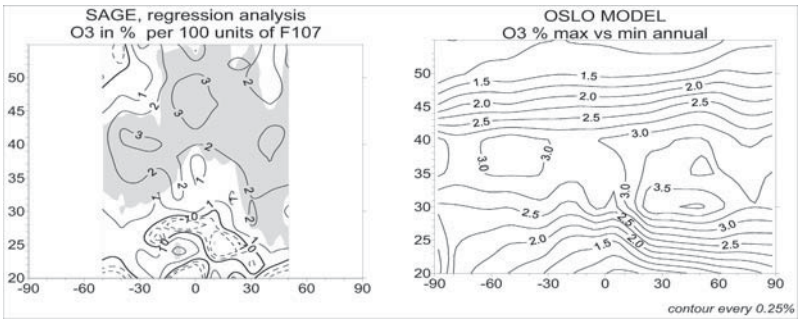


Figure 10. Percentage increase in zonal mean O_3 concentration (solar minimum to maximum) as a function of latitude and height. Left: estimated from SAGE data (shaded areas statistically significant at the 95% level). Right: estimated by 2D model. For further details of both observational and model studies see Haigh et al. (2004)

Response of Stratospheric Temperature to Solar UV Variability

The response of atmospheric temperatures to solar variability is large in the upper atmosphere, with, for example, variations of 400 K being typical at 300 km over the 11-year cycle, reflecting the large modulation of far and extreme ultraviolet radiation in that region. At lower altitudes the response is smaller, and less certain. Measurements made from satellites suggest an increase of up to about 1 K in the upper stratosphere at solar maximum; a minimum, or possibly even a negative change, in the mid-low stratosphere with another maximum, of a few tenths of a degree, below. However, precise values, as well as the position (or existence) of the negative layer, vary between datasets. GCM simulations of the solar influence on the temperature of the middle atmosphere are fairly successful in reproducing the magnitude of warming in the tropical upper stratosphere but less so in the lower stratosphere where they fail to simulate the minimum seen in the data. Models which incorporate an interactive photochemical scheme, and so predict ozone as well as temperature, do not at present seem to be any more successful than those with prescribed ozone changes.

Vertical Coupling Through the Middle and Lower Atmosphere

Northern hemisphere winter polar stratosphere

During the winter the high latitude stratosphere becomes very cold and a polar vortex of strong westerly winds is established. The date in spring when this vortex finally breaks down is very variable, particularly in the northern hemisphere, but plays a key role in the global circulation of the middle atmosphere. Because variations in solar UV input change the latitudinal temperature gradient in the upper stratosphere, the evolution of the winter polar vortex may be affected. Satellite data suggest that the vortex strengthens in November and December in response to solar activity. This positive perturbation to zonal mean zonal winds then propagates polewards and downwards, until by February it is replaced by an easterly anomaly (Kodera et al. 1990; Kodera 1995). However, the picture is complicated by the apparent modulation of the solar signal by the QBO in tropical stratospheric zonal winds (Labitzke 1987; Gray et al. 2004).

Planetary-scale waves induce a large-scale meridional circulation (Haynes et al. 1991) that is strengthened when the winter polar vortex is more disturbed. The meridional circulation is therefore a prime route for winter polar events to influence the lower stratosphere, not only in polar latitudes but throughout the winter hemisphere and even the equatorial and summer subtropical latitudes, through a modulation of the strength of equatorial upwelling. A dynamical feedback via the meridional circulation would serve to amplify the direct TSI and indirect UV solar signal since the less disturbed early winter conditions in solar maximum lead to a weaker meridional circulation, weaker equatorial upwelling and hence a warmer equatorial lower stratosphere at solar maximum than solar minimum. Recent advances in modelling these phenomena have been made (Matthes et al. 2004) but there are still several aspects that are not adequately reproduced or understood.

Stratosphere-troposphere coupling

The solar effects seen in the NCEP temperature and wind data (Fig. 6) have been reproduced by GCM simulations of the effects of increased solar UV variability (Haigh 1996, 1999; Larkin et al. 2000; Matthes et al. 2004): see Fig. 11. These model studies also predict a weakening and expansion of the tropical Hadley cells in response to solar activity. Recent analysis of NCEP vertical velocity data has confirmed that this effect is present in the real atmosphere (Gleisner and Thejll 2003).

The similarity of the signals found in the observational data and model runs is intriguing and, by the nature of the model experiments, suggests that changes in the stratosphere, introduced by modulation of solar ultraviolet radiation and ozone, are key. It does not, however, explain the mechanisms whereby such a change in tropospheric circulation is brought about by thermal perturbations to the stratosphere. Haigh et al. (2005) have carried out some experiments with a simplified GCM designed to elucidate some of these mechanisms. In these runs changes in radiative heating are imposed only in the stratosphere but a response is found extending throughout the troposphere. Figure 12 shows the response in Northern hemisphere zonal mean zonal wind found in two such experiments: U5, in which a 5K heating is imposed throughout the stratosphere, and E5, in which a stratospheric heating of 5K at the equator decreases by $\cos^2$(latitude) to zero at the poles.

From these experiments it was concluded that imposed changes in the lower stratospheric temperature forcing lead to coherent changes in the latitudinal location and width of the mid-latitude jetstream and its associated storm-track, and that wave/mean-flow feedbacks are crucial to these changes. Imposed stratospheric warming, and an associated lowering of the tropopause, weakens the jet and

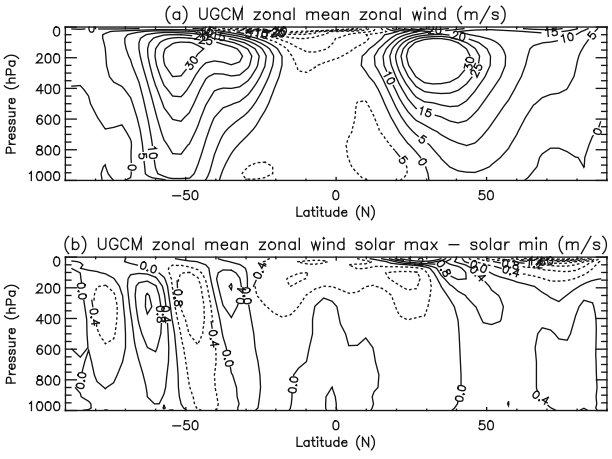


Figure 11. Zonal mean zonal wind in January from a GCM study **(a)** Mean, **(b)** Signal induced by solar cycle variation in UV (Haigh 1996, 1999)

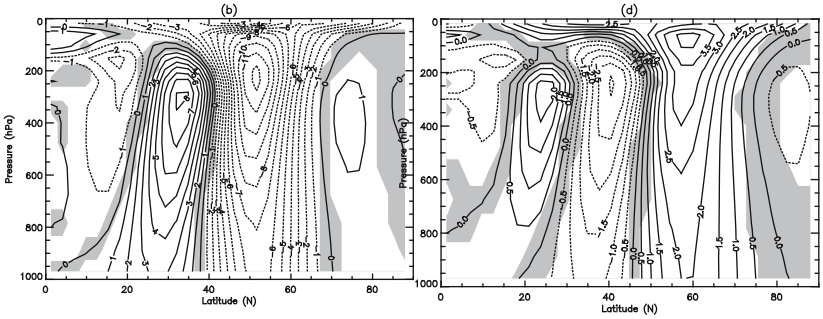


Figure 12. Difference from control run of zonal mean zonal winds in (Left) run U5 (contour interval $1\,\mathrm{ms}^{-1}$) and (Right) E5 ($0.5\,\mathrm{ms}^{-1}$). Average of 2 hemispheres. Regions in which the signal does not reach the 95% confidence level are shaded. From Haigh et al. (2005)

storm-track eddies; equatorial stratospheric warming displaces the jet polewards while uniform warming displaces it markedly equatorwards. It appears that the observed climate response to solar variability is brought about by a dynamical response in the troposphere to heating predominantly in the stratosphere. The effect is small, and frequently masked by other factors, but not negligible in the context of the detection and attribution of climate change. The results also suggest that, at the Earth's surface, the climatic effects of solar variability will be most easily detected in the sub-tropics and mid-latitudes.

Details of the mechanisms involved in the transfer of the effects of the stratospheric forcing to the troposphere are still not clear but further such investigations may provide the key to unraveling the mechanisms of how a solar (or indeed any other) influence in the stratosphere may influence tropospheric climate.

SUMMARY

Radiation from the Sun ultimately provides the only energy source for the Earth's atmosphere and thus changes in solar activity clearly have the potential to affect climate. There is now statistical evidence for solar influence on various meteorological parameters on a wide range of timescales, although extracting the signal from the noise in a naturally highly variable system remains a key problem. Changes in solar irradiance undoubtedly impact the Earth's energy balance, thermal structure and composition but in a complex and non-linear fashion and questions remain concerning the detailed mechanisms which determine to what extent, where and when these impacts are felt. Advances in understanding are being made through the use of global climate models and it is only by further investigation of the complex interactions between radiative, chemical and dynamical processes in the atmosphere that these difficult questions will be answered.

REFERENCES

Anderson, K.: The weather prophets: science and reputation in Victorian meteorology. Hist Sci, **37**, 179–216 (1999)

Bond, G., Kromer, B., Beer, J., Muscheler, R., Evans, M.N., Showers, W., Hoffmann, S., Lotti-Bond, R., Hajdas, I., Bonani, G.: Persistent solar influence on north Atlantic climate during the Holocene. Science, **294**, 2130–2136 (2001)

Crowley, T.J.: Causes of climate change over the past 1000 years. Science, **289**, 270–277 (2000)

Eddy, J.A.: The Maunder Minimum. Science, **192**, 1189–1202 (1976)

Friis-Christensen, E., Lassen, K.: Length Of The Solar-Cycle – An Indicator Of Solar-Activity Closely Associated With Climate. Science, **254**, 698–700 (1991)

Gleisner, H., Thejll, P.: Patterns of tropospheric response to solar variability. Geophys Res Lett 30:art no 17129 (2003)

Gray, L.J., Crooks, S.A., Pascoe, C., Sparrow, S.: Solar and QBO influences on the timings of stratospheric sudden warmings. J Atmos Sci, **61**, 2777–2796 (2004)

Haigh, J.D.: The impact of solar variability on climate. Science, **272**, 981–984 (1996)

Haigh, J.D.: A GCM study of climate change in response to the 11-year solar cycle. Quart J Roy Meteorol Soc, **125**, 871–892 (1999)

Haigh, J.D., Blackburn, M., Day, R.: The response of tropospheric circulation to perturbations in lower stratospheric temperature. J Clim, **18**, 3672–3691 (2005)

Haigh, J.D.: The effects of solar variability on the Earth's climate. Phil Trans Roy Soc A, **361**, 95–111 (2003)

Haigh, J.D., Austin, J., Butchart, N., Chanin, M.-L., Crooks, S., Gray, L.J., Halenka, T., Hampson, J., Hood, L.L., Isaksen, I.S.A., Keckhut, P., Labitzke, K., Langematz, U., Matthes, K., Palmer, M., Rognerud, B., Tourpali, K., Zerefos, C.: Solar variability and climate: selected results from the SOLICE project. SPARC Newsletter, **23**, 19–29 (2004)

Haynes, P.H., Marks, C.J., McIntyre, M.E., Shepherd, T.G., Shine, K.P.: On the 'downward control' of extratropical diabatic circulations by eddy-induced mean zonal forces. J Atmos Sci, **48**, 651–678 (1991)

Hood, L.L., Zhou, S.: Stratospheric effects of 27-day solar ultraviolet variations: The column ozone response and comparisons of solar cycles 21 and 22. J geophys Res, **104**, 26473–26479 (1999)

IPCC: Climate Change 2001: The Scientific Basis. CUP (2001)

Jackman, C.H., DeLand, M.T., Labow, G.J., Fleming, E.L., Weisenstein, D.K., Ko, M.K.W., Sinnhuber, M., Anderson, J., Russell, J.M.: The influence of the several very large solar proton events in years 2000–2003 on the neutral middle atmosphere. Adv. Space Res. **35**, 445–450 (2005)

Kodera, K.: On The Origin And Nature Of The Interannual Variability Of The Winter Stratospheric Circulation In The Northern-Hemisphere. J Geophys Res, **100**, 14077–14087 (1995)

Kodera, K., Yamazaki, K., Chiba, M., Shibata, K.: Downward propagation of upper stratospheric mean zonal wind perturbation to the troposphere. Geophys Res Lett, **17**, 1263–1266 (1990)

Labitzke, K.: Sunspots, the QBO and the stratospheric temperature in the north polar region. Geophys Res Lett, **14**, 535–537 (1987)

Larkin, A., Haigh, J.D., Djavidnia, S.: The effect of solar UV irradiance variations on the Earth's atmosphere. Space Sci Rev, **94**, 199–214 (2000)

Laut, P., Gundermann, J.: Solar cycle lengths and climate: A reference revisited. J Geophys Res, **105**, 27489–27492 (2000)

Lean, J., Rind, D.: Climate forcing by changing solar radiation. J Clim, **11**, 3069–3094 (1998)

Mann, M.E., Bradley, R.S., Hughes, M.K.: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. Geophys Res Lett, **26**, 759–762 (1999)

Matthes, K., Langematz, U., Gray, L.J., Kodera, K., Labitzke, K.: Improved 11-year solar signal in the Freie Universitat Berlin climate middle atmosphere model (FUB-CMAM). J Geophys Res, doi:10.1029/2003/D004012 (2004)

Meehl, G.A., Washington, W.M., Wigley, T.M.L., Arblaster, J.M., Dai, A.: Solar and greenhouse forcing and climate response in the twentieth century. J Clim, **16**, 426–444 (2003)

North, G.R., Wu, Q.: Detecting climate signals using space-time EOFs. J Clim, **14**, 1839–1863 (2001)
Rind, D.: Relating paleoclimate data and past temperature gradients: Some suggestive rules. Quat Sci Rev, **19**, 381–390 (2000)
Stott, P.A., Jones, G.S., Mitchell, J.F.B.: Do models underestimate the solar contribution to recent climate change? J Clim, **16**, 4079–4093 (2003)
Stott, P.A., Tett, S.F.B., Jones, G.S., Allen, M.R., Mitchell, J.F.B., Jenkins, G.J.: External control of 20th century temperature by natural and anthropogenic forcings. Science, **290**, 2133–2137 (2000)
Tinsley, B.A.: Influence of the solar wind on the global electric circuit and inferred effects on cloud microphysics, temperature and dynamics in the troposphere. Space Sci Rev, **94**, 231–258 (2000)
van Loon, H., Shea, D.J.: A probable signal of the 11-year solar cycle in the troposphere of the northern hemisphere. Geophys Res Lett, **26**, 2893–2896 (1999)
White, W.B., Lean, J., Cayan, D.R., Dettinger, M.D.: Response of global upper ocean temperature to changing solar irradiance. J Geophys Res, **102**, 3255–3266 (1997)
Willson, R.C., Mordinov, A.V.: Secular total solar irradiance trend during solar cycles 21 and 22. Geophys Res Lett, **30**, 1199–1202 (2003)

CHAPTER 2.2

INFLUENCE OF SOLAR ACTIVITY CYCLES ON EARTH'S CLIMATE

NIGEL D. MARSH
Center for Sun-Climate Research, Danish National Space Center, Copenhagen, Denmark

Abstract: In order to determine the influence of mankind on climate change it is important to understand the natural causes of *climate variability*. A natural effect that has been hard to understand physically is an apparent link between climate and *solar activity*. From historical and geological records there are strong indications that the sun has played an important role in the past climate of the Earth, but the *physical mechanism* is currently unknown. Whatever mechanism caused those earlier changes would most likely also be operating today and may have been active throughout the history of our planet. There have been several attempts to explain the link between solar activity and climate from variations in the sun's radiative output. These have tended to rely on simulations involving Global Climate Models (GCM), which are limited by our current understanding of the fundamental physics. In the following contribution, an outline of the current candidate mechanisms involving solar activity will be presented together with a description of the ESA funded project to study the Influence of Solar Activity cycles on Earth's Climate (ISAC)

INTRODUCTION

The observation that warm weather seems to coincide with high sunspot counts and cool weather with low sunspot counts was made 200 years ago by the astronomer William Herschel (Herschel 1801; Hoyt and Schatten 1992). Herschel noticed that the price of wheat in England was lower when there were many sunspots, and higher when there were few. Since the time of Herschel there have been numerous observations and non-observations of an apparent link between climate and the sunspot cycle, a large number of which have previously been recorded in various review articles and books on the subject (e.g., Dickinson 1975; Herman and Goldberg 1978; Hoyt and Schatten 1997).

J. Lilensten (ed.), Space Weather, 83–94.

There are three possible vectors between the Sun and the Earth that could lead to a solar imprint on climate; a) the electromagnetic radiation (*Total Solar Irradiance*) – or some component of it such as the ultra violet (*UV*), b) the direct *solar wind* through *magnetosphere/atmospheric coupling* and c) the *galactic cosmic radiation*, which, is modulated by the solar wind. Recently, a collaboration has been established, funded by ESA, to explore the Influence of Solar Activity Cycles on the Earth's climate (ISAC). The ISAC team, which includes the Space and Atmospheric Physics Department, Imperial College, UK, the Swedish Institute for Space Physics, Lund, Sweden, and the Center for Sun-Climate Research at the Danish National Space Center, Copenhagen, Denmark, intends to provide an up to date review of the current theories in solar modulation of climate. Our goals are threefold: 1) to describe how solar activity affects climate over a solar cycle, 2) to assess the likely impact of the currently proposed mechanisms linking solar activity to climate, and 3) to advise on possible methods for integrating these mechanisms into climate simulations. In the following an introduction to the ISAC project is presented together with a brief description of the candidate mechanisms that have been identified to date.

EVIDENCE FOR A SOLAR INFLUENCE ON EARTH'S CLIMATE

A solar influence on Earth's climate has been found in many climate parameters from the surface up to the top of the atmosphere. However, it is often average global temperature at the surface that is used as the key parameter for demonstrating variability in the average climate state. In the following sections, a summary is given of changes in observed temperature that coincide with variations in solar activity.

Ocean Temperatures

One example of a positive correlation is the apparent response of Sea Surface Temperatures to changing solar activity (Reid 1987; Reid 1991; Reid 2000). Sea Surface Temperatures (SSTs) have been obtained from ocean going ships since the middle of the 19th century. During the first part of the 20th century the observed SSTs increased, and then flattened out during the years 1940 and 1970, before continuing with the overall increasing trend. Figure 1 indicates that this long-term variability in SSTs is in phase with the 80–90 year envelope that modulates the approximately 11-year sunspot cycle.

White et al. (1997) confirmed this finding with two independent SST datasets, i.e., surface marine weather observations (1900–1991) and upper-ocean bathythermograph temperature profiles (1955–1994). They band-passed basin average temperatures, and found each frequency component to be in phase with changes in solar activity across the Indian, Pacific and Atlantic Oceans. Global averages yielded maximum changes of 0.08 ± 0.02 K on decadal (*ca.* 11-year period) scales and 0.14 ± 0.02 K on interdecadal (*ca.* 22-year period) scales in response to a $1\,\mathrm{Wm}^{-2}$ change in Total Solar Irradiance (TSI) reconstructed at the top of the atmosphere

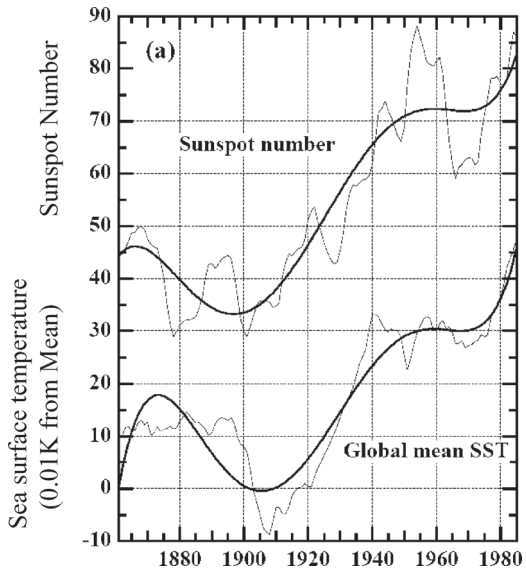


Figure 1. 11 year running mean of the annual sunspot numbers (upper thin curve), and the mean global sea-surface temperature anomaly (lower thin curve). The heavy curves represent a 7th degree polynomial least squares fit to the data. Units for the lower curves are 0.01K departures from the 1951–1980 average (adapted from Reid (2000)

by Lean et al. (1995). The highest correlations were obtained with ocean temperatures lagging solar activity by 1–2 years, which is roughly the time scale expected for the upper layers of the ocean ($< 100\,\mathrm{m}$) to reach radiative balance following a perturbation in TSI. From simple energy balance arguments White et al. (1997) estimated climate sensitivities due to changes in TSI at the ocean surface to be $0.2–0.4\,^{\circ}\mathrm{K}/(\mathrm{Wm}^{-2})$. This suggests that a $0.04–0.09\,^{\circ}\mathrm{K}$ change in SSTs would be expected from a $1\,\mathrm{Wm}^{-2}$ change in TSI at the top of the atmosphere. While these estimates are of a similar order of magnitude to the observed changes in global SSTs, they are on the low side, suggesting a possible amplification of the solar signal within the climate system.

Land Temperatures

Another example of a positive observation is the correlation between solar activity and northern hemisphere land temperatures. Friis-Christensen and Lassen (1991) used the sunspot cycle length as a measure of the Sun's activity. The cycle length averages 11 years but has varied from 7 to 17 years, with shorter cycle lengths corresponding to a more magnetically active Sun. A correlation was found between the sunspot cycle length and the change in land temperature of the northern hemisphere over the period 1861 to 1989 (latest update in Fig. 2). The land temperature of the northern hemisphere was used in order to avoid the lag by several years of air

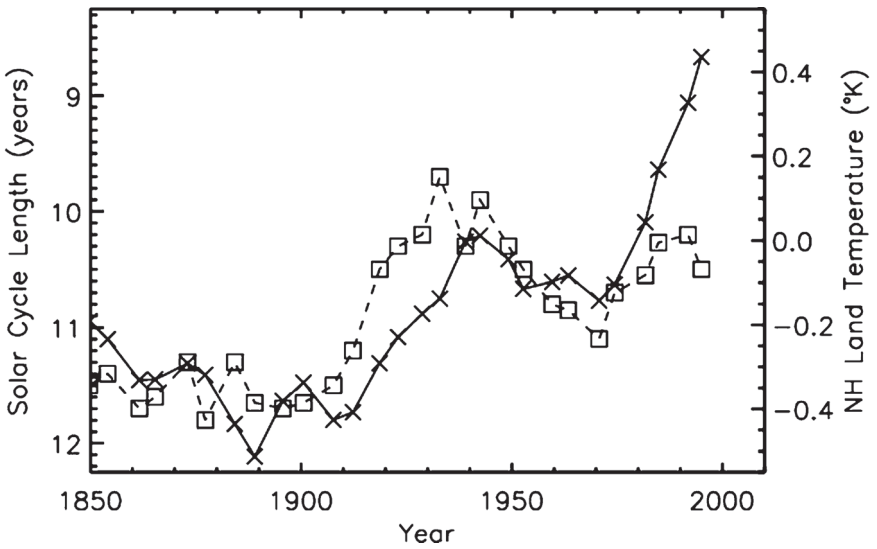


Figure 2. Sunspot cycle length smoothed with a 121 filter (dashed-squares, left hand scale) versus the northern hemisphere land temperature anomalies averaged over each solar cycle (solid-crosses, right hand scale) (adapted from Thejll and Lassen (2000))

temperatures over the oceans, due to their large heat capacity. Of particular note is the decrease between 1945 and 1970, which was also present in the Sea Surface Temperatures discussed above. This cannot easily be explained by the steadily rising concentrations of greenhouse gas in the atmosphere. It has been suggested that aerosols provided a counter radiative effect during this period, but it also appears to coincide with a decrease in the Sun's activity. The data plotted in Fig. 2 have been extended to include the most recent solar cycle. Clearly the correlation breaks down after 1990 and it has been suggested that this is an indication of the increasingly dominant effect which greenhouse gas emissions have had on global warming (Thejll and Lassen 2000).

However, the physical significance of the solar cycle length and whether it reflects changes in solar properties that in turn affect the Earth's environment are currently uncertain. Attempts have been made to attribute the solar cycle length with secular variations in the Sun's large-scale magnetic field, which influences the interplanetary shielding of cosmic rays arriving at Earth (Solanki et al. 2000), but there are still a number of open questions.

Tropospheric Temperatures

Radiosonde observations of tropospheric temperatures over the period 1958–2001 display significant variability at a number of different time-scales. From monthly data the effects of El Niño and volcanic eruptions are particularly evident. However,

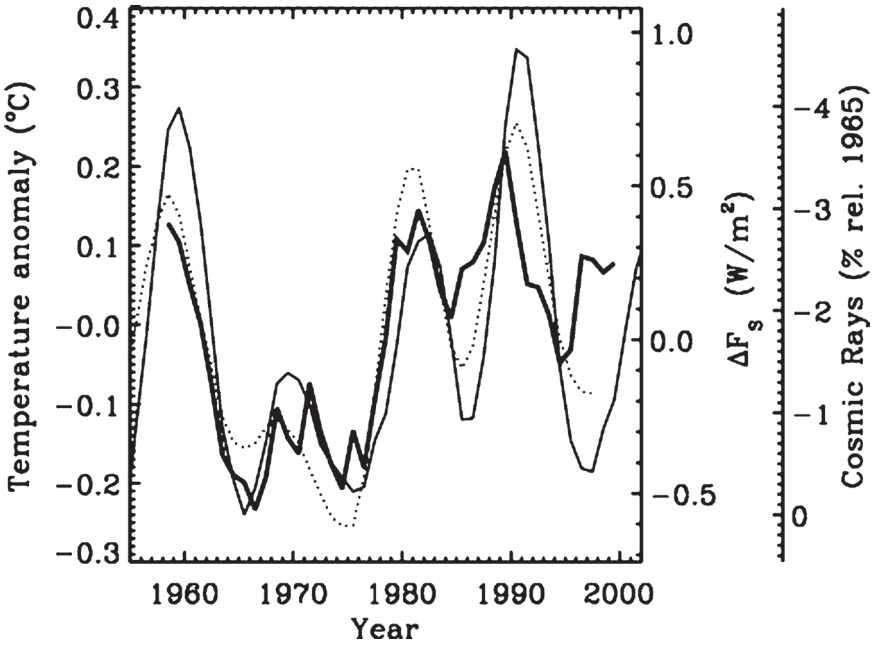


Figure 3. Tropospheric temperatures (thick), obtained from radiosondes, shown together with reconstructed TSI (dotted), ΔF_s, using re-scaled sunspot numbers as a proxy and cosmic rays (thin). All data sets have been low pass filtered with a 5 year running mean (adapted from, Marsh and Svensmark, 2003a)

these features are largely removed when filtering with a 3-year running mean (Marsh and Svensmark 2003a), and the low-pass tropospheric temperatures show a remarkably good agreement with changes in reconstructed TSI. Figure 3 indicates that an increase in reconstructed TSI of $1\,\mathrm{Wm}^{-2}$ coincides with an increase of $\sim 0.4\,^{\circ}\mathrm{K}$ in tropospheric temperatures.

An expected temperature response from the reconstructed changes in TSI can be estimated with a simple climate sensitivity analysis. A climate sensitivity of 0.6–$0.8\,\mathrm{K}/(\mathrm{Wm}^{-2})$, is obtained from the average response of climate models to a doubling of CO_2 (e.g., Appendix 9.1, Houghton et al. 2001). This predicts that a $1\,\mathrm{Wm}^{-2}$ change in TSI at the top of the atmosphere would result in only $\sim 0.1\,^{\circ}\mathrm{K}$ change in tropospheric temperature. Clearly, changes in TSI alone are too small to explain the observed tropospheric temperature variability and an amplification factor is required (Marsh and Svensmark 2003a).

OBSERVATIONAL EVIDENCE FOR INDIRECT MECHANISMS

Although a solar influence on climate is apparent, model studies have indicated that variations in the TSI are too small to explain the observed changes in recent climate. Stott et al. (2003) found that current climate models underestimate the

observed climate response to solar forcing over the 20th century as a whole, and concluded that the climate system has a greater sensitivity to solar forcing than models currently indicate. This is consistent with other studies that indicate the response to solar forcing in tropospheric temperatures is under estimated by a factor of 2 to 3 (Hill et al. 2001), and near-surface temperatures by a factor 2 (North and Wu 2001). These studies assumed the solar imprint on climate originated from direct changes in TSI at the top of Earth's atmosphere and did not include any possible indirect mechanism. Clearly an amplification of the solar signal, via some indirect mechanism(s), is required to explain the observed correlations with climate, and resolve the inconsistency with models.

The ISAC team has identified five external forcing parameters that are modulated by solar variability and have the potential to influence Earth's lower atmosphere below 50 km. These include: Total Solar Irradiance (TSI), the Ultra-Violet (UV) component of solar radiation, the direct input from the Solar Wind (SW), the total Hemispheric Power Input (HPI) reflecting properties of precipitating particles within the magnetosphere, and Galactic Cosmic Rays (GCR). In the table below, estimates of the energy input directly into Earth's environment is provided for each solar modulated parameter.

The quantities in the three right hand columns of Table 1 are miniscule compared to the TSI. Their variations in absolute terms are even smaller over a solar cycle, where TSI varies by about 0.1%, and the variations in the other parameters are $\sim 5\%$ for UV and 3–20% for GCR. Observational and modelling studies have suggested that the response to TSI over a solar cycle is too small to have had a significant impact on climate. Therefore, for the other solar-modulated quantities to play a role in climate change, some mechanism for magnifying their effect must be in place. These are briefly outlined below.

Table 1. Energy received at Earth from five Solar Modulated Parameters. The solar irradiance (TSI, Fröhlich 2000), ultra-violet radiation between 200–300nm (UV, http://rredc.nrel.gov/solar/spectra/am0/), solar wind (SW, ftp://nssdcftp.gsfc.nasa.gov/spacecraft_data/omni/omni2_2003.dat) and galactic cosmic ray (GCR, Bazilevskaya 2000) quantities are estimates at the top of the atmosphere outside the magnetosphere. The precipitating energy of charged particles within the magnetosphere has been estimated from the Total Hemispheric Power Input (HPI, http://spidr.ngdc.noaa.gov/spidr/dataset.do). The energy input from cosmic rays is nearly isotropic, whereas the solar irradiance impinges only on the dayside. The solar wind impinges on the much larger area of the magnetosphere, compared to the surface of the Earth, but only in the region of open field lines (polewards of the cusp) can the solar wind penetrate directly down to the uppermost layers of the dense atmosphere

Solar Modulated Parameter	TSI	UV	SW	HPI	GCR
Number Density (cm^{-3})	–	–	6	–	$5 \cdot 10^{-10}$
Energy/Particle(eV)	–	–	10^3	10^4	10^8
Energy Density($eVcm^{-3}$)	$2.8 \cdot 10^7$	–	$6 \cdot 10^3$	–	0.5
Velocity(ms^{-1})	$3 \cdot 10^8$	$3 \cdot 10^8$	$4.5 \cdot 10^5$	–	$3 \cdot 10^8$
Energy Flux (Wm^{-2})	1366	15	$4.4 \cdot 10^{-4}$	$6 \cdot 10^{-5}$	$2.4 \cdot 10^{-5}$

Solar UV

Various modelling studies have suggested that a response in atmospheric circulation can amplify the terrestrial effect of solar irradiance changes, possibly via the influence of solar UV variability on ozone concentrations and a corresponding response in stratospheric temperatures (Haigh 1996; Haigh 1999; Haigh 2003). Model results further suggest that this amplified stratospheric response to solar variability has the potential to influence tropospheric circulation patterns (Matthes et al. 2004). This aspect of solar modulated forcing is addressed by Haigh (this volume).

Direct Influence of the Solar Wind

The solar wind is able to generate significant heating of the lower thermosphere at high latitudes by direct charged particle precipitation, as well as generating upper atmospheric ionisation that may influence the global electric circuit. Through coupling with the magnetosphere the solar wind also drives ionospheric currents at high latitudes that in turn accelerate the neutral atmosphere. An intensification of both thermospheric and tropospheric flows following strong geomagnetic activity has been observed by Bucha and Bucha (1998). A mechanism connecting these two phenomena is currently uncertain, but they have suggested that downward winds generated in the polar cap of the thermosphere can penetrate to the stratosphere and finally couple with the troposphere. Arnold and Robinson (2001) have shown from modelling studies that planetary waves provided a means for coupling the solar-induced changes in the thermosphere down to the stratosphere during winter. Their modelled stratospheric response is qualitatively similar to that observed from UV changes.

Recent studies (Boberg and Lundstedt 2002, Boberg 2003) have shown a strong relationship between the electric field of the solar wind and a pressure phenomenon in the north Atlantic termed the North Atlantic Oscillation (NAO) (Marshall et al. 2001). A substantial portion of the climate variability in the Atlantic sector is associated with the NAO which varies over a wide range of timescales. A possible scenario for the solar wind/NAO interaction could include an electromagnetic disturbance induced by the solar wind in the ionosphere. This global disturbance could dynamically propagate downwards through the atmosphere, a scenario similar to the one proposed by Baldwin and Dunkerton (2001). This downward motion takes several weeks and during this time it would be affected by different atmospheric circulations from the equator toward the poles (Peixoto and Oort 1984), resulting in a pressure pattern more concentrated at high latitudes. Due to this slow, dynamic, propagation, it may be possible to make forecasts of the weather/climate in the Atlantic sector based on the solar wind state.

Magnetosphere/Ionosphere

The dynamics of the magnetosphere is driven by the solar wind, and thus is strongly correlated with solar activity. The physical state of the ionosphere is

determined predominantly by solar illumination and precipitating charged particles, so it too reflects variations in solar activity, but it is also sensitive to changes in the atmosphere below. For there to be a significant effect on the climate of the lower atmosphere, a strong amplifying mechanism is required which couples the solar wind-magnetosphere-ionosphere-lower atmosphere. This coupling could take various forms, e.g., through precipitating particles or through magnetospheric current systems closing in the ionosphere thereby depositing Joule heating to the upper layers of the atmosphere. While precipitating particles fluctuate by several orders of magnitude between quiet times and magnetic storms/substorms, magnetospheric current systems are more persistent existing even during very quiet periods. However, the energies involved are extremely small compared to TSI or UV (see Table 1); thus, a currently unknown amplifying mechanism is required if the magnetosphere – ionosphere system is to have a significant impact on the climate of the lower atmosphere.

Galactic Cosmic Rays

Cosmic rays are the main source of ionization in the troposphere (Bazilevskaya 2000), and there is increasing evidence that cosmic rays, which are modulated by the solar wind, can noticeably affect Earth's climate, via an influence on tropospheric cloud properties (Carslaw et al. 2002; Marsh and Svensmark 2000a, b, 2003b; Svensmark and Friis Christensen 1997).

The latest update of low cloud amount (LCA) and cosmic rays is shown in Fig. 4. However, variability in LCA correlates equally well with TSI or solar UV, and cannot be uniquely ascribed to a single mechanism when using globally-averaged data. This has led to suggestions that the cosmic ray-low cloud link is a result of a tropospheric circulation response to TSI or solar UV (Kristjansson et al. 2002). However, the ISCCP cloud data now span two solar cycles which allows for a qualitative comparison over the "22 year" cycle. This cycle is the result of a change in the Sun's magnetic polarity between consecutive solar cycles and affects the shielding of GCR as they enter the heliosphere. Under solar minimum conditions this results in a sharp GCR peak during the late 1980's with a more blunted GCR peak during the 1990's, a feature which is also apparent in LCA, but not in TSI or UV (see Fig. 4). Another way to distinguish between these processes is to utilize the property that cosmic rays arriving at Earth are additionally modulated by the geomagnetic field whereas solar irradiance is not. Recent observational evidence indicates that the solar cycle amplitude in LCA, over the period 1984–2000, increases polewards and possesses a similar latitudinal dependence to that found in cosmic ray-induced ionization (CRII) of the troposphere (Usoskin et al. 2004). This supports a physical mechanism involving cosmic rays rather than solar irradiance.

Ionization from GCR possibly influences the atmospheric aerosol distribution on which clouds form, and/or affects the global atmospheric electric circuit with the

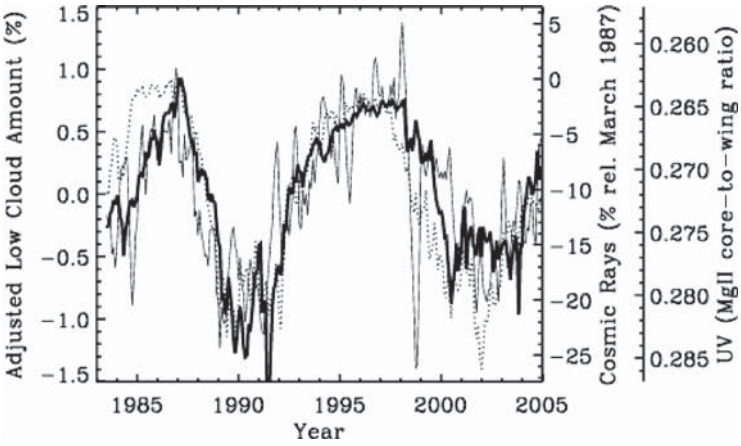


Figure 4. Monthly averages of ISCCP-D2 IR global Low Cloud Amount derived from a combination of polar orbiting and geostationary satellites (thin), cosmic rays (thick), and solar UV (dotted). The Low Cloud Amount has been adjusted to take account of a possible inter-calibration problem after 1994 (see Marsh and Svensmark 2003b)

potential to enhance aerosol-droplet collision efficiencies at cloud boundaries. Both mechanisms are briefly outlined below.

***Ion Induced Nucleation (IIN)*:** Ions produced through the nucleonic cascade of cosmic rays in the troposphere rapidly interact with atmospheric molecules and are converted to complex cluster ions (aerosols) (Gringel et al. 1986; Hoppel et al. 1986). It is thought that these cluster ions grow through ion-ion recombination or ion-aerosol attachment, and thus affect the number of aerosols acting as cloud condensation nuclei (CCN) at typical atmospheric supersaturations of a few percent (Viggiano and Arnold 1995). Recent atmospheric observations indicate a role for ion induced nucleation (IIN) in ultra-fine aerosol formation (sizes < 10 nm, Eichkorn et al. 2002; Lee et al. 2003). But it remains an open question as to whether aerosol concentrations at CCN sizes (~ 100 nm) are sensitive to a perturbation in ionisation and capable of significantly influencing cloud properties. Nucleation modelling studies suggest that it is the lower troposphere (below 5 km) that is most sensitive to changes in IIN (Yu 2002). Under such conditions, an increase in GCR would lead to an increase in aerosol and hence a decrease in cloud droplet sizes. Ferek et al. (2000) have shown that an increase in aerosol concentrations due to ship exhaust can lead to drizzle suppression which has implications for cloud lifetimes. If ionization from GCR can be shown to have a similar affect on the lower tropospheric aerosol distribution, and subsequently prolong a cloud's lifetime, it would be consistent with the cosmic ray – low cloud correlation outlined above. However, ship tracks are a large perturbation locally, whereas a possible GCR – CCN mechanism will be a small perturbation globally. The possible link between atmospheric ionization and aerosols acting as CCN requires confirmation by experiment to determine its potential implications for climate. Currently two such experimental efforts are

underway at DNSC, Copenhagen, Denmark (http://www.dsri.dk/sun-climate/) and at CERN, Geneva, Switzerland (http://cloud.web.cern.ch/cloud/). Initial results from the Copenhagen experiment suggest that ionisation does play a role in aerosol nucleation in the atmosphere (Svensmark et al. 2006).

Global Atmospheric Electric Circuit: A further suggestion is that the amplification of cosmic rays on climate could be through changes in the global atmospheric electric circuit (Rycroft et al. 2000). Current flowing in the global atmospheric electrical circuit substantially decreased during the 20th century, which has been quantitatively explained by a decrease in cosmic rays (CR) reducing the ionospheric potential as solar activity increased (Harrison 2002). This potentially affects aerosol-cloud interactions at the edges of clouds, e.g., Tinsley (2000), (see a review of possible mechanisms in Harrison and Carslaw (2003)). Highly-charged droplets are generated at cloud boundaries in the troposphere due to the weak vertical currents of the global electric circuit. Once these droplets have evaporated, highly charged CCNs remain, and the presence of this charge enhances collision efficiencies when interacting with other liquid droplets. The process of nucleation and evaporation repeats itself continuously and is thought to aid in the formation of ice particles in supercooled liquid water clouds; as a result it is referred to as 'electroscavaging' (Tinsley 2000). There is some limited observational evidence to suggest that this process can have an additional influence on atmospheric dynamics (Roldugin and Tinsley 2004).

SUMMARY

This introduction to the ISAC project has emphasized the need to better understand the impact which solar variability can have on Earth's climate. Traditionally climate studies have focused on solar irradiance or some component of the solar spectrum such as UV. Here a very brief review of other potential mechanisms linking solar variability to climate has been presented. The goal of the ISAC team is to try to quantify the relative impact from all five solar modulated processes. The atmospheric response to solar variability will be characterized at various timescales by comparing the five solar parameters with various climate parameters from the stratosphere down to the surface. By identifying the nature of the climate response both temporally and spatially, better constraints can be placed on the physical mechanisms that may link changes in Earth's climate to solar variability.

REFERENCES

Arnold, N.F., Robinson, T.R.: Solar magnetic flux influences on the dynamics of the winter middle atmosphere. Geophys. Res. Lett. **28**(12), 2381–2384 (2001)

Baldwin, M.P., Dunkerton, T.J.: Stratospheric harbingers of anomalous weather regimes. Science, **294**(5542), 581–584 (2001)

Bazilevskaya, G.A.: Observations of variability in cosmic rays. Space Sci Rev, 94(1–2), 25–38 (2000)

Boberg, F., Lundstedt, H.: Solar wind variations related to fluctuations of the North Atlantic Oscillation. Geophys Res Lett, **29**(15), 10.1029/2002GL014903 (2002)

Boberg, F., Lundstedt, H.: Solar wind electric field modulation of the NAO: A correlation analysis in the lower atmosphere. Geophys Res Lett, **30**(15), 10.1029/2003GL017360 (2003)

Bucha, V., Bucha, V.: Geomagnetic forcing of changes in climate and in the atmospheric circulation. J Atmos Sol-Terr Phys, **60**(2), 145–169 (1998)

Carslaw, K.S., Harrison, R.G., Kirkby, J.: Cosmic rays, clouds, and climate. Science, **298**(5599), 1732–1737 (2002)

Dickinson, R.E.: Solar variability and the lower atmosphere. Bull Am Met Soc, **56**(12), 1240–1248 (1975)

Eichkorn, S., Wilhelm, S., Aufmhoff, H., Wohlfrom, K.H., Arnold, F.: Cosmic ray-induced aerosol-formation: First observational evidence from aircraft-based ion mass spectrometer measurements in the upper troposphere. Geophys Res Lett, **29**(14), 10.1029/2002GL014903 (2002)

Ferek, R.J., Garrett, T., Hobbs, P.V., Strader, S., Johnson, D., Taylor, J.P., Nielsen, K., Ackerman, A.S., Kogan, Y., Liu, Q.F., Albrecht, B.A., Babb, D.: Drizzle suppression in ship tracks. J Atmos Sci, **57**(16), 2707–2728 (2000)

Friis Christensen, E., Lassen, K.: Length of the solar-cycle – An indicator of solar-activity closely associated with climate. Science, **254**(5032), 698–700 (1991)

Fröhlich, C.: Observations of irradiance variability. Space Sci. Rev. 94, 15–24 (2000)

Gringel, W.J., Rosen, M., Hofmann, D.J.: Electrical Structure from 0 to 30 kilometers. In Officer, C.B., Cronin, L.E. (eds) Earth's electrical environment. National Academy Press, Washington DC, pp 166–182 (1986)

Haigh, J.D.: The impact of solar variability on climate. Science, **272**(5264), 981–984 (1996)

Haigh, J.D.: A GCM study of climate change in response to the 11-year solar cycle. Quar J Royal Met Soc, **125**(555), 871–892 (1999)

Haigh, J.D.: The effects of solar variability on the Earth's climate. Phil Trans Royal Soc A, **361**(1802), 95–111 (2003)

Harrison, G.: Twentieth century secular decrease in the atmospheric potential gradient. Geophys Res Lett, **29**(14), 10.1029/2002GL014878 (2002)

Harrison, R.G., Carslaw, K.S.: Ion-aerosol-cloud processes in the lower atmosphere. Rev Geophys, **41**(3), 10.1029/2002RG000114 (2003)

Herman, J.R., Goldberg, R.A.: Sun, weather, and climate. NASA, Washington DC (1978)

Herschel, W.: Observations tending to investigate the nature of the Sun, in order to find the causes or symptoms of its variable emission of light and heat; With remarks on the use that may possibly be drawn from solar observations. Phil. Trans. Royal Soc. **91**, 265–318 (1801)

Hill, D.C., Allen, M.R., Stott, P.A.: Allowing for solar forcing in the detection of human influence on tropospheric temperatures. Geophys Res Lett, **28**(8), 1555–1558 (2001)

Hoppel, W.A., Anderson, R.V., Willet, J.C.: Atmospheric electricity in the planetary boundary layer. In Officer, C.B., Cronin, L.E. (eds) Earth's electrical environment. National Academy Press, Washington DC, pp 149–165 (1986)

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Xiaosu, D.: Climate Change 2001: The Scientific Basis. Cambridge University Press (2001)

Hoyt, D.V., Schatten, K.H.: New Information on Solar-Activity, 1779–1818, from Herschel,William Unpublished Notebooks. Astrophys J, **384**(1), 361–384 (1992)

Hoyt, D.V., Schatten, K.H.: The role of the sun in climate change. Oxford University Press, New York (1997)

Kristjansson, J.E., Staple, A., Kristiansen, J., Kaas, E.: A new look at possible connections between solar activity, clouds and climate. Geophys Res Lett, **29**(23), 10.1029/2002GL015646 (2002)

Lean, J., Beer, J., Bradley, R.: Reconstruction of Solar Irradiance since 1610: Implications for Climate Change, Geophys Res Lett, **22**(23), 10.1029/95GL03093 (1995)

Lee, S.H., Reeves, J.M., Wilson, J.C., Hunton, D.E., Viggiano, A.A., Miller, T.M., Ballenthin, J.O., Lait, L.R.: Particle formation by ion nucleation in the upper troposphere and lower stratosphere. Science, **301**(5641), 1886–1889 (2003)

Marsh, N., Svensmark, H.: Cosmic rays, clouds, and climate. Space Sci Rev, **94**(1–2), 215–230 (2000a)

Marsh, N., Svensmark, H.: Low cloud properties influenced by cosmic rays. Phys. Rev. Lett. **85**(23), 5004–5007 (2000b)

Marsh, N., Svensmark, H.: Solar influence on Earth's climate. Space Sci Rev, **107**(1–2), 317–325 (2003a)

Marsh, N., Svensmark, H.: Galactic cosmic ray and El Nino Southern Oscillation trends in International Satellite Cloud Climatology Project D2 low-cloud properties. J Geophys Res, **108**(D6), 10.1029/2001JD001264 (2003b)

Marshall, J., Kushner, Y., Battisti, D., Chang, P., Czaja, A., Dickson, R., Hurrell, J., McCartney, M., Saravanan, R., Visbeck, M.: North Atlantic climate variability: Phenomena, impacts and mechanisms. Int J Clim, **21**(15), 1863–1898 (2001)

Matthes, K., Langematz, U., Gray, L.L., Kodera, K., Labitzke, K.: Improved 11-year solar signal in the freie universitat Berlin climate middle atmosphere model (FUB-CMAM). J Geophys Res, **109**(D6), 10.1029/2003JD004012 (2004)

North, G.R., Wu, Q.G.: Detecting climate signals using space-time EOFs. J Clim, **14**(8), 1839–1863 (2001)

Peixoto, J.P., Oort, A.H.: Physics of Climate. Rev Mod Phys, **56**(3), 365–429 (1984)

Reid, G.C.: Influence of solar variability on global Sea-Surface Temperatures. Nature, **329**(6135), 142–143 (1987)

Reid, G.C.: Solar total irradiance variations and the global Sea-Surface Temperature record. J Geophys. Res. **96**(D2), 2835–2844 (1991)

Reid, G.C.: Solar variability and the Earth's climate: Introduction and overview. Space Sci. Rev. **94**(1–2), 1–11 (2000)

Roldugin, V.C., Tinsley, B.A.: Atmospheric transparency changes associated with solar wind-induced atmospheric electricity variations. J Atmos Sol-Terr Phys, **66**(13–14), 1143–1149 (2004)

Rycroft, M.J., Israelsson, S., Price, C.: The global atmospheric electric circuit, solar activity and climate change. J Atmos Sol-Terr Phys, **62**, 1563–1576 (2000)

Solanki, S.K., Schussler, M., Fligge, M.: Evolution of the Sun's large-scale magnetic field since the Maunder minimum. Nature, **408**(6811), 445–447 (2000)

Stott, P.A., Jones, G.S., Mitchell, J.F.B.: Do models underestimate the solar contribution to recent climate change? J Clim, **16**(24), 4079–4093 (2003)

Svensmark, H., Friis Christensen, E.: Variation of cosmic ray flux and global cloud coverage – A missing link in solar-climate relationships. J Atmos Sol-Terr Phys, **59**(11), 1225–1232 (1997)

Svensmark, H., Pedersen, J.O.P., Marsh, N.D., Enghoff, M.B., Uggerhøj, U.I.: Experimental evidence for the role of ions in particle nucleation under atmospheric conditions. Proc. R. Soc. A, doi: 10.1098/rspa.2006.1773 (2006)

Thejll, P., Lassen, K.: Solar forcing of the Northern hemisphere land air temperature: New data. J Atmos Sol-Terr Phys, **62**(13), 1207–1213 (2000)

Tinsley, B.A.: Influence of solar wind on the global electric circuit, and inferred effects on cloud microphysics, temperature, and dynamics in the troposphere. Space Sci. Rev. **94**(1–2), 231–258 (2000)

Usoskin, I.G., Marsh, N., Kovaltsov, G.A., Mursula, K., Gladysheva, O.G.: Latitudinal dependence of low cloud amount on cosmic ray induced ionization. Geophys Res Lett, **31**(16), 10.1029/2004GL019507 (2004)

Viggiano, A.A., Arnold, F.: Ion Chemistry and Composition of the Atmopshere. In Volland, H. (ed.) Handbook of atmospheric electrodynamics. CRC Press, Washington DC, pp 1–26 (1995)

White, W.B., Lean, J., Cayan, D.R., Dettinger, M.D.: Response of global upper ocean temperature to changing solar irradiance. J Geophys Res, **102**(C2), 3255–3266 (1997)

Yu, F.Q.: Altitude variations of cosmic ray induced production of aerosols: Implications for global cloudiness and climate. J Geophys Res, **107**(A7), 10.1029/2001JA000248 (2002)

CHAPTER 2.3

UNRAVELLING SIGNS OF GLOBAL CHANGE IN THE IONOSPHERE

THOMAS ULICH[1], MARK A. CLILVERD[2], MARTIN J. JARVIS[2] AND HENRY RISHBETH[3]

[1]*Sodankylä Geophysical Observatory, Sodankylä, Finland*
[2]*British Antarctic Survey, Cambridge, UK*
[3]*School of Physics and Astronomy, University of Southampton, Southampton, UK*

Abstract: As a consequence of alterations of atmospheric chemical composition due to anthropogenic emissions, Earth's ionosphere and thermosphere are expected to change. A number of authors tried to detect signs of global change in their ionospheric data, but many findings remain controversial. We briefly review long-term trends observed in the critical frequencies of the ionospheric E and F2 layers as well as in the height of the F2-peak, i.e. the layer of maximum electron density. Using 48 years of F2-layer critical frequency data from Sodankylä, Finland, we demonstrate how the sign and amplitude of the detected trends depend upon the choice of model fitted to the data and suggest a method to choose the best possible model

INTRODUCTION

Predictions of changes of atmospheric chemical composition due to anthropogenic greenhouse gas emissions led Brasseur and Hitchman (1988) to study the effects of these changes on the upper atmosphere using a 2D numerical model. They anticipated the stratopause to cool by 8 to 15 K for doubled concentrations of carbon dioxide (CO_2) and methane (CH_4). Roble and Dickinson (1989) followed up on this idea using their global mean model of the upper atmosphere and ionosphere. They predicted the mesosphere and thermosphere to cool by 10 K and 50 K, respectively, for doubled CO_2 and CH_4 at 60 km altitude. Since ionospheric temperatures are expected to change much more dramatically with greenhouse gas increases than surface temperature, which is expected to rise only by a few degrees K, this "greenhouse effect" should thereby be more easily detectable in the

J. Lilensten (ed.), Space Weather, 95–105.

ionosphere than on the ground. Rishbeth (1990), using a brief calculation based on scale heights, estimated the changes expected in some standard ionospheric parameters. He concluded that only the F2-layer peak height could be expected to show significant change. Later Rishbeth and Roble (1992) used the NCAR computational Thermosphere-Ionosphere General Circulation Model and largely verified Rishbeth's earlier estimates.

Changes in temperature affect both heights and critical frequencies of ionospheric layers, but for different reasons. The 'photochemical' E and F1 layers form where the ionizing radiations that produce them reach unit optical depth, which depends on solar zenith angle, the absorption cross-sections for these radiations, and atmospheric pressure. The F2 peak (normally the level of highest electron density) is controlled by a balance between photochemical and transport processes, and this, too, tends to occur at some given pressure-level (Garriott and Rishbeth 1963; Rishbeth and Edwards 1989). The height $h(p)$ of any fixed pressure-level p depends on how the scale height H varies with height, and may be computed from an integration from the ground (height $h = 0$, pressure p_o) up to the required pressure-level p, namely $\ell n(p_o/p) = \int (dh/H)$.

With increasing greenhouse gas concentrations, the thermosphere cools as the lower atmosphere warms. There will be an "isopycnic level" in the middle atmosphere where the pressure stays constant. The E layer is only a few scale heights above that level, so its height will not decrease much as greenhouse gases increase. But the F2 layer is many scale heights above the isopycnic level, and the cumulative effects of the decreases of H below it should lower its peak height hmF2. The maps of Rishbeth and Roble (1992) indeed show that in most places hmF2 drops by 10–20 km if atmospheric CO_2 and CH_4 are doubled.

As regards the peak electron density of a layer, the total number of ions and electrons produced along a slant path from the Sun through the ionosphere is determined by the flux of ionising photons and the chemical composition of the atmosphere on which the radiation acts. The thickness of any layer depends on the local scale height, which is proportional to temperature. By this simple consideration, the peak electron density should vary inversely with temperature, so that $N\mathrm{m} \propto 1/T$. Beyond that, any further change due to "greenhouse cooling" depends on how the rate coefficients of the chemical and transport processes vary with temperature. Furthermore, both layer height and peak electron density could be further affected if global change alters the chemical composition of the thermosphere, e.g. by changing the level of turbulence in the turbopause region.

Over the years a number of authors have used their various data sets to detect long-term change consistent with theoretical estimates. These include most prominently the F2-peak height, and the F2- and E-layer critical frequencies (foF2 and foE). Furthermore, there have been studies of radio path reflection (e.g. Taubenheim et al. 1990) and ionospheric absorption (Serafimov and Serafimova 1992). Many differing and partly conflicting results have been obtained, possibly as a result of the trend analysis procedures. To highlight this we briefly summarise the published research

on E- and F2-layer trends. Then we demonstrate the difficulties of comparing the results of different authors and suggest an optimised method for determining trends.

TRENDS IN F2-LAYER PEAK HEIGHT

Even though Rishbeth (1990) suggested F2-layer peak height to be the most obvious parameter in which to expect an ionospheric greenhouse effect, it possibly is the most complicated one to analyse. Usually, *h*mF2 is not routinely scaled from ionograms; it has to be estimated using other regularly scaled parameters. Shimazaki (1955) showed that it is approximated by a formula involving the maximum usable frequency factor $M(3000)$F2 for a 3000-km radio path of the form $A + B/M(3000)$F2, where A, B are numerical constants. Several authors published improved empirical formulae, adding to $M(3000)$F2 a correction term ΔM that allows for the effect of ionisation below the F2 peak (see Dudeney 1983). The choice of some formulae can even reverse the sign of the observed trends (Jarvis et al. 2002). Consequently, the applicability of any such formula to a given ionosonde needs to be verified (Ulich 2000).

Bremer (1992) was the first to derive trends from ionosonde data; he found a negative trend much stronger than predicted ($-0.24\,\mathrm{km/yr}$) for *h*mF2 from Juliusruh, Germany. Later on, Ulich and Turunen (1997) reported negative *h*mF2 trends for Sodankylä ($-0.39\,\mathrm{km/yr}$). Jarvis et al. (1998) studied data from Argentine Island, Antarctica, and Port Stanley, Falkland Is., and found the trends to depend upon season and time of day, however most trends were negative (between -0.1 and $-0.4\,\mathrm{km/yr}$). Bremer (1998) arrived at the puzzling result of negative trends west and positive trends east of 30 °E. Upadhyay and Mahajan (1998) published the first global study but their results were inconclusive, finding seven positive and as many negative *h*mF2 trends, some in disagreement with Bremer's results. Trends at Tucumán, Argentina, turned out to be negative (Ortiz de Adler et al. 2002) and so did trends derived by Clilverd et al. (2003) for 11 stations. Xu et al. (2004) tested several trend models against *h*mF2 of Kokubunji, Japan, and found negative trends. Ulich (2000) studied 66 stations worldwide and found negative and positive trends, some changing sign with the trend model or empirical *h*mF2-formula used. He did not find any geographic consistency.

TRENDS IN F2-LAYER CRITICAL FREQUENCY

Cooling of the thermosphere should cause little change in foF2: the computations of Rishbeth and Roble (1992) found changes (mainly decreases) of no more than 0.5 MHz for doubled CO_2. Bremer (1992) did not find any significant changes in electron density at Juliusruh. Upadhyay and Mahajan (1998) reported only small trends in foF2 (17 negative, 14 positive) and concluded that ionospheric data do not contain definitive evidence of a global trend. Bremer (1998) found the sign of

trends to be negative (positive) to the west (east) of 30°E. Sharma et al. (1999) found a rather strong negative trend (−40 kHz/yr) for foF2 at Ahmedabad, India. In the same year, Danilov and Mikhailov (1999) studied 22 stations and found all of them to have a negative trend. Thereafter, Mikhailov and Marin (2000) reported the trends of 30 northern hemisphere stations to depend upon geomagnetic latitude. Their "geomagnetic control concept" suggests geomagnetic effects to dominate over possible climate change effects.

TRENDS IN E-LAYER CRITICAL FREQUENCY

Trends in the E layer are suggested to be practically non-detectable (Rishbeth, 1990). However, a few attempts have been made to observe long-term change in critical E-layer frequency. Bremer (1998) studied 25 stations and found only 11 trends to be significant at 90% confidence level and their magnitude was typically a few kHz/yr, positive as well as negative. Mikhailov and de la Morena (2003) found that the geomagnetic control concept applies to foE, too. Peculiarly, they find an anti-correlation of long-term changes of geomagnetic activity (Ap index) with foE long-term trends, but only before the early 1970s. Thereafter this dependency breaks down and the authors speculate whether this could be due to anthropogenic effects such as chemical pollution.

PROBLEMS OF TREND DETERMINATION

Several problems need to be addressed when considering long-term trends, especially if one wants to compare results by different authors: for instance the resolution of the data used; if smoothing filters are used; if and how known variations have been removed from the time series. Ulich et al. (2003) highlighted some principal problems with trend studies and pointed out that quality and consistency of historic geophysical time series must be ensured. Changes of instrument hardware or location as well as changes of personnel can introduce discontinuities or gaps in the data set. Moreover, the observed trends are often smaller than the operational accuracy of the instruments and thus the significance of trends has to be verified and error bars should accompany trends.

In order to demonstrate the impact of data treatment and model composition on trend analysis, the present work focuses on long-term trends in F2-layer critical frequency, which is readily obtained from ionosondes and does not require the use of empirical formulae. However, many issues pointed out here are equally valid for other (ionospheric) time series. Here we have tested a number of physically plausible long-term trend models against the foF2 record from the Sodankylä Geophysical Observatory, Finland. Located in the auroral zone in Northern Finland (67°22′N, 26°38′ E, L = 5.2), the observatory, established in 1913, began vertical ionospheric soundings in August 1957. Since then the ionosonde was only changed twice, in February 1977 and in November 2005. However, the present work includes only data obtained by the first two instruments. From the beginning to February

2002, almost all ionograms were scaled by the same person. The Sodankylä ionosonde data are thus of very high quality and consistency and cover more than 48 years.

Multi-parameter Models

Known components have to be removed from a time series in order to reveal the unknown components, which make up the variation of the residual. These could be changes of atmospheric chemical composition, the Earth's magnetic field, thermospheric winds or atmospheric dynamics in general. But the residual also includes measurement errors and possibly elements thus far not considered at all.

Some authors remove known variabilities in separate steps, making it very difficult to estimate an error for the resulting trends. Here we use linear combinations of base functions $f_k(t)$ to model the data, which replicates the models used in published trend analyses. All of these models include a constant x_1 and a linear trend $x_2 t$. The parameters x_k of the models were then fitted to the foF2 data in the least-squares sense by means of singular value decomposition. The main advantage of this method is that it fits all model components in a single step giving the most probable solution for the unknown x_k and their standard error. Here we assume foF2 to be accurate to within 1 MHz. If t_i are the sampling times and N is the number of functions included, any of these models is given by

$$M(t_i) = x_1 + x_2 t_i + \sum_{k=3}^{N} x_k f_k(t_i).$$

The inclusion of the sampling times is another major advantage of this method, because it removes the need for interpolating gaps and thereby introducing possibly misleading pseudo data into the time series. The most basic model contains only the first two terms. Additionally, these models can contain variations of solar or geomagnetic activity, or seasonal cycles.

Choice of the Proxy of Solar Activity

Most prominently, ionospheric data are affected by variations of solar activity. Jarvis et al. (2002) and Clilverd et al. (2003) showed that solar activity variations need to be removed from the data prior to determining a trend. They found that the trend depends on the phase of the solar cycle at the beginning and the end of the data set: if the data start at solar maximum and end at solar minimum, then a strongly negative trend is likely in foF2 and hmF2, which correlate positively with solar activity. Since most ionosondes started during the IGY in 1957, the year of the highest solar activity in the 20th century, any time series strongly modulated by solar activity and starting at that time will show a negative trend.

However, solar activity can be expressed in a number of different ways. The simplest approach would be to simulate it by a sinusoidal wave having a period

of 11 years. This would certainly be too crude, since solar activity cycles are not symmetrical. Thanks to modern computers, one can directly use, e.g., Zürich sunspot numbers or sunspot group numbers in the tradition of the Royal Greenwich Observatory (RGO) to name but a few. Besides counting spots on the Sun's surface, which might appear a rather unphysical "measurement" of solar activity, there are records of solar 10.7-cm radio fluxes in form of "adjusted" values normalised to a Sun–Earth distance of 1 AU and as "observed" values including the 3.3% distance variation. Here we use monthly means of both fluxes (Fadj and Fobs) and of Zürich sunspot counts (SSN). However, since the ionosphere is formed largely by extreme UV and X radiation from the Sun, other proxies are possible, too, e.g. the E10.7 index (Tobiska et al. 2000).

Geomagnetic Activity and Seasonal Variations

Geomagnetic activity variations, which also show long-term changes (Clilverd et al. 1998), directly affect the ionosphere, which can be taken into account by means of geomagnetic indices. Most readily available are the planetary *A*p and *K*p indices. *K*-based indices are not suitable, because *K* is a logarithmic scale. However, the linear *A*p index can directly be used for correlation studies. One could also use the local *A*k from the ionosonde site, if this is available. Here we use monthly means of both *A*p and Sodankylä *A*k indices.

Ionospheric foF2 is modulated by annual and semi-annual variations (Rishbeth et al. 2000). They are included in the trend models in form of their Fourier components with periods of 1 year (Ann) and half a year (S-Ann).

Resolution and Filtering of the Data

The task is to find a trend over several solar activity cycles and thus it seems inappropriate to use high-resolution data such as hourly or daily values. On the other hand, any averaging or filtering removes features and makes the data less physical. A completely featureless time series will show a high degree of correlation with another featureless data set even if they lack any physical relation. Thus the resolution of the data sets has to be chosen carefully. Typically daily, monthly or annual averages are used for trend studies, with monthly data being most common. Often data sets are smoothed by low-pass filters: monthly data might be smoothed using a filter window of 1 year in order to remove seasonal variability. Some authors use 11-year filters to remove solar activity variation. Sometimes an 11-year running mean filter is applied to annual averages.

When using filters, two issues need to be tackled: (a) to which point of the filter window the result is assigned and (b) how parts of the time series are treated, where the sliding window is not entirely populated with data. The former question (a) merely leads to a shift in time of the filtered with respect to the unfiltered data set, but the trend magnitude is not affected. In order to avoid the time shift, the result of the filter is usually assigned to the centre of the filtering window. The latter

issue (b) is more important. If the filter window is not completely covered with data, artefacts might be introduced. This happens whenever there are large gaps, but also (for a centred window) at the ends of the data set. Thus the filter can only be applied to data half a window length from each end of the time series and thus it reduces the amount of usable data.

Furthermore, the level of filtering of model components and foF2 needs to be consistent. For instance it is not appropriate to fit unfiltered monthly values of, say, solar activity to monthly foF2 filtered by a 1-year running mean filter.

In the present work we used exclusively monthly averages, but for comparison we smoothed these by centred 13-month and 131-month (11-year) running mean filters. Besides solar and geomagnetic activity parameters, the original foF2 data were filtered. In all cases, the filter window had to be almost completely populated with data, which means, in the case of the 11-year running mean, that the filtered time series is shortened by 9 years (requiring fully-populated filter windows would have been too restrictive due to gaps in the foF2 data). While all available data were used for filtering, all time series were thereafter cut to the longest common interval, so that smoothed as well as unsmoothed data sets are of identical duration covering 40 years from October 1961 to September 2001.

UNRAVELLING LONG-TERM TRENDS

Subsequently, we selected 133 physically plausible combinations of these model components all of which are given in the top panel of Fig. 1. Every row refers to a component given above. Crosses, circles, and stars refer to unsmoothed, 13-month and 131-month filtered data, respectively. Constant and slope are always present. Semi-annual and annual variations appear only together. The first model includes only constant and slope. Models 2–7 neglect solar but include geomagnetic activity. Models 8–70 combine solar and geomagnetic activity, while 71–133 include also seasonal variations.

All of these models were fitted to the Sodankylä foF2 time series, which itself was used in the form of monthly medians as well as 13 and 131 month running means thereof. The fitted model was subtracted from the original data and the standard deviation of the residual was obtained. This value (2nd panel) represents the "goodness of the fit:" the smaller it is, the better the model represents the data.

The fit was done only when the degree of smoothing of the measured foF2 was less than that of the model. For instance, all models containing semi-annual and annual variation were fitted only to unfiltered foF2, because any filter would have removed the seasonal variation. In the lower panels of Fig. 1, the symbols indicate the type of foF2 used for the fit: crosses mean "raw" monthly medians, circles and stars refer to 13 and 131 month running means.

The third panel of Fig. 1 depicts the probable error of the slope parameter in kHz per year and the fourth panel shows the slope x_2 of the foF2 time series. If the fit of the slope was not confident to at least 95% according to the Fisher criterion (Bremer, 1992), the result was excluded from the last panel.

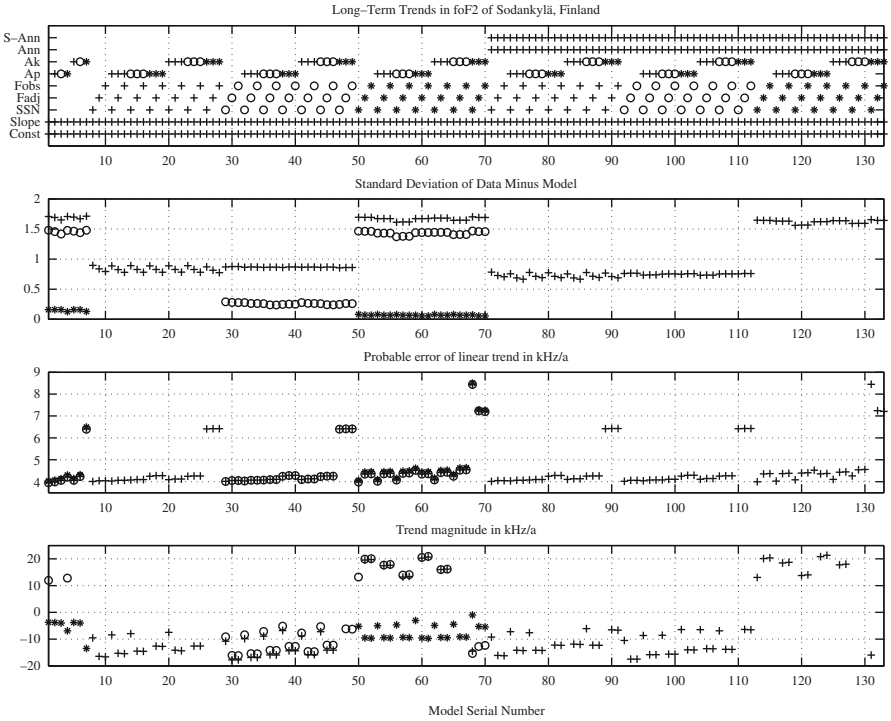


Figure 1. Top panel: composition of the model (crosses: monthly averages; circles and stars: 13 and 131 month running mean). *Second panel*: standard deviation of the residual after subtracting the model from the data. *Third panel*: probable error of the slope parameter in kHz/year. *Fourth panel*: trend magnitudes in kHz/year at confidence limits $\geq 95\%$. In all panels the horizontal axis is the serial number of the model used for the fit. The model was fitted to monthly foF2 (crosses), 13-month smoothed foF2 (circles), or 11-year smoothed foF2 (stars)

The trends determined by fitting these models to the same data set vary greatly: the smallest trend (at 95% confidence level) is $-18 \pm 4\,\mathrm{kHz/yr}$ and the largest is $21 \pm 4\,\mathrm{kHz/yr}$. Most trends are negative, i.e. foF2 decreases over time, but some of the trends are positive. Comparing top and bottom panel, it is apparent that the trends turn positive whenever an 11-year running mean of solar activity is used (Models 50–70 and 113–133) but only if foF2 was not smoothed using the same filter. Models 29–49, which use 13-month running means of solar activity, do not show this behaviour: the resulting trends are very similar to one another, but those fitted to the non-smoothed foF2 data are consistently smaller. This demonstrates that (Sodankylä) foF2 is mainly modulated by solar activity and that the same level of smoothing should be applied to both solar activity and foF2. Moreover, Models 1–7 (no solar activity) do not yield a confident trend for monthly data and only two confident trends for 13-month smoothed foF2. They can establish a confident trend only after smoothing over one solar cycle.

In analogy to Jarvis et al. (2002), we studied the trends determined by Models 1–7 as a function of the length of the data set and found very strong "ringing" behaviour. This means that the underlying cyclicity is not sufficiently removed from the data. By fitting a damped oscillator, we estimate the underlying trend to be $+40\,\mathrm{kHz/yr}$, which is much stronger that of $+12 \pm 4\,\mathrm{kHz/yr}$ given by Model 1 but only because the model is still oscillating. This indicates that a more sophisticated model is required to determine a trend.

Another obvious feature of Fig. 1 is the consistently larger error whenever 11-year running means of geomagnetic activity were included. This is clear for the Sodankylä *A*k index, but it is also observed in the planetary *A*p. Further inspection shows that all *A*k models lead to slightly larger trend errors than the *A*p models while both fit the data equally well (2nd panel), independent of which index is included or whether it is omitted altogether.

The quality of the fit does, however, depend upon the type of solar activity proxy, as long as unsmoothed monthly values are considered (Models 8–28, 71–91). While the errors of the trend do not vary much within each group of three models, which are identical except for solar activity, the model fit is worst for sunspot numbers, better for adjusted and best for observed radio fluxes, which makes perfect sense since solar radiative energy received on the Earth is modulated by the Earth–Sun distance variation.

The choice of solar activity proxy also affects the trend magnitude (4th panel). Practically all sunspot-based models (8–112) obtain a trend considerably less negative than the radio-flux based models. Inclusion of annual and semi-annual variability does not affect the trend error much, but it does make the fit better.

In order to decide which model is the most suitable for trend determination, we use the "goodness" of the fit as well as the probable error of the slope. By themselves these parameters are not sufficient to make a good decision. This is evident from Fig. 1 when looking at, say, Models 89–91, which fit the data very well (2nd panel), but their slope has an error twice as large as the almost equally well fitting Models 86–88. The reverse argument is true, too: Models 113–130 have small slope errors (3rd panel), but they do not fit the data very well. Therefore we use the product of the probable slope error and the standard deviation of the residual to decide which model suits our foF2 time series best.

In case of unsmoothed foF2 values, the overall best model (76) gives a trend of $-14 \pm 4\,\mathrm{kHz/yr}$. It contains observed monthly radio fluxes, monthly planetary *A*p and seasonal variability, which is consistent with our previous conclusions. When the models are fitted to 13-month running means of foF2, the best model (37) yields a slope of $-14 \pm 4\,\mathrm{kHz}$, too. It consists of 13-month running means of both observed radio fluxes and *A*p, which shows that equal levels of data filtering should be used in a trend analysis. Models containing seasonal variation (71–133) have not been fitted to the filtered foF2, because the filter removes this seasonal variability from the data. Finally, for 131-month running means of foF2, the best model is 61 ($-10 \pm 4\,\mathrm{kHz}$), with observed fluxes and *A*p filtered in the same way.

The trends from Models 37, 61 and 76 as functions of the length of the data set did not exhibit any obvious 11-year ringing, indicating that the solar cycle has been removed effectively. Moreover, fitting the damped oscillator yielded a trend of $-14\,\mathrm{kHz/yr}$ for Models 37 and 76 and $-9\,\mathrm{kHz/yr}$ for Model 61, which agrees with the trend from the multi-parameter fit (Fig. 1).

CONCLUSIONS

In order to determine the long-term trend in Sodankylä foF2, we fitted a selection of 133 physically plausible models to the data. We demonstrated that the trend greatly depends on which model is used. We multiplied the error of the trend with the standard deviation of the residual (data minus fitted model) and used this product to find the best model for our data. We isolated one model for each level of filtering of the original foF2. Two models using monthly averages and 13-month running means thereof yielded trends of $-14 \pm 4\,\mathrm{kHz}$ while the model fitted to 131-month running means of foF2 returned a trend of $-10 \pm 4\,\mathrm{kHz/yr}$. All three models contain the planetary geomagnetic Ap index as well as observed 10.7-cm fluxes, which fit the data better than sunspot numbers or adjusted radio fluxes. When using unfiltered monthly averages, annual and semi-annual variation should be included. The results show that, for best results, the degree of smoothing of the model components must match that of the time series to be studied.

These conclusions apply to Sodankylä foF2. We did not test whether the same models work best for other ionosonde stations, too. Each site is different in latitude, balance of solar and other inputs, influence of thermospheric winds, etc., and therefore it is strongly recommended that for every foF2 observatory the procedure described in this paper be used for finding the best possible model.

ACKNOWLEDGEMENTS

We thank Mrs Mirja Hämäläinen and Mrs Nina Riipi for their excellent work of scaling the Sodankylä ionograms. Th.U. was supported by the Academy of Finland and the Thule Institute, University of Oulu, Finland.

REFERENCES

Brasseur, G., Hitchman, M.H.: Stratospheric response to trace gas perturbations: Changes in ozone and temperature distribution. Science, **240**, 634–637 (1988)

Bremer, J.: Ionospheric trends in mid-latitudes as a possible indicator of the atmospheric greenhouse effect. J Atm Terr Phys, **54**, 1505–1511 (1992)

Bremer, J.: Trends in the ionospheric E- and F-regions over Europe. Ann Geophys, **16**, 986–996 (1998)

Clilverd, M.A., Clark, T.G.C., Clarke, E., Rishbeth, H.: Increased magnetic storm activity from 1868 to 1995. J Atm Sol-Terr Phys, **60**, 1047–1056 (1998)

Clilverd, M.A., Ulich, Th., Jarvis, M.J.: Residual solar cycle influence on trends in ionospheric F2-layer peak height. J Geophys Res, 108, doi:10.1029/2003JA009838 (2003)

Danilov, A.D., Mikhailov, A.V.: Spatial and seasonal variations of foF2 long-term trend. Ann Geophys, **17**, 1239–1243 (1999)

Dudeney, J.R.: The accuracy of simple methods for determining the height of the maximum electron concentration of the F2 layer from scaled ionogram characteristics. J Atm Terr Phys, **45**, 629–640 (1983)

Garriott, O.K., Rishbeth, H.: Effects of temperature changes on the electron density profile in the F2 layer. Planet Space Sci, **11**, 587–590 (1963)

Jarvis, M.J., Clilverd, M.A., Ulich, Th.: Methodological influences on F-region peak height trend analysis. Phys Chem Earth, **27**, 589–594 (2002)

Jarvis, M.J., Jenkins, B., Rogers, G.A.: Southern hemisphere observations of a long-term decrease in F region altitude and thermospheric wind providing possible evidence for global thermospheric cooling. J Geophys Res, **103**(20), 774–20,787 (1998)

Mikhailov, A.V., de la Morena, B.A.: Long-term trends of foE and geomagnetic activity variations. Ann Geophys, **21**, 751–760 (2003)

Mikhailov, A.V., Marin, D.: Geomagnetic control of the foF2 long-term trends. Ann Geophys, **18**, 653–665 (2000)

Ortiz de Adler, N., Elias, A.G., Heredia, T.: Long-term trend of the ionospheric F2-layer peak height at a southern low latitude station. Phys Chem Earth, **27**, 613–615 (2002)

Rishbeth, H.: A greenhouse effect in the ionosphere? Planet Space Sci, **38**, 945–948 (1990)

Rishbeth, H., Edwards, R.: The isobaric F2 layer. J Atm Terr Phys, **51**, 321–338 (1989)

Rishbeth, H., Roble, R.G.: Cooling of the upper atmosphere by enhanced greenhouse gases – Modelling of thermospheric and ionospheric effects. Planet Space Sci, **40**, 1011–1026 (1992)

Rishbeth, H., Sedgemore-Schulthess, K.J.F., Ulich, Th.: Semiannual and annual variations in the height of the ionospheric F2-peak. Ann Geophys, **18**, 285–299 (2000)

Roble, R.G., Dickinson, R.E.: How will changes in carbon dioxide and methane modify the mean structure of the mesosphere and thermosphere? Geophy Res Lett, **16**, 1441–1444 (1989)

Serafimov, K., Serafimova, M.: Possible radio indications of anthropogenic influences on the mesosphere and lower thermosphere. J Atm Terr Phys, **54**, 847–850 (1992)

Sharma, S.S., Chandra, H., Vyas, G.D.: Long-term ionospheric trends over Ahmedabad. Geophys Res Lett, **26**, 433–436 (1999)

Shimazaki, T.: World-wide variations in the height of the maximum electron density of the ionospheric F2 layer. J Radio Res Labs Japan, **2**, 85–97 (1955)

Taubenheim, J., von Cossart, G., Entzian, G.: Evidence of CO_2-induced progressive cooling of the middle atmosphere derived from radio observations. Adv Space Res, 10:(10)171–(10)174 (1990)

Tobiska, W.K., Woods, T., Eparvier, F., Viereck, R., Floyd, L., Bouwer, D., Rottman, G., White, O.R.: The SOLAR2000 empirical solar irradiance model and forecast tool. J Atm Sol-Terr Phys, **62**, 1233–1250 (2000)

Ulich, Th.: Solar variability and long-term trends in the ionosphere. PhD thesis, Sodankylä Geophysical Observatory, Sodankylä, Finland (2000)

Ulich, Th., Turunen, E.: Evidence for long-term cooling of the upper atmosphere in ionosonde data. Geophys Res Lett, **24**, 1103–1106 (1997)

Ulich, Th., Clilverd, M.A., Rishbeth, H.: Determining long-term change in the ionosphere. EOS Trans, **84**, 581, 585 (2003)

Upadhyay, H.O., Mahajan, K.K.: Atmospheric greenhouse effect and ionospheric trends. Geophys Res Lett, **25**, 3375–3378 (1998)

Xu, Z.-W., Wu, J., Igarashi, K., Kato, H., Wu, Z.-S.: Long-term ionospheric trends based on ground-based ionosonde observations at Kokubunji, Japan. J Geophys Res, 109, doi:10.1029/ 2004JA010572 (2004)

CHAPTER 2.4

THERMOSPHERE DENSITY MODEL CALIBRATION

EELCO DOORNBOS

Delft Institute for Earth Observation and Space Systems (DEOS), Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
e.n.doornbos@tudelft.nl

Abstract: Thermospheric density models are a main source of error in the orbit determination and prediction of low Earth satellites. The empirical models that are in wide use today show large systematic errors when compared with data derived from accelerometers and spacecraft tracking. This accuracy limit is inherent in their model formulation, which is based on an imperfect correlation of observed thermosphere density with a limited set of certain space weather proxy indices. It has been demonstrated that a substantially higher accuracy can be reached by model calibration using concurrent observations of satellite drag. Such drag observations can be obtained by processing freely available Two-Line Element (TLE) data, which are used for representing and distributing satellite orbit trajectories. Several aspects of this data processing require specific attention. These include the selection of suitable space objects, determining their ballistic coefficients, and taking into account thermospheric winds and radiation pressure accelerations

INTRODUCTION

Empirical neutral density models of the thermosphere are crucial in applications of satellite orbit determination and prediction. For low orbit satellites, neutral drag is the most important disturbing force. Even if in some situations the other major non-gravitational force, radiation pressure, is slightly larger in magnitude, the effect of drag on the orbit is often dominant, due to its energy-dissipating nature. Drag is also the most difficult force to model, partly due to the lack of accurate data on particle-surface interactions, but mainly because of the complexity of neutral atmosphere variations with solar extreme ultraviolet radiation and geomagnetic heating inputs.

Thermosphere density models are applied in many types of satellite orbit calculations, including re-entry prediction, collision risk analysis, manoeuvre planning for ground-track maintenance and precise orbit determination for geodetic applications (Doornbos and Klinkrad 2005). The accuracy of density models therefore does not only affect scientific results, but also the requirements on operations, tracking systems and propellant consumption of many satellite missions.

J. Lilensten (ed.), Space Weather, 107–114.

EMPIRICAL MODELS

Density models used routinely in orbit determination applications include CIRA-72 (Jacchia 1972), DTM-78 (Barlier et al. 1978), DTM-94 (Berger et al. 1998), MSIS-86 (Hedin 1987) and NRLMSISE-00 (Picone et al. 2002). Each of these models is based on a corresponding database of historical observations, to which parametric equations have been fitted, representing the known thermospheric variations with local time, latitude, season, etc. Although not all models incorporate all data types, the observational databases can consist of density derived from drag computations based on satellite tracking or accelerometer data, in-situ mass spectrometer measurements and incoherent scatter radar measurements. Location and time inputs are used to model the diurnal bulge and semi-annual variations in density. Changes in solar and geomagnetic activity are represented by their proxies $F_{10.7}$ and A_p or K_p with model specific combinations of lag-times, interpolation and smoothing applied.

Despite recent enhancements in model formulation, a better choice of proxies and expanded databases of observations from precise tracking and accelerometers (Picone et al. 2002; Bruinsma et al. 1999, 2004), the accuracy of models for operational use has not significantly improved in over 30 years. An accuracy barrier of approximately 15% is apparently inherent in density models of this type (Marcos 1990). In the absence of significantly better models, many orbit analysts currently still use CIRA-72 or MSIS-86 density models.

Figure 1 shows a comparison of typical model- and tracking-derived densities from the analysis of ERS-2 orbits, with precise tracking from satellite laser ranging

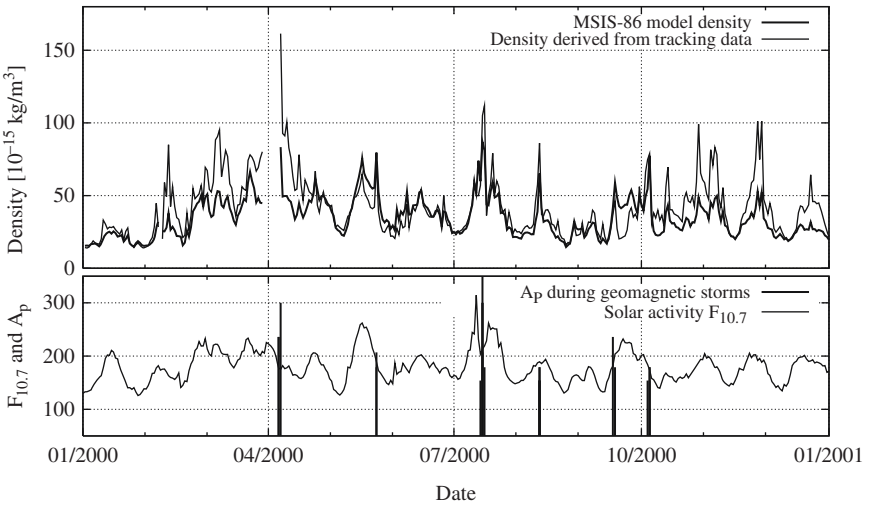


Figure 1. Daily average densities along the ERS-2 orbit (top panel) and solar and geomagnetic indices (bottom panel)

and radar altimetry (Doornbos et al. 2005). During this year of high solar activity, and on the time scales shown here (daily to yearly), the variations associated with the 27-day solar rotation period dominate. The comparison of modelled and measured density shows that the model often under- or overestimates the density by up to a factor of two. The MSIS-86 model was chosen to generate this plot, but a very similar figure is achieved using any of the other above-mentioned models.

THERMOSPHERE DENSITY MODEL CALIBRATION

Density calibration involves the calculation of corrections to existing density models using an ongoing analysis of drag accelerations on a number of low perigee satellites and debris objects. The correction parameters determined from assimilated daily drag data can then largely replace the role of solar and geomagnetic activity indices in driving all irregular variations in density (Marcos et al. 1998). Using this method, an accuracy level well below the 15% barrier can be reached. A feasibility study for the European calibrated density model presented here was inspired by two distinct initiatives that were started by scientists in the space tracking community several years ago.

A US-Russian collaborative density calibration project involved the estimation of daily scale factors, varying linearly with altitude, which multiply the original model density to better represent orbit data. The analysis was performed for both the Russian GOST model (Cefola et al. 2003) and the US NRLMSISE-00 model (Yurasov et al. 2005a). Trajectory information for the calibration objects was taken from publicly available Two-Line Element (TLE) data. TLEs are a way of describing and distributing satellite orbits using a very limited number of parameters. They will be further described in the next section. A statistical analysis, investigating the possibility to predict correction parameters into the future has also been performed (Yurasov et al. 2005b). As expected, the accuracy behaviour of predictions depends on the forecasting interval. While near-term results are significantly improved, forecasts over more than a few days can apparently no longer provide additional density accuracy over uncalibrated models.

Another project, the US Air Force Space Command's Dynamic Calibration Atmosphere (DCA) for the High-Accuracy Satellite Drag Model (HASDM) (Storz et al. 2005), uses the more accurate tracking data from the Space Surveillance Network (SSN), to adjust simultaneously the trajectories of around 75–80 calibration objects with spherical harmonic expansions of two temperature parameters from the Jacchia-70 thermosphere model (Casali and Barker 2002). Within Jacchia's models, these two temperatures define the vertical density profile completely. The spherical harmonic expansion allows corrections to the modelling of the diurnal variation. HASDM also includes a method for predicting the corrections 3 days into the future as a function of solar and geomagnetic indices. A follow-on project, named Sapphire Dragon, is aimed at improving the prediction capabilities of HASDM through a series of enhancements, including an increase in the number of calibration objects and a more sophisticated use of various space weather proxies. Unfortunately, the

resulting calibrated model or the underlying precise SSN tracking data are not made publicly available.

These developments have prompted a feasibility study into a European density calibration project based on satellite data that is openly available for scientific use. One of the conclusions of this study is that since no tracking data with an accuracy and wide availability comparable to that of the SSN are freely available, only TLEs supply information on a large enough number of objects. At least 50 to 100 objects are required to supply sufficient spatial coverage for a low degree and order spherical harmonic adjustment in latitude and local solar time.

Other data sources, such as the accelerometer instruments on CHAMP (Bruinsma and Biancale 2003), GRACE and the future GOCE and SWARM, or tracking techniques such as satellite laser ranging, DORIS (Willis et al. 2005) and GPS (Van den IJssel and Visser 2005) can provide more accurate drag data, and a higher temporal resolution. However, the number and spatial distribution of their data as well as their data latency is too limited compared with TLEs for a near real-time adjustment of a global model (Doornbos et al. 2005). Nevertheless, they can be used as valuable sources of supplementary data for calibration, or for validation of TLE-calibrated models.

PROCESSING OF TLE DATA FOR DENSITY CORRECTIONS

Two-Line Elements consist of a set of orbit parameters that have been fitted to Space Surveillance Network tracking data using a specific orbital model. An estimate of the orbital position at a given epoch can then be obtained by propagating the orbit using a combination of the model and parameters of a nearby TLE (Hoots and Roehrich 1988). TLE sets are currently provided at least once per day for most objects suitable for density calibration, offering an orbit accuracy of a few hundred metres to many kilometres, depending on the magnitude of the perturbation forces acting on the object. Since TLE data are provided without any indication of accuracy, methods such as orbit overlap comparisons, in which the orbit from multiple nearby TLEs are propagated over the same time interval, have to be applied in order to determine data quality.

An efficient algorithm for processing TLE data for density studies has been described by Picone et al. (2005), following a study of long-term change in the thermosphere by Emmert et al. (2004). This algorithm was adopted for the European near real-time calibrated density model. Although the algorithm is much more efficient than older methods, automatically applying it to a large number of objects and creating long time series of accurate density data can still be a labour intensive task.

First of all, a set of suitable objects has to be obtained, so that their TLE data can be downloaded from the Internet. The US Air Force has recently set up a website at www.space-track.org for distribution of TLEs. Slowly tumbling objects with non-spherical shape have changes in cross-sectional area that are difficult to separate from changes in density. This makes these objects unsuitable for calibration. Only

after generating long time-series of data and by comparison with other objects can these be identified and removed. For the remaining objects, the ballistic coefficient, comprising the area, mass and drag coefficient, is often unknown. When these quantities are stable over long time periods, as is often the case, the ballistic coefficient can be determined to within a few percent by assuming that the long-term average of the ratio of observed over modelled density equals one (Bowman 2002; Yurasov et al. 2005a). The more accurate the density model used, the more accurate the ballistic coefficient estimate, so that an iterative scheme can be devised in which a calibrated density model is used in order to find more accurate ballistic coefficients for objects which in turn can improve the calibration (Granholm et al. 2000).

For orbits with a perigee lower than approximately 200 km, the drag perturbations and therefore the error in the TLE-derived positions can become too large. In addition, changes in the aerodynamic flow regime could affect the ballistic coefficient. At the other end of the spectrum, for perigees higher than approximately 400–500 km, depending on the level of solar activity and orbit geometry, either the perturbations by solar radiation pressure become non-negligible, or the TLE data are not sensitive enough to detect the small changes in the orbit due to the

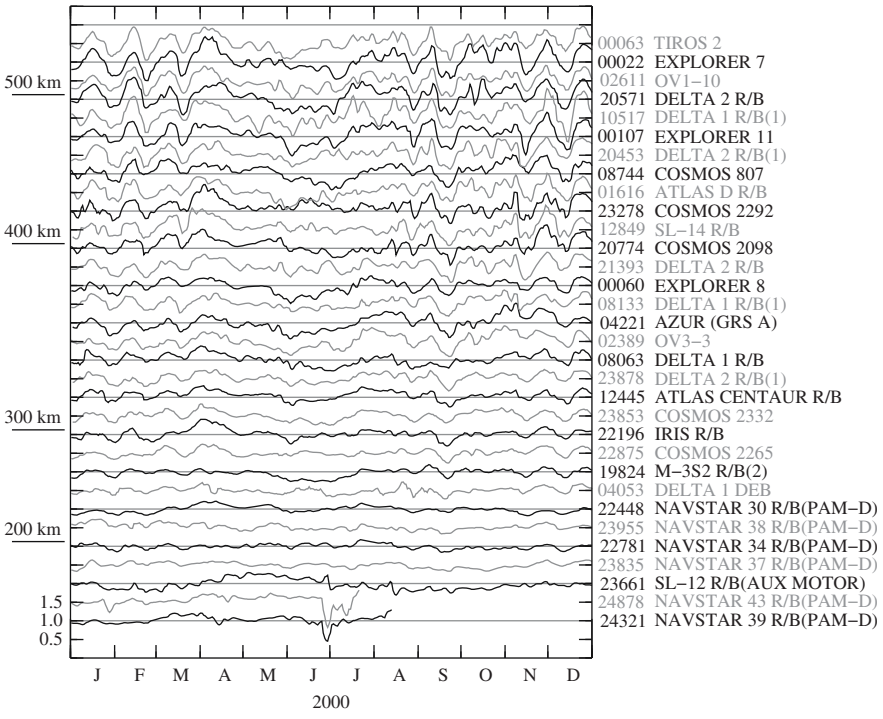


Figure 2. Time series of density ratios (TLE-derived over MSIS-86) for 32 space objects. The objects (satellites and rocket parts) are sorted by their perigee heights in mid-2000, and their time series are offset from one another by 0.5 on the y-axis

remaining drag. Accurate modelling of solar radiation pressure perturbations for subsequent removal is often not possible, as the geometry and optical properties of the spacecraft surfaces are unknown. An order of magnitude estimate of the ratios between the effects of drag and radiation pressure on the orbit must then be used to remove suspect data.

Once the appropriate data editing is complete, the algorithm of Picone et al. (2005) will result in density estimates along the object's trajectory with an accuracy of a few percent, depending on the accuracy of the ballistic coefficient estimate as well as the remaining errors introduced by solar radiation pressure and the influence of neutral thermosphere winds.

Figure 2 shows the results of the algorithm of Picone et al. (2005) applied to TLE data from 32 selected objects. The derived daily densities shown here are for a wide range of perigee locations in the lower thermosphere. The values have therefore been divided by equivalent MSIS-86 model densities so that the resulting ratios could be plotted, with offsets, in a single graph. The large similarity between the time series of density ratios is immediately apparent, as is an increase in amplitude with perigee height. Solar activity-related variations in the thermosphere increase in amplitude with increasing height, as do the errors in the MSIS-86 model. The figure demonstrates why the 2-parameter fit of the US-Russian density correction (Yurasov et al. 2005a) is already able to remove a large part of the model errors. Figures such as this one are also very suitable for identifying problematic time periods and objects in the data.

CONCLUSIONS AND OUTLOOK

Development of neutral density models for use in orbit determination is moving towards the assimilation of near real-time satellite drag data. Two independent projects by Russian and US satellite tracking experts have published accuracies far below the 15% barrier of traditional empirical models. This text has given a brief overview of these developments and discussed TLE processing and data editing aspects of a new European initiative for an automated near real-time model calibration service.

The calibration parameters that are estimated from satellite data can, to a large extent, compensate for the imperfect correlation between space weather proxies and neutral density, which has been a main obstacle for model improvement in the past. This by no means indicates that the role of space weather observations has ended in density model research. The calibration parameters can only be directly determined from satellite tracking for past epochs, while orbit determination applications, such as re-entry analysis and manoeuvre planning, rely on density predictions. The renewed interest in satellite drag data generated for density calibration, combined with space weather data from recent missions, has already sparked new investigations in physical modelling of the thermosphere (Fuller-Rowell et al. 2003), as well as new research into the interaction between various aspects of solar activity and neutral density (Bowman and Tobiska 2006).

REFERENCES

Barlier, F., Berger, C., Falin, J., Kockarts, G., Thuillier, G.: A, thermospheric model based on satellite drag data. Annales de Geophysique, **34** (1), 9–24 (1978)

Berger, C., Biancale, R., Ill, M., Barlier, F.: Improvement of the empirical thermospheric model DTM: DTM-94 – a comparative review of various temporal variations and prospects in space geodesy applications. Journal of Geodesy, **72** (3), 161–178 (1998)

Bowman, B.: True satellite ballistic coefficient determination for HASDM, AIAA/AAS Astrodynamics Specialist Conference and Exhibit, 5–8 August 2002, Monterey, California, AIAA 2002–4887 (2002)

Bowman, B.R., Tobiska, W.K.: Improvements in modelling thermospheric densities using new EUV and FUV solar indices, 16th AAS/AIAA Space Flight Mechanics Conference, Tampa, Florida, January 22–26, 2006, AAS 06–237 (2006)

Bruinsma, S., Biancale, R.: Total densities derived from accelerometer data, Journal of Spacecraft and Rockets, **40** (2), 230–236 (2003)

Bruinsma, S., Exertier, P., Biancale, R.: An assessment of new satellite total density data for improving upper atmosphere models. Planetary and Space Science, **47**, 1465–1473 (1999)

Bruinsma, S., Tamagnan, D., Biancale, R.: Atmospheric densities derived from CHAMP/STAR accelerometer observations, Planetary and Space Science, **52** (4), 297–312, doi:10.1016/j.pss.2003.11.004 (2004)

Casali, S.J., Barker, W.N.: Dynamic calibration atmosphere (DCA) for the high accuracy satellite drag model (HASDM), AIAA/AAS Astrodynamics Specialist Conf. (Monterey, CA), August 2002, AIAA–2002–4888 (2002)

Cefola, P.J., Nazarenko, A.I., Proulx, R.J., Yurasov, V.S.: Atmospheric density correction using two line element sets as the observation data, AAS/AIAA Astrodynamics Specialists Conference, Big Sky, Montana, August 3–7, 2003, AAS 03–626 (2003)

Doornbos, E., Klinkrad, H.: Modelling of space weather effects on satellite drag, Advances in Space Research, Article in press, available online at www.sciencedirect.com, doi:10.1016/j.asr.2005.04.097 (2005)

Doornbos, E., Klinkrad, H., Visser, P.: Atmospheric density calibration using satellite drag observations, Advances in Space Research, **36** (3), 515–521, doi:10.1016/j.asr.2005.02.009 (2005)

Emmert, J.T., Picone, J.M., Lean, J.L., Knowles, S.H.: Global change in the thermosphere: Compelling evidence of a secular decrease in density, Journal of Geophysical Research, **109**, A02301, doi:10.1029/2003JA010176 (2004)

Fuller-Rowell, T., Minter, C., Codrescu, M.: On the use of physics-based models in data assimilation for neutral density specification and forecast, AAS/AIAA Astrodynamics Specialists Conference, Big Sky, Montana, August 3–7, 2003, AAS 03–627 (2003)

Granholm, G.R., Proulx, R.L., Cefola, P.J., Nazarenko, A.I., Yurasov, V.: Requirements for accurate near-real time atmospheric density correction, AIAA/AAS Astrodynamics Specialist Conference, Denver, CO, Aug. 14–17, 2000, AIAA–2000–3932 (2000)

Hedin, A.E.: MSIS–86 thermospheric model, Journal of Geophysical Research, **92** (A5), 4649–4662 (1987)

Hoots, F.R., Roehrich, R.L.: Models for propagation of NORAD element sets, Spacetrack Rep. 3, Aerosp. Def. Command (Available at http://celestrak.com/NORAD/documentation/spacetrk.pdf) (1988)

Jacchia, L.G.: Atmospheric models in the region from 110 to 2000 km, in Stickland, A.C. (ed.) CIRA 1972. Akademie-Verlag, Berlin

Marcos, F.: Accuracy of atmospheric drag models at low satellite altitudes. Advances in Space Research, **10** (3), 417–422 (1990)

Marcos, F., Kendra, M.J., Griffin, J.M., Bass, J.N., Liu, J.J.F., Larson, D.R.: Precision low Earth orbit determination using atmospheric density calibration, Advances in Astronautical Sciences, **97** (1), AAS, 515–527 (1998)

Picone, J., Hedin, A., Drob, D., Aikin, A.: NRLMSISE-00 empirical model of the atmosphere: statistical comparisons and scientific issues. Journal of Geophysical Research, **107** (A12), 1468, doi:10.1029/2002JA009430 (2002)

Picone, J.M., Emmert, J.T., Lean, J.L.: Thermospheric densities derived from spacecraft orbits: Accurate processing of two-line element sets, Journal of Geophysical Research, 110, A03301, doi:10.1029/2004JA010585 (2005)

Storz, M.F., Bowman, B.R., Branson, J.I., Casali, S.J., Tobiska, W.K.: High accuracy satellite drag model (HASDM), Advances in Space Research, **36** (12), 2497–2505 (2005)

Van den IJssel, J., Visser, P.: Determination of non-gravitational accelerations from GPS satellite-to-satellite tracking of CHAMP, Advances in Space Research, **36**(3), 418–423, doi:10.1016/j.asr.2005.01.107 (2005)

Willis, P., Deleflie, F., Barlier, F., Bar-Sever, Y.E., Romans, L.J.: Effects of thermosphere total density perturbations on LEO orbits during severe geomagnetic conditions (Oct–Nov 2003) using DORIS and SLR data, Advances in Space Research, **36**, 522–533, doi:10.1016/j.asr.2005.03.029 (2005)

Yurasov, V.S., Nazarenko, A.I., Cefola, P.J., Alfriend, K.T.: Density corrections for the NRLMSIS-00 atmosphere model, AAS/AIAA Space Flight Mechanics Conference, Copper Mountain, Colorado, January 23–27, 2005, AAS 05–168 (2005a)

Yurasov, V.S., Nazarenko, A.I., Cefola, P.J., Alfriend, K.T.: Application of the ARIMA model to analyze and forecast the time series of density corrections for NRLMSIS-00, AAS/AIAA Astrodynamics Specialists Conference, Lake Tahoe, CA, August 7–11, 2005, AAS 05–256 (2005b)

CHAPTER 2.5

NUMERICAL SPACE WEATHER PREDICTION: CAN METEOROLOGISTS FORECAST THE WAY AHEAD?

M. KEIL
UK Met Office

INTRODUCTION

Space weather prediction lags significantly behind the current weather forecasting capability. Despite the capability gap there are large areas of overlap between these fields. However, these areas have yet to be fully explored or exploited. Many of the issues, questions and techniques related to the analysis and forecasting of an evolving environmental system are similar. In the middle and upper atmosphere the domains of meteorology and space weather coincide and common problems are required to be tackled, albeit from different perspectives.

The emphasis of meteorologists can vary considerably because the subject area is driven by three distinct groups. The majority of the "blue-sky" research is carried out by the academic community, in a similar way to space weather research. A large number of meteorologists are associated with a National Met Service (NMS). The broad purpose of a NMS is to provide weather and climate information for the general public and governments. They often have wider remits to benefit the economy and the environment. They also positively contribute to the global meteorological and scientific communities. There are also a growing number of private sector meteorologists. Their work can range from sub-contracted academic research to providing user-specific forecasts e.g. for off-shore oil platforms.

The main focus of this paper is to concentrate on NMSs and how they could potentially contribute to space weather research and development. Within Europe there is no space weather parallel to, or equivalent of, NMSs. This capability gap could be addressed by working with meteorologists for the benefit of both communities.

National Met Services tend to focus on the production of operational forecasts and applied research in order to fulfil their remit. Most large NMSs produce their own analyses and forecasts on supercomputers in a process known as Numerical

J. Lilensten (ed.), Space Weather, 115–124.

Weather Prediction (NWP), which is illustrated on Fig. 1. A vast and diverse array of observations is combined with a first guess of the atmospheric state in a process known as Data Assimilation. The first guess is usually provided by a previous forecast, indicated by the model in Fig. 1, and is know as the background. Data Assimilation takes the two sources of information (observations and background) and uses them to produce an analysis which is the best estimate of the state of the atmosphere at that time. A good analysis is crucial as it acts as the initial conditions (ICs) for further forecasts. The forecast also provides the background for the next cycle of data assimilation to be performed. The process in Fig. 1 is the cornerstone of operational weather forecasting and central to the work of a NMS. This capability is very expensive and underpins not only the work of a NMS, but also much of the research carried out in academia. The data assimilation process can be applied to other areas such as space weather analysis and forecasting.

This paper considers some of the key scientific and practical issues from the view of the meteorologist. The structure of this paper is as follows: some historical aspects of the development of numerical weather analysis and forecasting and how they relate to space weather analysis and forecasting are explored in Section 2; the lessons the space weather community can learn from issues tackled by meteorologists are highlighted in Section 3; finally, a potential way ahead for space weather forecasting is discussed in Section 4.

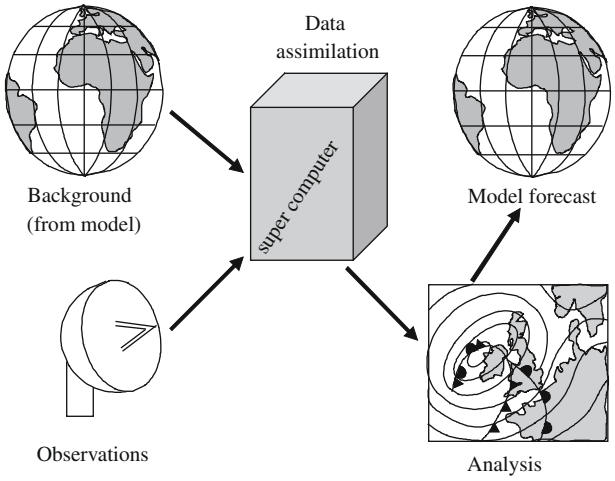


Figure 1. Illustration of the NWP process. A background first guess (from a model forecast) and observations provide information sources for data assimilation to produce an analysis from which a new forecast is produced

NUMERICAL WEATHER PREDICTION (NWP): PAST AND PRESENT

The NWP process is more mature than current space weather analysis and forecasting capability. The development of modern NWP began over a 100 years ago. Much progress was made by Vilhelm Bjerknes; a substantial part of his research was spent investigating how hydrodynamics and thermodynamics could be applied to determine the state of the atmosphere and in 1917 he set up the Bergen School of Meteorology. Bjerknes (1911) referred to weather forecasting as the ultimate problem in meteorology. He broke the problem down into two key components that must be satisfied. The first is that the present state of the atmosphere must be analysed and understood. The second is that laws are specified which govern how future atmospheric states are related to prior atmospheric states. Bjerknes essentially described the NWP process illustrated in Fig. 1; however at the time he did not possess the resources to implement his ideas successfully.

Many of the themes developed by Bjerknes were taken forward by L.F. Richardson. In the period 1916–1918 Richardson simplified Bjerknes's equations (Richardson 1922) and solved them by hand to produce the first weather forecasts based on physical laws that govern the atmosphere. The dates surrounding Richardson's calculations are vague because he was serving as an ambulance driver in World War I at the time and was tackling his meteorological problem in his spare moments. Richardson's first forecast was a major achievement despite turning out to be highly inaccurate. Richardson did not know at the time that the primary cause of the errors was due to poor specification of the initial conditions (Platzman 1967). This acts as a reminder that without good initial conditions even an excellent model will produce a poor forecast.

The advent of computers led to a rapid decrease in the time taken to produce a forecast. It was not until the middle of the 20th century when Charney, Fjortoft and von Neumann (1950) ran the first computer based forecast using the ideas of Bjerknes and Richardson. However, the initial conditions for this forecast were still produced subjectively and it took longer to generate the initial conditions than to run the forecast.

Although there has been no explicit mention of data assimilation, it is apparent that the initial conditions are crucial in producing an accurate forecast. Numerical techniques to combine data from a number of disparate sources have existed for many centuries. Around the same time as the first computer-based weather forecast was run, Panofski (1949) created gridded objective meteorological analysis of observations using polynomial expansion techniques with coefficients found by least squares fit. Gridded analysis of this form are used for initialising numerical weather prediction models.

Further advances were subsequently made in data assimilation. Gilchrist and Cressman (1954) introduced the concept of the "region of influence" in which observations affect a number of grid points within a local area. They also recognised that a previous forecast could be used as a source of information in addition to observations. This is the concept of a background state described in the previous section. Bergothorsson and Doos (1955) introduced the concept of analysing observation

increments – departures between the background and the observations – rather than observations themselves. The idea that regions containing consistent data voids could benefit from information propagated from data rich areas through using a background from a forecast based on a prior analysis was developed by Thompson (1961). The assumption is that over time information from the well-analysed regions propagates into the poorly-analysed regions via the background forecasts.

These ideas make up much of the basic frame work of data assimilation today. The development of more sophisticated numerical models and increased computer power has led to more complex assimilation schemes. Many National Met Services use variational data assimilation (know as "Var"). This method finds the best fit by minimizing the square of the deviation between the analysis and the background/observations in an iterative fashion. The Var procedure allows observations to be assimilated which are not necessarily model variables; thus radiances, as measured by many satellites, can be assimilated directly, rather than relying on secondary derived products. The UK Met Office currently use 4D-Var (Lorenc and Rawlins 2006), in which observations are assimilated at their validity time (as opposed to 3D-Var group, which requires less computer power, where observations within a time window are grouped together and treated as simultaneous). A schematic of 4D-Var can be found in Fig. 2 in which an increment is found at analysis time which produces a new forecast which best fits the observations within a 6 hour time window.

The process underpinning the progress made in the meteorological community over the last hundred years can be summarised by the "virtuous cycle", proposed by Tony Hollingsworth, a former Director of the European Centre for Medium Range Weather Forecasting. The virtuous cycle, illustrated in Fig. 3, contains four linked components of NWP that lead to effective development. Observations are the crucial link to reality, hence they appear at the top of the cycle; without observations there would be no need for expensive data assimilation to be performed. Assimilation and modelling developments are related, better analyses lead to more sophisticated

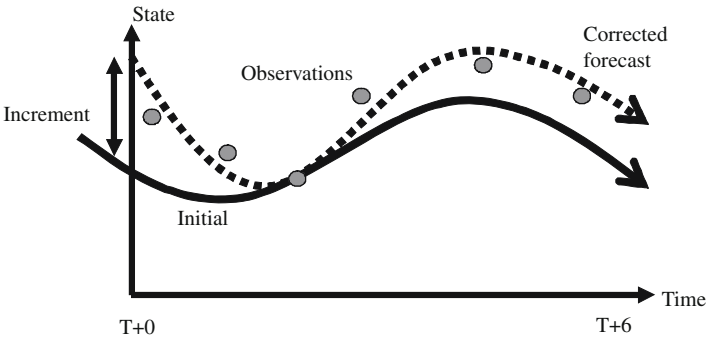


Figure 2. A schematic of 4D-Var assimilation. An increment is applied to an initial background forecast (solid line) to produce a new forecast (dashed line) which best fits the background and the observations (circles) within a 6 hour window

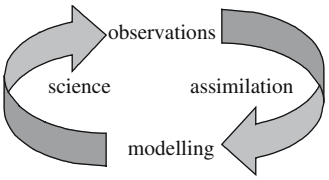


Figure 3. The virtuous cycle leading to effective scientific development

models (and vice-versa). For example work is currently in progress at the UK Met Office to develop assimilation on a 1 km grid. As a result, high resolution convective systems can be represented in the analysis; however for effective forecasting to be performed these high resolution processes need to be included in models. Improved modelling reflects an increased understanding of the underlying science. Models offer a research platform for conducting experiments to test and stretch current scientific knowledge. New scientific theories can be verified through modelling followed by comparison with observations, thus completing the virtuous cycle.

If one of the components is missing, or greatly reduced, overall progress is reduced. This cycle is well established in meteorology, primarily due to the existence of National Met Services which supports observations, assimilation and modelling. However, it can be argued that this cycle is not currently functioning effectively in the space weather community and is subsequently hampering elements of development.

LESSONS LEARNED FROM METEOROLOGISTS

There are two key points to be drawn from the historical development of NWP. Bjerknes identified that a physical modelling approach was necessary to accurately produce a weather forecast based on an analysis of the current situation. This is known as a deterministic approach, as future states are determined by past states. This is the foundation of today's NWP and offers a sensible approach to numerical space weather forecasting development. Empirical models can be used, although they offer very few options for future development as there is a limit to the amount of tuning that can be applied and there is a limit to the range of situations these models can handle. One of the most important aspects of environmental modelling is the ability to capture extreme events. A well-tuned empirical model may capture common situations with a reasonable degree of accuracy, but will ultimately fall short when faced with unusual events. These issues also prevent empirical models from contributing to the virtuous cycle as effectively as physical models.

Secondly, data assimilation techniques can be directly applied to space weather analysis and meteorological research in this area can be exploited for the benefit of the space weather community. Some work in this area has taken place, for example Khattotov et al. (2004) and Schunk et al. (2005). The background state

for data assimilation can come from a number of sources, for example climatology, subjective analysis, or empirical models. For meteorological purposes, the background state is provided from a model forecast from the previous data assimilation cycle. Physical models are used because they organise and evolve previous data in a consistent and realistic way. The background state needs to come from a physically-based model in order to fully exploit the power of data assimilation.

There are other areas of skill and expertise that the meteorology community possess that could benefit the space weather community. National Met Services continually receive and process vast amounts of data from a wide array of sources. NMSs are linked via the Global Telecommunications System infrastructure which allows data to be access and shared quickly. The World Meteorological Organisation (WMO) co-ordinates the global observation system and aims to ensure that sensible strategic decisions are made. The WMO is made up of member NMSs, and is therefore a self-regulating community. Space agencies, such as ESA and NASA, are the closest space weather equivalent to the WMO, although there are significant differences between their aims.

Satellites provide the clearest common observation types between Meteorology and Space Weather. Currently NMSs regularly assimilate data from around 25 satellite instruments, for example Kelly et al. (2004) and English et al. (2004) both highlight the impact from the range of satellite data used. Most of these are operational satellites, which form part of an agreed programme of satellites, providing consistent data in real time. An increasing amount of data is assimilated from research satellites although they essentially provide data as if they were operation satellites and follow WMO criteria regarding timeliness of data.

Global Positioning System (GPS) Radio Occultation (RO) data are an example of observations that can be used for both meteorological and space weather applications. Techniques to retrieve both temperature and humidity profiles from GPS signals were developed in the mid-1990s. The first realistic assimilation of this data was carried out at the UK Met Office (Healy et al. 2005). Plans are in place to assimilate this data in operational weather forecasting models during 2006. The techniques developed by the UK Met Office can be applied to obtain Total Electron Content (TEC) information that could be assimilated in to space weather (ionosphere/thermosphere) models. In the next few years the volume of GPS RO data will increase with the advent of COSMIC (Constellation Observing System for Meteorology, Ionosphere and Climate), which consists of 6 spacecraft measuring TEC and will provide data in near real time for scientific research (Lee et al. 2000). In addition the European Galileo system (Dow 2005) will offer another source of similar observations.

There are other issues that the meteorology community face which are relevant to space weather analysis and forecasting. A crucial aspect for operational weather forecasts is the timeliness of data. Meteorological conditions can change rapidly and data quickly loses its information value. Observations must be taken, transferred, processed and assimilated within a few hours in order for a useful forecast to be produced and disseminated to end-users. However, this issue is more important in

the space weather community than the meteorological community as the situation can change even more rapidly. The whole area of "nowcasting", which is the analysis of the current situation and predicting a few hours further into the future, could offer valuable techniques for space weather forecasting. Other relevant areas include how to effectively take biases in satellite data into account when assimilating and assessing the value of assimilating raw satellite data rather than retrieved products. These issues should be considered by the space weather forecasting community in order to prevent the unnecessary duplication of work.

THE FUTURE

Many operational weather forecasting models have upper boundaries which push beyond the stratosphere. The reason for extending the vertical domain is to improve the way satellite data is exploited by increasing the number of satellite channels and to model the upper atmosphere more accurately. The UK Met Office have recently moved to an operational system with 50 vertical levels, with a top at around 63 km and have plans to increase the vertical resolution to 70 levels. Recent comparisons between 38 and 50 level systems have shown that increasing the vertical resolution has a bigger impact on tropospheric weather forecasts than increasing the horizontal resolution. In addition to an operational model that spans the stratosphere, the UK Met Office run a research model which extends into the mesosphere and has an upper boundary of 86 km. Other centres' research models extend higher, for example the Canadian Middle Atmosphere Model (Polavarapu et al. 2005).

A natural consequence of extended vertical domains is that meteorologists will become interested in space weather assimilation and modelling issues. Most of this paper has focussed on how meteorology and NWP can help with space weather assimilation and modelling; however, there is much that the space weather community can teach the meteorology community, e.g. how to model, or parameterise, the effect of magnetic fields in the upper atmosphere. Science in the crossover domain can be advanced most effectively with both communities working together. A logical strategy should be to have a joined-up approach to common issues and challenges.

There is a growing requirement for commercial operational space weather forecasting capabilities (Lilensten and Bornarel 2006), with many of the applications highlighted in other papers within this book. In addition, there are operational requirements from various areas of within governments including public health and defence. As our reliance on technology susceptible to space weather affects increases, so our need for space weather analysis and forecasting increases. It is logical to assume that fully operational centres dedicated to numerical space weather prediction will be established. Partnerships with National Met Services are an attractive and sensible route forward when establishing such centres.

Section 2 of this paper described how the field of meteorology presents a path for development of numerical space weather prediction. In addition, NMSs offer practical support through their facilities and infrastructure for operational space

weather prediction centres. Most NMSs have access to a supercomputer on which their NWP is performed; these supercomputers could also be used for numerical space weather prediction. NMSs have connectivity to the Global Telecommunications System and this infrastructure could easily be adapted to receive additional space weather related observations. Weather forecasting is a 24/7 operation and this functionality would be necessary for operational space weather forecasting. Finally, a key element of the work of a NMS is to produce user applications and to disseminate relevant information in a timely manner. Similar methods would be required to fully exploit space weather forecasts.

If governments choose to support operational space weather centres the most cost effective way of achieving their goal is to utilise the underpinning capabilities and structures of existing NMS. An example for working in this manner exists in the field of operational oceanography in the United Kingdom. The National Centre for Ocean Forecasting (NCOF) is a strategic partnership between the UK Met Office and the Proudman Oceanographic Laboratory, Plymouth Marine Laboratory, National Oceanography Centre, Southampton and the Environmental Systems Science Centre at Reading. NCOF's mission is to establish ocean forecasting as part of the national infrastructure, based on world-class research and development. The UK Met Office is the home of NCOF where the infrastructure and skills are exploited to produce operational ocean forecasts. The modelling and research expertise of the various partners is coupled with the assimilation expertise and facilities at the Met Office. In a similar manner operational space weather forecasts could be produced by utilising the strengths of NMSs and fusing them with research and modelling skills from the space weather community. This idea has been fully embraced in the United States where the Space Environment Centre is run by NOAA, which is the organisation responsible for the US National Weather Service.

Like many environmental factors, the influence of space weather is independent of national boundaries. Co-ordination on a multinational or continental level would offer a sensible approach to establishing an operational space weather centre within Europe. Potential agencies for spearheading such action include the European Space Agency (ESA), or the European Environmental Agency. A parallel organisation exists for meteorology through the European Centre for Medium-Range Weather Forecasting. A space weather equivalent would bring significant benefits to the European space weather community and beyond.

SUMMARY AND CONCLUSIONS

The emergence of numerical weather forecasting over the last hundred years suggests a clear development path for space weather prediction. Crucial elements include the production of an accurate analysis of the current state of the atmosphere; this analysis subsequently provides the initial conditions for a physically-based model to determine future states of the atmosphere. These two factors lie at the heart of modern meteorological NWP.

Meteorological data assimilation allows expensive observations to be fully exploited and can be applied to space weather analysis. Some observations types, such as Global Positioning System Radio Occultation measurements, are common to both meteorology and space weather domains and similar data assimilation techniques can be used. Much of the data assimilation expertise that has been developed over many years within National Met Services could be transferred to other areas, such as space weather.

The need for operational space weather analysis and forecasting capability is increasing. The infrastructure required for operational space weather (24/7 capability, observation supply, supercomputing, data dissemination, etc.) already exist in the meteorological community. This capability could be exploited through the creation of operational space weather centres working in partnerships with National Met Services. Examples of co-operation of this kind exist in operational oceanography in the UK through the National Centre for Ocean Forecasting and, more directly, in the USA through the Space Environment Centre which is part of the National Oceanographic and Atmosphere Administration which runs the US National Weather Service. Central co-ordination from an organisation, such as ESA, would be necessary for the establishment of single European Space Weather Centre.

As meteorological models increase their vertical domain and as the need for numerical space weather forecasting increases, the meteorology and space weather communities will be drawn closer together. The common ground will become larger as the domains overlap and lessons can be learned from both communities. The challenge is to work together by breaking down organisational, academic and funding barriers in order to further space weather research and development in the most effective way.

REFERENCES

Bergthorsson, P., and Doos, B.: Numerical weather map analysis. Tellus, 7, 329–340 (1955)

Bjerknes, V.: Dynamical Meteorology and hydrography. Part II Kinematics. Gibson Bros, New York: Carnegie Institute (1911)

Charney, J., Fjortoft, R., von Neumann, J.: Numerical integration of the barotropic vorticity equation. Tellus, 2, 237–254 (1950)

Dow, J.M., Neilan, R.E., Gendt, G.: The International GPS Service: Celebrating the 10th anniversary and looking to the next decade. Advances in Space Research, **36** (3), 320–326 Sp Iss 2005 (2005)

English, S., Saunders, R., Candy, B., Forsythe, M., Collard, A.: Met Office Satellite Data OSEs. In Bottger, H., Menzel, P., Pailleux, J. (eds) Proceedings of the Third WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction, World Meteorological Organisation, WMO/TD 1228, pp 146–156 (2004)

Gilchrist, B., and Cressman, G.: An experiment in objective analysis. Tellus, 6, 309–318 (1954)

Healy, S.B., Jupp, A.M., Marquardt, C.: Forecast Impact Trial with GPS radio occultation measurements. Geophys Res Lett, **32** (3), L0380410.1029/2004GL020806 (2005)

Kelly, G., McNally, T., Thepaut, J.-N., Szyndel, M.: OSEs of all main data types in the ECMWF operational system. In Bottger, H., Menzel, P., Pailleux, J. (eds) Proceedings of the Third WMO Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction, World Meteorological Organisation, WMO/TD 1228, pp 63–94 (2004)

Khattotov, B., Murphy, M., Cruikshank, B., Fuller-Rowell, T.: Ionospheric Corrections from a Prototype Operational Assimilation and Forecast System. In Proceedings of the IEEE Position, Location and Navigation Symposium, (PLANS), Monterrey CA April 26–29, 2004 (2004)

Lee, L.C., Rocken, C., Kursinski, E.R.: Applications of the Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC). Terrestrial Atmospheric and Oceanic Sciences, **11** (1), I–III (2000)

Lilensten, J., Bornarel, J.: Space Weather, Environment and Societies. Springer, The Netherlands (2006)

Lorenc, A.C., Rawlins, F.: Quart. J. Roy. Met. Soc., Accepted for publication (2006)

Panofski, H.: Objective weather map analysis. J Appl Met, **6**, 386–392 (1949)

Platzman, G.: A retrospective view of Richardson's book on weather predictions. Bull Am Met Soc, **60**, 302–312 (1967)

Polavarapu, S., Ren, S., Rochon, Y., Sankey, D., Ek, N., Koshyk, J., Tarasick, D.: Atmosphere-Ocean, **43** (1), 77–100 (2005)

Richardson, L.F.: Weather Prediction by numerical process. Cambridge University Press, Cambridge (1922)

Schunk, R., Scherliess, L., Sojka, J.J., Thompson, D.C., Zhu, L.: Ionospheric weather forecasting on the horizon – Models of the ionosphere using state-of-the-art data assimilation techniques are nearing operational use. Space Weather, **3** (8): Art. No. S08007 (2005)

CHAPTER 3.0

IONOSPHERE/POSITIONING AND TELECOMMUNICATIONS

SANDRO M. RADICELLA

INTRODUCTION

Ionosphere shows, at all latitudes and longitudes, very complex dynamical behaviors that define its weather like features. Part of this ionosphere weather is originated by internal processes but a large fraction of it is driven by solar and magnetosphere conditions and can be considered an important component of Space Weather. Ionosphere weather is defined as the inter-hourly, hour-to-hour, day-to-day and week-to-week variability and it has important effects on both communications and satellite positioning and navigation systems operation. Its observation, specification and prediction are key scientific objectives of space weather physics.

The ionosphere is very responsive to solar activity and geomagnetic storms. At middle latitudes the peak electron density in the ionosphere can be strongly altered by the occurrence of solar and geomagnetic processes in a rather intricate way. Weather disturbances at high latitudes occur during geomagnetic storms and substorms as well as during changes in the interplanetary magnetic field. Ionosphere weather features at those latitudes include propagating plasma patches, polar cap arcs and ionosphere convection vortices. Ionosphere changes at equatorial low latitudes are particularly sensitive to electro-dynamical phenomena and the effect of geomagnetic storms at those latitudes is not clearly understood. Variations of the equatorial plasma drifts affect the development and strength of the so-called equatorial electron density anomaly. In turn, plasma drifts are influenced by sudden magnetospheric electric fields. As a result, the ionosphere at low latitudes is highly variable particularly between sunset and midnight and the presence there of irregularities of different scales produces scintillation of satellite signals and spread of the ionogram echoes.

Ionosphere weather has a variety of effects on communication and satellite positioning and navigation systems. HF communications are controlled by the ionosphere behavior and respond to space weather phenomena like solar flares and geomagnetic storms. Severe disturbances in the ionosphere will affect the

J. Lilensten (ed.), Space Weather, 125–127.

propagation of HF radio waves by modifying the usable frequencies and the plasma irregularities may result in signals traveling in more than one path producing radio interference and other communication difficulties. At frequencies above 30 MHz, unexpected reflections of the radio waves by the ionosphere may also cause radio interference. The performance of single-frequency GNNS operation can be considerably degraded by the ionosphere propagation delays changes due to severe conditions of the ionosphere weather affecting satellite positioning. Also differential and real time precise positioning can be influenced by weather changes in the ionosphere. If the electron density along a signal path from a satellite to a receiver varies very rapidly, due to the presence of irregularities, it results in rapid change in the phase of the radio wave that may cause difficulties at GNNS signal receiver level, in the form of "loss of lock". Temporary loss of lock results in "cycle slip", a discontinuity in the phase of the signal.

The articles of this chapter describe the research status in Europe in some key areas related to ionosphere weather and its effects on telecommunications and positioning. Two of them cover topics more geophysical in nature: short-term forecast of electron density peak (represented by foF2), geomagnetic storms effects on the ionosphere over Europe. The other three touch on more technology related subjects: recent improvements in HF ionosphere communication and direction findings systems, the influence of ionosphere weather on GNSS satellite navigation and precise positioning, and effects of scintillations in GNSS operation.

On the contributions dealing with geophysical issues, Mikhailov et al. present their views on the problem of the short-term forecast of foF2, a measure of the peak electron density in the ionosphere. They define as short-term forecast the prediction of the ionosphere parameter value from 1 to 24 hours in advance. The authors concentrate their effort in defending their own approach to the problem that is based on a semi-empirical method that combines theoretical and empirical methods, in contrast with approaches based on neural networks. The method presented by the authors is based on FoF2 relationship with geophysical indices. They discuss the efficiency of the available indices.

In addition Buresova et al. give a well-structured analysis of the ionosphere response to strong to severe geomagnetic storms over Europe, using data from stations located from 38° to 70 ° N. At F2 layer level they concentrate on the occurrence frequency of positive and negative phases of the storm effects. With reference to possible model reproducibility (using IRI storm model) and positive or negative phase predictability of the storm effect the authors admits that still serious difficulties exist. An interesting contribution of their work is the study of the geomagnetic storm effect on the electron density at different heights in the F region of the ionosphere. This type of analysis could help understanding the overall storm effect behavior in the ionosphere and opens a new avenue of research. Previous results by the authors on the storm effects in the F1 layer are also mentioned and they analyze too some results of an Arctic ionosphere scintillations observation campaign related to storm effects.

In the line of communication and positioning application oriented articles, Bertel et al. give a well ordered presentation about new improvements in HF digital

communications and direction finding systems. They analyze recent results of ground-to-ground HF links, including the ionosphere models and prediction tools. They show that is possible to increase the transmission bit rate with respect to what is considered currently possible. They present also propagation channel model that take into account ionosphere effects like polarization, Doppler shift and time delay as well as type of antennas used. Several experiments are described and analyzed indicating the outstanding progresses reached recently in the field of HF digital communications by the authors.

With reference to ionosphere weather effects on satellite navigation and positioning, Warnant et al. give a clear contribution to the understanding of the problem. They start by describing the basic GNSS observables and the effect of the ionosphere on the satellite signals. They describe briefly the total electron content variability and analyze in details the effect of space weather (ionosphere weather) on two GNSS applications: real-time differential positioning and precise relative positioning with real time kinematic technique. The authors call the attention on the fact that ionosphere weather effects on GNSS operation depend strongly on the type of application. Differential positioning is mainly influenced by total electron content gradients observed at high solar activity during severe geomagnetic storms. The largest errors in positioning are observed in the low latitude regions where they can reach 25 m, being smaller at middle latitudes with the errors up to 6 m. On the other hand precise relative positioning is affected by ionosphere small-scale structures (generating what are known as Traveling Ionosphere Disturbances). These structures introduce errors that can reach several decimetres.

Finally Beniguel and Adam present a comprehensive analysis of the GNSS signal scintillation observed as effect of the propagation of radio waves through ionosphere irregularities at low latitudes. Their study is based on data obtained during two measurements campaigns in Cameroon and Brazil. The African campaign included the measurements of the signals from a geostationary satellite that are affected only by ionosphere movements. Their experimental results have been compared with predictions from a scintillation model developed by the authors obtaining reasonable agreement between experimental and modeled data. The authors analyze the behavior of the amplitude scintillation index and calculate the probability of loss of lock and the positioning errors as a function of the scintillation index observed values. They have also calculated the spatial extent of the irregularities region finding values of few hundreds kilometers.

CHAPTER 3.1

SPACE WEATHER INFLUENCE ON SATELLITE-BASED NAVIGATION AND PRECISE POSITIONING

R. WARNANT, S. LEJEUNE AND M. BAVIER
Royal Observatory of Belgium, Avenue Circulaire, 3, B-1180 Brussels, Belgium
R. Warnant@oma.be

Abstract: Global Navigation Satellite Systems (GNSS) are widely used to measure positions with accuracies ranging from a few mm to about 20 m. The effect of the Earth ionosphere on GNSS signal propagation is one of the main error sources which limits the accuracy and the reliability of GNSS applications. In particular, disturbed Space Weather conditions can be the origin of strong variability in the ionosphere Total Electron Content (TEC) which itself degrades the accuracy of GNSS applications. Space Weather effects on GNSS depend very much on the type of application. In this paper, we discuss the effects of Space Weather conditions on differential positioning with the Differential GPS (DGPS) technique and on relative positioning with the Real Time Kinematic (RTK) technique. We show that DGPS is affected by medium to large-scale gradients in TEC mainly observed at solar maximum when RTK will be degraded by smaller-scale ionospheric variability due to scintillations, TEC noise-like behaviour and Travelling Ionospheric Disturbances

FROM NAVIGATION TO HIGH PRECISION GEODESY

Nowadays, Global Navigation Satellite Systems (GNSS) are used to measure positions in the frame of a wide variety of applications. In addition, the deployment of the European Galileo constellation which will be fully operational in 2010 should be the origin of a new increase of the number of GNSS users.

The effect of the ionosphere on GNSS signal propagation is one of the main factor which limits the accuracy and the reliability of GNSS applications. In particular, Space Weather conditions can be the origin of strong variability in the ionospheric plasma which itself degrades the quality of positions measured by GNSS. The influence of the ionosphere on GNSS applications depends very much on the type of positioning technique used.

The different applications of GNSS can be classified in several categories:

- real-time or post-processing: in real-time mode, the user needs to know his position immediately after the measurement; in post-processing mode, there is a delay between the measurements and the moment where the results are available;

J. Lilensten (ed.), Space Weather, 129–146.

- absolute, differential or relative positioning: in absolute mode, the observer measures his absolute position with only one receiver; the differential mode is a particular case of the absolute mode – the observer still wants to measure his absolute position with only one receiver but he makes use of differential corrections broadcast by a reference station. These corrections allow to improve the quality of the measured positions. In relative mode, the observer combines the measurement collected by at least 2 receivers. The absolute position of one of these 2 receivers must be known. Based on the combined measurements, it is possible to compute the vector (often called the baseline) between the 2 receivers. Then, the absolute position of the second receiver can be obtained.
- type of observable: two types of measurements can be performed on GNSS signals – code or carrier phase measurements. Precise applications (cm-level) are always based on carrier phase measurements when navigation mainly uses code measurements which can provide a meter-level accuracy.

At the beginning of the 80's, only two main positioning techniques were available:

- absolute positioning in real-time with code measurements; at that time, the accuracy was ranging between 10 and 100 m (in the horizontal component). This technique was used for navigation.
- relative positioning in post-processing mode with carrier phase measurements for high accuracy applications (a few centimetres) in geodesy.

At the present time, there is a wide variety of positioning techniques which allow to reach a nearly continuous spectrum of accuracies ranging from a few millimetres (high precision geodesy in post-processing mode with phase measurements) to about 20 m (navigation with code measurements). The ionospheric and Space Weather conditions responsible of degradations in GNSS accuracy are very different depending on the positioning technique. In particular, applications which require real-time results are much more sensitive to Space Weather conditions.

In this paper, we address ionospheric and Space Weather effects on differential navigation based on the Differential GPS (DGPS) technique and on precise real-time positioning with the so-called Real Time Kinematic (RTK) technique. Both are real-time positioning techniques. In addition, as the US Global Positioning System (GPS) is, at the present time, the most widely used GNSS, we shall discuss Space Weather effects on GPS to fix ideas. Our discussion is nevertheless representative of Space Weather effects on other GNSS (GLONASS, Galileo) and other positioning techniques.

GNSS OBSERVABLES

GPS signal is composed of 2 ranging codes which are modulated on 2 carrier frequencies called L1 (1575.42 Mhz) and L2 (1227.6 Mhz). The basic GPS observables are measures of the geometric distance between a given GPS satellite and a receiver. They are often called pseudo-range measurements due to the fact that they

do no represent exactly the geometric distance between the satellite and the receiver. Indeed, they are affected by different error sources: satellite and receiver clock synchronisation errors, atmospheric effects, multipath effects, …These pseudo-range measurements can be made based on both code or carrier signals. GPS receivers are able to generate reference code and carrier signals which are in phase with the signals emitted by GPS satellites. They can measure the time delay which exists between the code received from a satellite and their own reference code: this time delay is a measure of the code signal propagation time (and consequently the distance) between the satellite and the receiver. Similarly, GPS receivers can measure the phase difference between the carrier signal received from a satellite and their own reference carrier signal. This measurement also depends on the signal travel time. Nevertheless, phase measurements are ambiguous – they measure the satellite-to-receiver distance modulo an integer number of wavelengths. GPS signals wavelength is about 20 cm.

In practice, if we neglect multipath effects, the simplified mathematical model of code and phase measurements made by receiver p on satellite i, respectively P_p^i (in meters) and φ_p^i (in cycles), can be written as follows (Seeber 2003; Leick 2004):

$$P_p^i = D_p^i + T_p^i + I_p^i + c\left(\Delta t^i - \Delta t_p\right) \tag{1}$$

$$\varphi_p^i = \frac{f}{c}\left(D_p^i + T_p^i - I_p^i + c\left(\Delta t^i - \Delta t_p\right)\right) + N_p^i \tag{2}$$

with:

D_p^i, the geometric distance between receiver p and satellite i;
I_p^i, the ionospheric error;
T_p^i, the tropospheric error;
Δt_p, the receiver clock error (the synchronisation error of the receiver time scale with respect to GPS time scale);
Δt^i, the satellite clock synchronisation error (the synchronisation error of the satellite time scale with respect to GPS time scale);
N_p^i, the phase ambiguity (integer number).

Let's add that GPS satellites also broadcast an ephemeris message which contains a model allowing to compute the position of all GPS satellites with an accuracy of 1–2 m. From equation (1) and (2), it can be seen that the observation equations of code and carrier phase measurements are very similar except that:

- phase measurements are ambiguous;
- the effect of the ionosphere has a different sign: this is due to the fact that the ionosphere is a dispersive medium; carrier signals propagate at phase velocity when code signals propagate at group velocity.

In addition, the precision of phase measurements is much better (around 1 mm) than the precision of code measurements (around 1 meter). More details on GPS observables can be found in Seeber (2003), Leick (2004) and Hoffman-Wellenhof et al. (2001).

EFFECT OF THE IONOSPHERE ON GNSS SIGNALS

Range Error Due to the Ionosphere

The ionosphere is a dispersive medium. The phase n_p and group n_g refraction index can be written (Seeber 2003):

$$n_p = 1 - 40.3\frac{N_e}{f^2} + \frac{c_1}{f^3} + \frac{c_2}{f^4} \tag{3}$$

$$n_g = 1 + 40.3\frac{N_e}{f^2} + \frac{c_3}{f^3} + \frac{c_4}{f^4} \tag{4}$$

with N_e, the free electron concentration (in $e^- m^{-3}$), f, the signal frequency in Hz and c_i, coefficients independent of signal frequency. As carrier frequencies used by GNSS are larger than 1 GHz, we shall neglect the term in f^{-3}, f^{-4}, …

The range error due to the ionosphere can be easily computed. If we neglect path bending effects, the difference (due to the ionosphere) between the measured range using code observations and the geometrical range (in vacuum) from satellite i to receiver p is given by:

$$I_g = \int_p^i n_g \mathrm{d}s - \int_p^i \mathrm{d}s = \int_p^i (n_g - 1)\mathrm{d}s \tag{5}$$

Using the expression of n_g given in equation (4):

$$I_g = \frac{40.3}{f^2}\int_p^i N_e \mathrm{d}s \tag{6}$$

In same way, the range error in carrier phase measurements is given by:

$$I_p = \int_p^i \left(n_p - 1\right)\mathrm{d}s = -\frac{40.3}{f^2}\int_p^i N_e \mathrm{d}s \tag{7}$$

We can see that:

$$I_g = -I_p \tag{8}$$

For this reason, we will omit subscripts g and p:

$$I \equiv I_g = -I_p \tag{9}$$

We define the ionosphere Total Electron Content (TEC):

$$\mathrm{TEC} = \int_p^i N_e \mathrm{d}s \tag{10}$$

Using this definition,

$$I = \frac{40.3}{f^2}\mathrm{TEC} \tag{11}$$

The Total Electron Content is measured in $e^{-}m^{-2}$ or in TEC Units (TECU) with $1\,\mathrm{TECU} = 10^{16}\,e^{-}m^{-2}$. From this equation, it can be seen that a TEC of 1 TECU is responsible of a range error of about 16 cm for the L1 frequency.

Let's assume that the ionosphere can be modelled as a layer of infinitesimal thickness (Fig. 1) at a mean height, h_{iono} (between 350 and 400 km). The intersection between the satellite line of sight and this infinitesimal layer is called the ionospheric point (IP). The relationship between slant TEC and vertical TEC (VTEC) is given by:

$$\mathrm{TEC} \equiv \frac{\mathrm{VTEC}}{\cos Z_{IP}} \tag{12}$$

with $\cos Z_{IP}$, the cosine of the satellite zenith angle measured at the ionospheric point.

TEC Variability

The TEC which depends on the ionosphere electron concentration is the key parameter which influences GNSS signals. In particular, TEC spatial variability on different scales will affect differential and relative applications. The TEC is

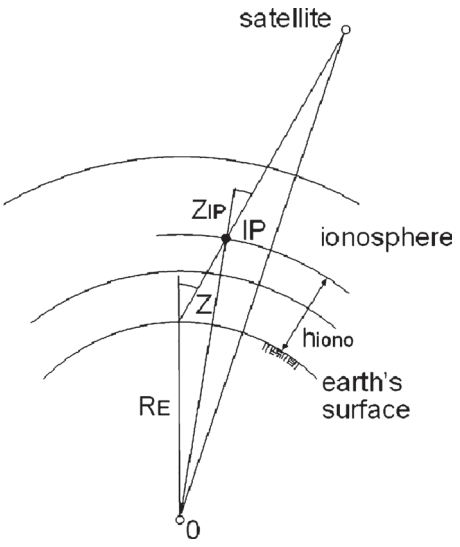


Figure 1. Definition of the ionospheric point

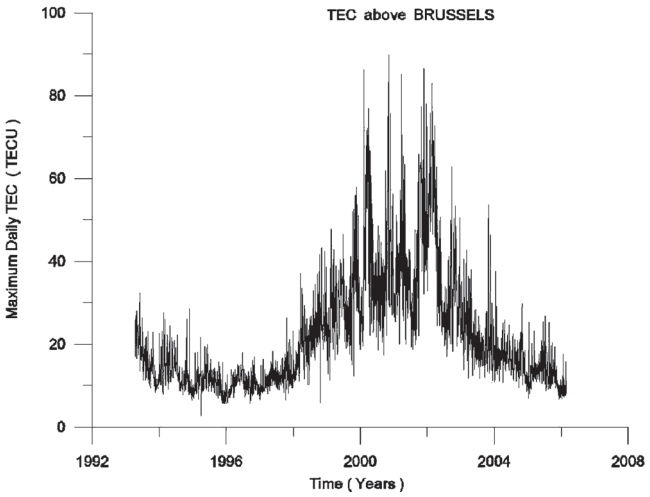


Figure 2. Maximum daily TEC at Brussels from April 1993 to February 2006

very variable in space and time and depends on local time, on the season, on the region and on Space Weather conditions (Solar activity, geomagnetic activity, …). Warnant et al. (2000) have developed a technique allowing to reconstruct TEC and to detect ionospheric smaller-scale disturbances along the track of all the observed satellites using a the so-called geometric free combination of GPS dual frequency measurements. In this paper, this technique is used to characterize ionospheric conditions which are the origin of degradations in GNSS positions. Figure 2 illustrates TEC dependence on solar activity: it shows the maximum daily TEC value observed at Brussels from April 1993 to February 2006. This time series has been

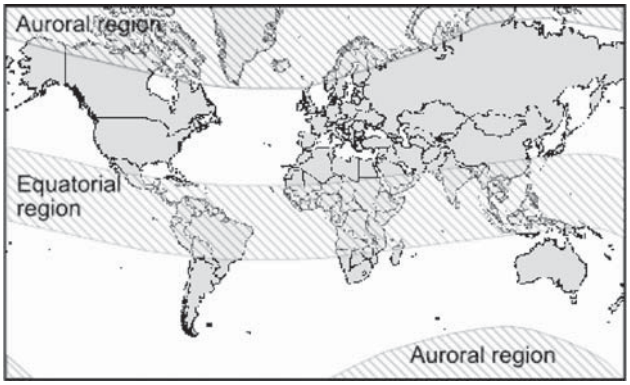


Figure 3. The regions of the ionosphere

obtained at the Royal Observatory of Belgium from GPS measurements using the above mentioned reconstruction technique. This is the longest GPS-TEC time series in Europe.

The ionosphere is usually divided in 3 regions (Fig. 3): the equatorial region (a belt of ± 30 degrees around the magnetic equator), the auroral or polar region and the mid-latitude region which is the region between the polar and equatorial regions. The highest TEC variability is observed in the equatorial and polar regions where ionospheric scintillations are regularly observed (Béniguel et al., same volume). The largest TEC values and the largest TEC gradients are observed in the equatorial region: TEC values of more than 200 TECU and North-South TEC gradients of up to 30 TECU/100 km can be observed at solar maximum (Skone et al. 2002; Wanninger 1993).

On the one hand, differential navigation is particularly affected by larger-scale TEC gradients. On the other hand, precise positioning is sensitive to smaller-scale variability in TEC. Such a variability can be induced mainly by 3 types of phenomena: Travelling Ionospheric Disturbances (TID's), scintillations or "noise-like" variability in TEC. TID's appear as waves in the ionosphere free electron concentration (and consequently in TEC) which are due to interactions between the neutral atmosphere and the ionosphere. They have wavelengths ranging from a few km to more than thousand km. The smaller-scale TID's can be the origin of TEC variability of the order of 1 TECU/min even on distances of a few km (Warnant et al. (2006–1), Warnant et al. (2006–2)). Figures 4a and 4b show fluctuations in vertical TEC (in TEC/min) due to TID's detected using our TEC reconstruction technique on day of year (DOY) 359 in 2004 at Brussels on the track of satellites 17 and 21. Scintillations are fluctuations in phase and amplitude

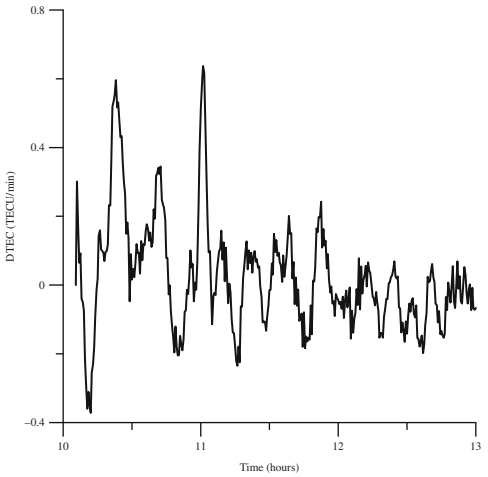


Figure 4a. TEC variability (in TECU/min) due to a TID detected at Brussels on DOY 359 in 2004 along the track of satellite 17

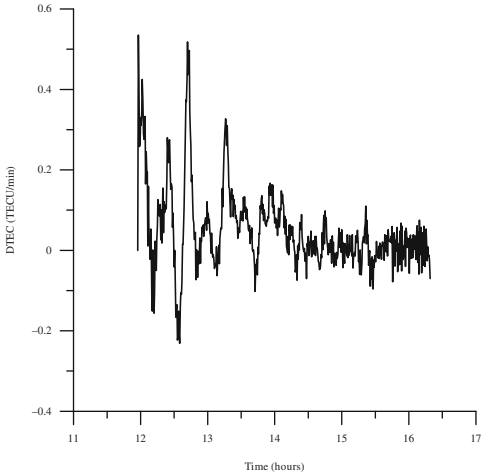


Figure 4b. TEC variability (in TECU/min) due to a TID detected at Brussels on DOY 359 in 2004 along the track of satellite 21

of GPS signals which are due to the presence of small-scale irregularities in the electron concentration (Béniguel et al., same volume). They can severely degrade or even prevent RTK positioning. Scintillations are observed in the polar and in the equatorial ionosphere. In mid-latitude stations like Brussels (51°L), "noise-like" variability in TEC can also be observed. Figures 5a and 5b show vertical TEC variability detected at Brussels on day of year 324 in 2003 on the track of satellites

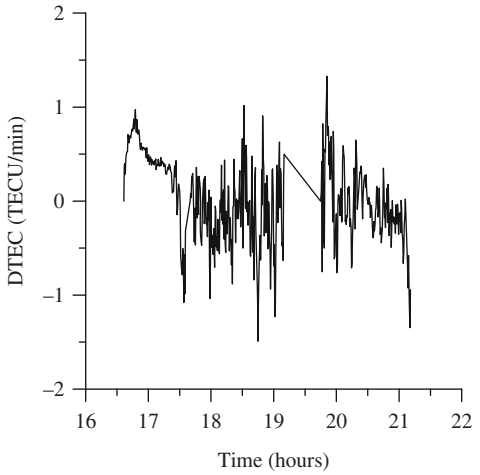


Figure 5a. Noise-like behaviour in TEC observed during a severe geomagnetic storm at Brussels on DOY 324 in 2003 along the track of satellite 15

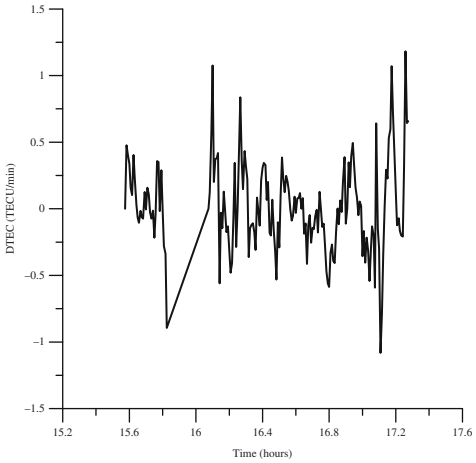


Figure 5b. Noise-like behaviour in TEC observed during a severe geomagnetic storm at Brussels on DOY 324 in 2003 along the track of satellite 16

15 and 16. On that day, a severe geomagnetic storm ($Kp = 9$) was observed. During this storm, vertical TEC variability up to 2.5 TECU/min was detected.

SPACE WEATHER EFFECT ON REAL-TIME DIFFERENTIAL AND RELATIVE APPLICATIONS

In this paper, we focus on 2 particular GNSS applications: navigation with the so-called Differential GPS (DGPS) technique and precise relative positioning with the RTK technique. The study of Space Weather effects on these applications is, in practice, representative of most of the problems which can be encountered when using GNSS under disturbed Space Weather conditions.

Differential Navigation with DGPS

The basic concept of differential positioning is the following: a mobile observer wants to measure his absolute position using only one GNSS receiver but he improves the quality of his computed position by making use of differential corrections broadcast by a reference station of which the position is well-known.

Let's consider the case of differential positioning with code measurements, the so-called DGPS technique. We assume that the mobile observer and the reference station observe the same satellites. For a given satellite i, the code observation equation for the reference station (subscript r) is given by (see equation (1)):

$$P_r^i = D_r^i + T_r^i + I_r^i + c\left(\Delta t^i - \Delta t_r\right) \tag{13}$$

The position of satellite i can be computed based on the broadcast navigation message (of which the accuracy ranges between 1 and 2 m). As the reference station

position is precisely known, it is possible to compute a "theoretical" estimation of the geometric distance D_r^i:

$$D_r^i = \left(D_r^i\right)_{eph} + E_r^i \tag{14}$$

Where $\left(D_r^i\right)_{eph}$ is the estimation of D_r^i, the geometrical distance between satellite i and the reference station, obtained using the broadcast ephemeris and E_r^i the error on this estimation due to the error on the satellite position computed using broadcast ephemeris.

Based on equation (14), equation (13) can be rewritten:

$$P_r^i = \left(D_r^i\right)_{eph} + E_r^i + T_r^i + I_r^i + c\left(\Delta t^i - \Delta t_r\right) \tag{15}$$

In the same way, we can write for the mobile user (subscript u):

$$P_u^i = \left(D_u^i\right)_{eph} + E_u^i + T_u^i + I_u^i + c\left(\Delta t^i - \Delta t_u\right) \tag{16}$$

At the reference station, $\left(D_r^i\right)_{eph}$ can be estimated as the position of this station is known. This allows to compute a differential correction DC:

$$\begin{aligned} DC &= \left(D_r^i\right)_{eph} - P_r^i \\ &= -E_r^i - T_r^i - I_r^i - c\left(\Delta t^i - \Delta t_r\right) \end{aligned} \tag{17}$$

We assume that the mobile user can receive and process instantaneously the differential correction DC:

$$\widehat{P_u^i} = P_u^i + DC \tag{18}$$

with $\widehat{P_u^i}$ the pseudo-distance corrected using DC.

If we rewrite equation (18) using equations (16) and (17), we obtain:

$$\widehat{P_u^i} = \left(D_u^i\right)_{eph} + E_{ur}^i + T_{ur}^i + I_{ur}^i + c\left(\Delta t_r - \Delta t_u\right) \tag{19}$$

with the notation:

$$*_{ur}^i = *_u^i - *_r^i \tag{20}$$

In an ideal case, the residual atmospheric effects T_{ur}^i, I_{ur}^i and the residual orbit error E_{ur}^i are negligible. In practice, these residual errors increase when the distance between the mobile user and the reference station increases. DGPS can be used on distances up to 1 000 km and usually provides an accuracy ranging from 1 to 4 m depending on the distance and on tropospheric and ionospheric activity.

For applications in navigation, users are interested in the horizontal component (East and North components) of their positions. Therefore, we discuss Space Weather effects only on the horizontal component of positions. Figure 6 compares

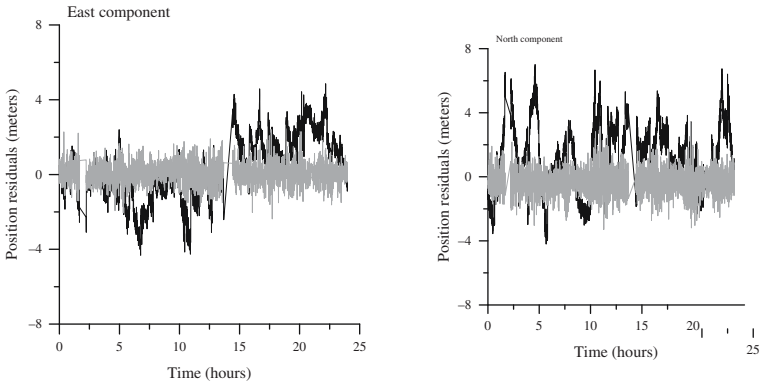


Figure 6. Residual error in horizontal (North and East) component of Dourbes station on DOY 120 in 2003 without (black line) and with (grey line) DGPS corrections generated by station Brussels

the North and East components of position residuals (i.e. computed position minus real position) for the station Dourbes (Belgium) computed using code measurements without differential correction (black line) and with differential corrections (grey line) generated by the Brussels station (the distance between Brussels and Dourbes is about 80 km). The ionospheric residual error I^i_{ur} depends mainly on TEC gradients which exist between the 2 stations considered or more precisely on the slant TEC difference between the ionospheric points observed in the 2 stations. On such a short distance (at mid-latitudes), the residual ionospheric error due to variability in TEC between the reference station and the user will be negligible with respect to the code measurement noise and other effects like multipath.

The influence of Space Weather conditions at mid-latitudes can be observed on larger distances. In the text which follows, we will "quantify" the ionospheric activity by giving the mean daily and maximum daily TEC value at Brussels for the day considered in the analysis. Let's consider the station Potsdam in Germany. We computed Potsdam position using DGPS corrections generated by the station Brussels (the distance Brussels-Potsdam is about 628 km). Figure 7a shows the East component of the Potsdam position residuals on DOY 001 in 2003: this figure represents usual conditions, the ionosphere activity is rather low (mean TEC: 6 TECU, maximum TEC: 21 TECU). In practice, at mid-latitude, strong gradients in TEC which can degrade DGPS accuracy will mainly be observed at solar maximum. Figure 7b shows Potsdam position residuals for DOY 063 in 2002 during the second maximum of Solar cycle 23 (mean TEC: 35 TECU, maximum TEC 62 TECU): it appears that the horizontal position error grows up to 5–6 meters what is not unusual at solar maximum. TEC gradients induced by severe geomagnetic storms can also degrade horizontal positions: for example, the severe storm ($Kp = 9$) of DOY 324 in 2003 was the origin of degradations of about 2–3 m during a few hours on the baseline Brussels-Potsdam.

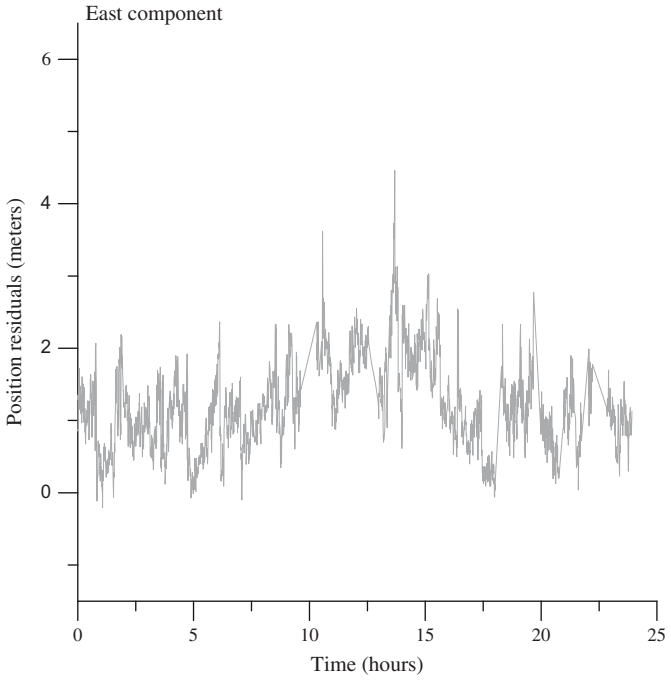


Figure 7a. Residual error on the East component of Potsdam station on DOY 001 in 2003 with DGPS corrections generated by station Brussels

The largest degradations are observed in the equatorial region where North-South gradients in TEC of up to 30 TECU/100 km are observed at solar maximum (Wanninger 1993). Skone et al. (2002) reported horizontal errors of up to 25 m on a 430 km baseline in Brazil.

Relative Positioning with Real Time Kinematic

Principle of RTK

Real Time Kinematic is a technique which allows to measure positions in real-time with a centimetre level accuracy. RTK users combine their own phase measurements with the measurements made by a reference station of which the position is precisely known. The distance between the user and the reference station should not be larger than 10–20 km mainly depending on ionospheric activity. The RTK technique measures the vector (baseline) between the reference station (denoted using subscript A) and the mobile user (subscript B) by forming single and double differences of phase measurements.

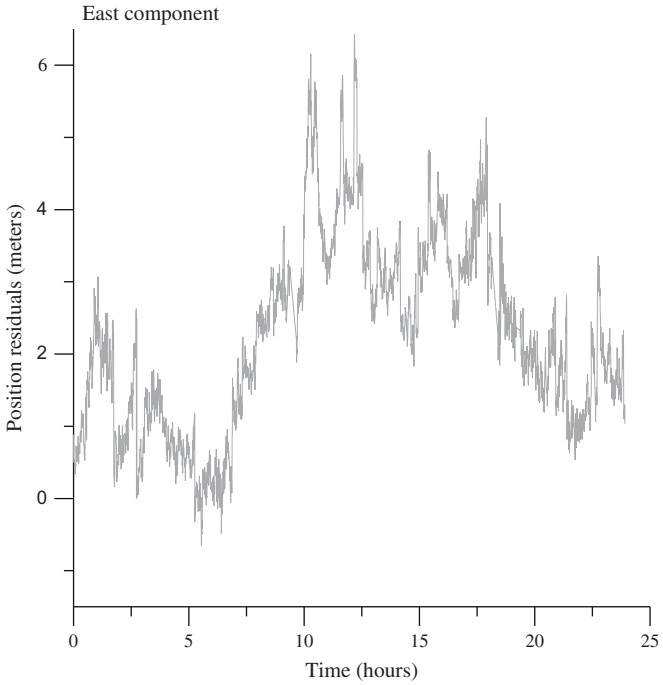


Figure 7b. Residual error on the East component of Potsdam station on DOY 063 in 2002 with DGPS corrections generated by station Brussels

If φ_A^i and φ_B^i are phase measurements made simultaneously by receivers A and B on satellite i, the single difference φ_{AB}^i is defined as:

$$\varphi_{AB}^i = \varphi_A^i - \varphi_B^i \tag{21}$$

If receivers A and B observe a second common satellite j, they can form a second single difference. Then, the double difference φ_{AB}^{ij} is defined as:

$$\varphi_{AB}^{ij} = \varphi_{AB}^i - \varphi_{AB}^j \tag{22}$$

Based on equation (2), equation (22) can be rewritten:

$$\varphi_{AB}^{ij} = \frac{f}{c}\left(D_{AB}^{ij} + T_{AB}^{ij} - I_{AB}^{ij}\right) + N_{AB}^{ij} \tag{23}$$

with the notation:

$$*_{AB}^{ij} = (*_A^i - *_B^i) - (*_A^j - *_B^j) \tag{24}$$

In the double differences, all the error sources which are common to the phase measurements performed by receivers A and B cancel, in particular, satellite

and receiver clock errors. In addition, in the case of RTK, which is used on short distances, orbit residual errors can be neglected. Residuals atmospheric effects T_{AB}^{ij} and I_{AB}^{ij} depend on the distance between A and B and also on the atmospheric "activity". Given the short distances considered, RTK data processing algorithms usually assume that residual atmospheric errors are negligible. In this case, equation (23) can be rewritten:

$$\varphi_{AB}^{ij} = \frac{f}{c} D_{AB}^{ij} + N_{AB}^{ij} \tag{25}$$

Let's recall that the term D_{AB}^{ij} contains the unknowns (i.e. the receiver B coordinates or more exactly the 3 components of the baseline vector).

Effect of ionospheric small-scale variability on ambiguity resolution

Precise positioning with RTK requires the resolution of the ambiguity N_{AB}^{ij} in real-time: RTK uses sophisticated data processing algorithms which allow to resolve phase ambiguities in a few minutes as long as residual atmospheric errors remain negligible with respect to GPS carriers wavelength (about 20 cm). This assumption is valid in most of the cases. Tropospheric residual errors are usually negligible. Nevertheless, disturbed ionospheric and Space Weather conditions can be the origin of smaller-scale (a few kilometres) variability in the Total Electron Content which can itself strongly degrade or even prevent ambiguity resolution. As outlined in paragraph 3, such a variability can be brought by 3 types of ionospheric disturbances: Travelling Ionospheric Disturbances (TID's), noise-like behaviour in TEC and scintillations.

To analyze the effects of these structures (in particular the structures shown in Figures 4 and 5) on double differences, we use measurements collected in 2 permanent GPS stations of which the position are precisely known (at a few millimetre level): Brussels and Saint-Gilles (baseline of about 4 km) in Belgium. The term D_{AB}^{ij} in equation (25) can be evaluated due to the fact that the station A and B and the satellite i and j positions are known. Therefore, we can form the combination $\varphi_{AB}^{ij} - \frac{f}{c} D_{AB}^{ij}$. If residual ionospheric errors are negligible, this combination should remain constant and be equal to the ambiguity N_{AB}^{ij}:

$$\varphi_{AB}^{ij} - \frac{f}{c} D_{AB}^{ij} = N_{AB}^{ij} \tag{26}$$

Figure 8 displays the combination $\varphi_{AB}^{ij} - \frac{f}{c} D_{AB}^{ij}$ on day of year 359 in 2004 on the Brussels-Saint Gilles baseline for satellite pair 29–26. This is an "usual" case, the combination remains nearly constant (and equal to an integer number) due to the fact the atmospheric errors are negligible. In such a situation, ambiguity resolution will be easy and successful. Figure 9a shows the same combination for the same day but later in the day and for another satellite pair 17–21 where TID's have been detected (Fig. 4): TEC variability due to these TID's is the origin of residual ionospheric effects which cause peak to peak fluctuations of about 1 cycle

in the double difference. In this case, the ambiguity is solved to a wrong integer what degrades positions up to several decimetres. Figure 9b shows the effect of noise like variability in TEC observed on DOY 324 in 2003 (Fig. 5) on the same baseline and on satellite pair 15–16. The signature of the double difference (Fig. 9b) is very well correlated with Fig. 5. Again, peak to peak residual effects of more than one cycle are observed with the same consequences on precise positioning with RTK.

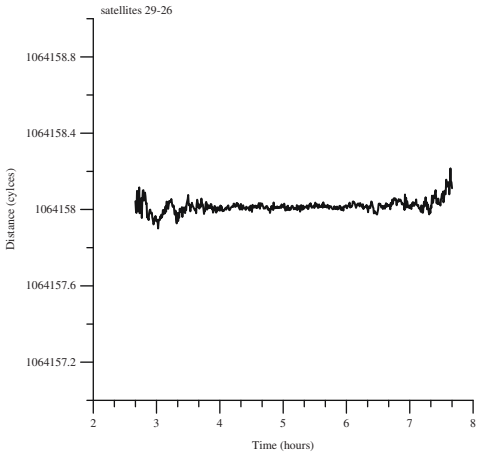


Figure 8. Double differences (ambiguity + residual errors) on the baseline Brussels-Saint Gilles (4 km) for DOY 359 in 2004 on satellite pair 29–26

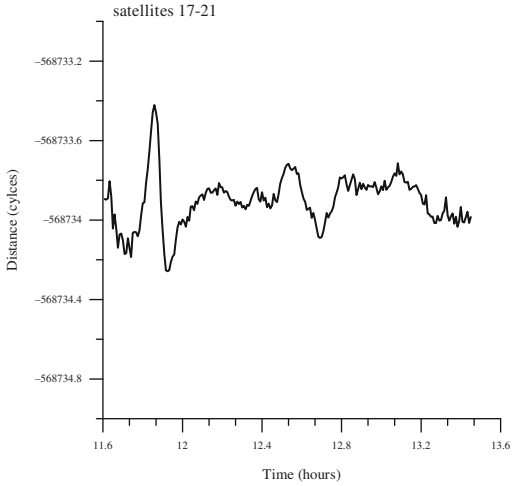


Figure 9a. Double differences (ambiguity + residual errors) on the baseline Brussels-Saint Gilles (4 km) for DOY 359 in 2004 on satellite pair 17–21 (in the presence of a TID)

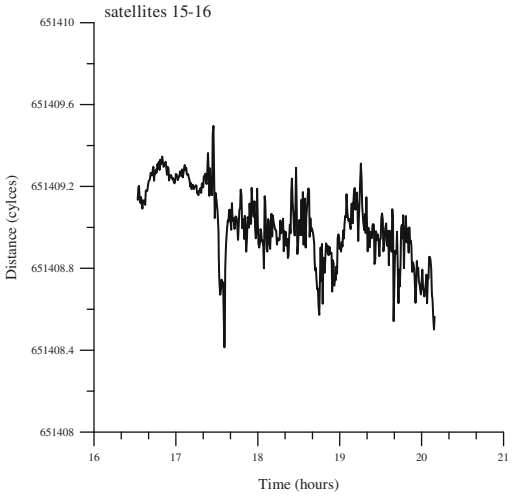


Figure 9b. Double difference (ambiguity + residual errors) on the baseline Brussels-Saint Gilles (4 km) for DOY 324 in 2003 on satellite pair 15–16 (in the presence of noise-like variability in TEC)

GNSS Space Weather Services

As showed in the previous paragraphs, active Space Weather conditions can be the origin of strong degradations in real time GNSS applications. Field GNSS users are usually not aware about Space Weather effects on their measurements. This is an important limitation to the reliability of GNSS applications: the user cannot be sure that he can trust his results. For example, RTK users who would have been on the field on DOY 359 in 2004 around Brussels would have experienced errors of several decimetres on their positions without being aware about that. In addition, even if strong TID's have been observed during that day, the "background" ionospheric conditions were very quiet (mean TEC: 5 TECU, maximum TEC: 10 TECU). Therefore, a service informing GNSS users about Space Weather effects on their measurements would be very useful. This is particularly true for the future European GNSS, Galileo, which will offer services with certified accuracy levels.

Since the beginning of the 90's, the Royal Observatory of Belgium and the Royal Meteorological Institute of Belgium are conducting a research program dedicated to the study of ionospheric and Space Weather effects in Space Geodesy. The goal of this project is to assess in real time, to forecast a few hours in advance and to mitigate Space Weather and ionospheric effects on GNSS. Based on the experience gained in this field, we decided to develop different Space Weather related products for GNSS users. These products, which are available at http://www.gpsatm.oma.be since April 2004, are mainly dedicated to DGPS and RTK users. For example, our service assesses the effects of the ionospheric activity on RTK using a colour scale:

- Black: Extreme small-scale variability. RTK applications are severely degraded or even are impossible.

- Red: Severe to extreme small-scale variability. Severe degradations of RTK applications are expected.
- Orange: Strong level of small-scale variability; degradations of RTK applications are expected.
- Green: low level of variability; no degradation due to the ionosphere expected for RTK applications.

When red or black conditions are observed, our service sends warning messages to our registered users. In the case of the strong TID's detected on DOY 359 in 2004, a warning message (via e-mail or SMS) has been sent to inform our users about the occurrence of red conditions. The service also makes forecasts about the expected positioning conditions in the near future: for example, if a severe geomagnetic storm is "under way", our users are informed about the fact that the positioning conditions could be degraded in the next hours.

In the future, we intend to further develop these products to warn Galileo users against ionospheric threats. Such products will allow to increase the reliability of Galileo applications: if the certified accuracy cannot be guaranteed any more due to unusual Space Weather conditions, Galileo customers will be automatically informed.

CONCLUSION

In this paper, we showed that Space Weather effects on GNSS depend very much on the type of application. Differential navigation is mainly affected by strong TEC gradients which are observed at solar maximum. In particular, the North-South gradients in TEC of up to 30 TECU/100 km observed at solar maximum in the equatorial region can degrade positioning accuracies up to 25 m. At mid-latitudes, degradations of up to 6 m are observed at solar maximum; extreme geomagnetic storms can also induce abrupt degradations of accuracies to about 2–3 m.

Precise relative positioning is affected by smaller-scale variability (a few km) in TEC. Such a variability can be induced by Travelling Ionospheric Disturbances, scintillations or noise-like behaviour in TEC. This variability can reach a level of 2.5 TECU/min even at mid-latitudes. Under such circumstances, residual ionospheric effects can induce fluctuations of more than 1 cycle in the double differences. In this case, the ambiguity resolution procedure can fail what degrades positioning accuracies up to several decimetres.

REFERENCES

Béniguel, Y., Adam, J.-P.: Effects of ionospheric scintillations on GNSS operations (this volume)

Leick, A.: GPS Satellite surveying, 3rd Edition, John Wiley & Sons, 435 (2004)

Seeber, G.: Satellite Geodesy, 2nd Edition, de Gruyter, 589 (2003)

Skone, S., Shrestha, S. M.: Limitations in DGPS positioning accuracies at low latitudes during solar maximum, Geoph. Res. Letters, Vol. 29, 10, 10.1029/2001GL013854 (2002)

Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J.: GPS. Theory and practice, 5th Edition, Springer, 382 (2001)

Wanninger, L., Effects of equatorial ionosphere on GPS, GPS World, July, pp 48–54 (1993)
Warnant, R., Kutiev, I., Marinov, P., Bavier, M., Lejeune, S.: Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 1. Monitoring variability in TEC which degrades the accuracy of Real Time Kinematic GPS applications, Adv. Space Res. (in press)
Warnant, R., Kutiev, I., Marinov, P., Bavier, M., Lejeune, S.: Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 2. RTK events during disturbed and quiet geomagnetic conditions, Adv. Space Res. accepted
Warnant, R., Pottiaux, E.: The increase of the ionospheric activity as measured by GPS, Earth, Planets and Space, Vol. 52, 11, pp 1055–1060 (2000)

CHAPTER 3.2

NEW IMPROVEMENTS IN HF IONOSPHERIC COMMUNICATION AND DIRECTION FINDING SYSTEMS

LOUIS BERTEL, CHRISTIAN BROUSSEAU, YVON ERHEL, DOMINIQUE LEMUR, FRANÇOIS MARIE AND MARTIAL OGER

IETR (Institut d'Electronique et de Télécommunication de Rennes), UMR CNRS 6164, Université de Rennes I (France)

INTRODUCTION

This chapter presents new HF (3–30 MHz) systems dedicated to ground to ground radio links with applications to ionospheric characterisation, channel modelling, radio communications, direction finding and single site localisation.

The received signals result from the vectorial addition of the multipaths generated by the ionosphere. Considering the acquisitions at the outputs of an array of identical antennas (homogeneous array), a high level of spatial and temporal correlation can be observed. Therefore, it appears relevant to additionally discriminate the incoming modes by considering their polarisations.

The purposes of the different systems which are described in the following sections are the use of a heterogeneous array. This polarisation-sensitive solution for array processing is principally characterized by the spatial distribution of non identical antennas.

Consequently, the applications to digital communication involve a multi channel processing in the receiver as a SIMO (single input multiple output) structures. The correlation factors depend on the polarisation characteristics of the incident wave. Moreover, the heterogeneous array is still efficient with a reduced space diversity (set up in a limited place), the differences in the polarisation parameters balancing the weak values of the differential geometrical phases.

In the following developments, the suggested techniques aim to take some better advantage of the ionospheric medium in several applications.

J. Lilensten (ed.), Space Weather, 147–168.

Such developments require a deterministic model of the polarisation at the exit point of the ionosphere. This model has been proposed by Bertel et al. (1989) and can be applied to communication systems, assuming only the limit conditions of Budden (1952).

Section 2 describes a model of received signal. It is the heart of the simulation software used in different developments. Its originality stands in the consideration of the vectorial nature of the incident signal.

Section 3 presents an operational HF radio link for digital transmission through the ionospheric channel. Thanks to a multi channel receiving system connected to a heterogeneous array, the data transfer rate attains 40 kbits/s and significantly exceeds the capabilities of standard HF modems (4.8 kbits/s).

Section 4 describes a system of Direction Finding (DF) operating on different heterogeneous arrays. The polarisation sensitivity allows a separated estimation of the DOA (Direction Of Arrival) for the two magneto ionic modes (Ordinary and eXtraordinary), improving by this way the angular resolution.

Section 5 gives a conclusion about the capabilities of these different systems operating on heterogeneous antenna arrays. Several outlooks are mentioned, as for example the possibility of video conferencing through the ionospheric channel.

MODEL OF THE HF RECEIVED SIGNAL

Spatial Response of an HF Active Receiving Antenna

A solution of the Maxwell equations in the ionospheric plasma has been proposed by Appleton and Hartree in the context of the magneto-ionic theory. It underlines the presence of two magneto-ionic propagation modes (Ratcliffe 1962, Davies 1990) named ordinary (denoted O) and extraordinary (denoted X) corresponding to two different refractive index.

For the following applications, only the polarisation at the exit point of the ionosphere is required for a given HF radio link. It is calculated considering the limit conditions of Budden (Budden 1952) which express that the electron density tends towards zero (and jointly the longitudinal component of the electric field) at the output points of the ionosphere. In these conditions, the incident wave is transverse electromagnetic (TEM) and elliptically polarized from the exit of the ionosphere to the receiving station. Its elliptical shape is fully described with two parameters η and α. The first parameter is the ellipticity ratio (real) η: its absolute value quantifies the respective lengths of the two axes of the ellipse along which the electrical field rotates; its sign indicates the clockwise $(+)$ or counter-clockwise $(-)$ rotation. The second parameter is the inclination angle α evaluated in the wave plane between the main axis and a horizontal direction.

The phasor vector, defined as $\mathbf{w} = (0\ 1\ j\eta)^T$, gives an expression of the electric field in the wave plane, including the received scalar signal $s_r(t, \mathbf{r})$ related to the transmitted signal:

$$\mathbf{E}(t,\mathbf{r}) = \begin{pmatrix} 0 \\ 1 \\ j\eta \end{pmatrix} s_r(t,\mathbf{r}) = \mathbf{w} s_r(t,\mathbf{r}) \quad (1)$$

where t is the time and **r** is the position vector.

For a given receiving site and a DOA characterized by the azimuth and elevation angles $\theta = (Az, El)$, the coordinates of the point at the exit of the ionosphere are easily estimated. Applying the results of the magneto-ionic theory with models of the electron density and a data base related to the Earth's geomagnetic field, it is then possible to calculate both parameters η and α as functions of the DOA θ and of the receiving station location. The expressions of η_O and η_X, respectively for the O and X mode, have been first derived by Appleton (Davies, 1990).

The knowledge of the incident polarisation yields an expression of the signal at the output of a receiving antenna. In the case of a wire antenna with a simple geometry (dipole antenna for example), the expression results in (Bertel et al., 1989):

- a rotation matrix $\mathbf{Q}(\alpha)$ characterizing, in the wave plane, the change of the system of coordinates from the main axis of the ellipse to a second system regarding an horizontal direction as a reference axis:

$$\mathbf{Q}(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix} \quad (2)$$

- a second rotation matrix $\mathbf{R}(\theta = (Az, El))$ characterizing the transition from the previous system to a topocentric system attached to the receiving antenna; its axis are for example the west-east, south-north and vertical directions:

$$\mathbf{R}(\theta) = \begin{pmatrix} \cos Az & -\sin El.\sin Az & -\cos El.\sin Az \\ -\sin Az & -\sin El.\cos Az & -\cos El.\cos Az \\ 0 & \cos El & -\sin El \end{pmatrix} \quad (3)$$

- an antenna specific vector **V** combining the three components of the electric field in the received signal. For example, the expression of **V** for a vertical monopole of length L is simply:

$$\mathbf{V} = (0\,0\,L) \quad (4)$$

With this description, the output signal $x_r(t)$ can be written as (Erhel et al. 2004):

$$x_r(t) = \mathbf{V}\mathbf{R}(\theta)\mathbf{Q}(\alpha) \begin{pmatrix} 0 \\ 1 \\ j\eta \end{pmatrix} s_r(t) \quad (5)$$

where $x_r(t)$ is the temporal expression of the received signal evaluated at the phase centre of the antenna. Assuming that the parameters α and η are DOA dependent in a deterministic way, the previous formula can be shortened as:

$$x_r(t) = F(\theta) s_r(t) \quad (6)$$

where $F(\theta)$ is named spatial response of the antenna (Rojas-Varela 1987, Bertel et al., 1989). $F(\theta)$ is generally complex due to the structure of the phasor **w**. It is real valued only for linear polarisations ($\eta = 0$).

Let us remark that this presentation applied to "ground to ground" ionospheric propagation, is valid for transionospheric applications as well.

Expression of the HF Received Signal

Introducing the carrier frequency f_0 and the complex envelope $m(t)$ of the transmitted signal and assuming a single propagation path, the output signal is expressed as:

$$x_r(t) = AF(\theta)m(t-\tau_g)e^{2j\pi(f_0+\delta f)(t-\tau_p)} \tag{7}$$

where A is the amplitude factor, τ_g, the group delay, τ_p, the phase delay and δf, the Doppler frequency shift.

In presence of *NS* multiple paths identified by the DOA $\{\theta_k\}$, this expression becomes, omitting the phase delays

$$x_r(t) = \sum_{k=1}^{NS} A_k F(\theta_k)m(t-\tau_{gk})e^{2j\pi(f_0+\delta f_k)t} \tag{8}$$

where A_k, τ_{gk} and δf_k are the same parameters than those previously defined but related to the path k.

Array processing involves a set of *NC* antennas with a common reference for the geometrical phase. In this context, the expression of the output signal for antenna i is:

$$x_{ri}(t) = \sum_{k=1}^{NS} A_k F_i(\theta_k)e^{j\varphi_i(\theta_k)}m(t-\tau_{gk})e^{2j\pi(f_0+\delta f_k)t} + n_i(t) \tag{9}$$

where $F_i(\theta_k)$ is the spatial response of antenna i for the DOA θ_k, $\varphi_i(\theta_k)$, the geometrical phase of this antenna relatively to θ_k, and $n_i(t)$, the additive noise.

A specific device for HF applications has been designed at the IETR laboratory: it is made up with different HF active antennas set up on a mast with the same phase centre. It appears as a particular heterogeneous array without any space diversity. In the corresponding expression of the acquisition on antenna i, the mention to the geometrical phase $\varphi_i(\theta_k)$ is then suppressed:

$$x_{ri\text{col}}(t) = \sum_{k=1}^{NS} A_k F_i(\theta_k)m(t-\tau_{gk})e^{2j\pi(f_0+\delta f_k)t} + n_i(t) \tag{10}$$

Validation of the Model

A model of received signal is considered as relevant if the experimental acquisitions captured on the sensors of an array are in a good accordance with the samples calculated according to the method under test.

With this point of view, the original device of seven HF collocated active antennas is considered. An optimization of the diversity of their complex spatial responses and a minimization of the mutual coupling are required for the final structure represented in Fig. 1: it contains 2 orthogonal vertical loop antennas, 1 horizontal loop, 2 V-shaped dipoles, 1 vertical dipole and 1 dipole with an original geometry.

Figure 2 plots the acquisitions (over 50 seconds) at the output of the receivers connected to the device, for Skelton (UK) – Rennes (F) radio link (Le Meins et al., 1999). A bandpass filtering selects a narrow bandwidth containing the carrier frequency. The presence of multi paths (generated by the ionospheric channel) induces fading. The minima of power do not appear at the same time on the seven channels. The reason is that the incident sources are combined with varying phases on each antenna: these are the arguments of the antenna responses $\{F_i(\theta_k)\}$ to the incoming polarisations.

Consequently, the diversity of the spatial responses results in a partial decorrelation of the seven received signals though there is no spatial diversity in this particular "array".

Moreover, these measurements can be compared with the results of simulations based on the proposed model of HF signal. The predicted values of parameters are

Figure 1. Array #1 of seven HF collocated antennas (patent n°99 16112)

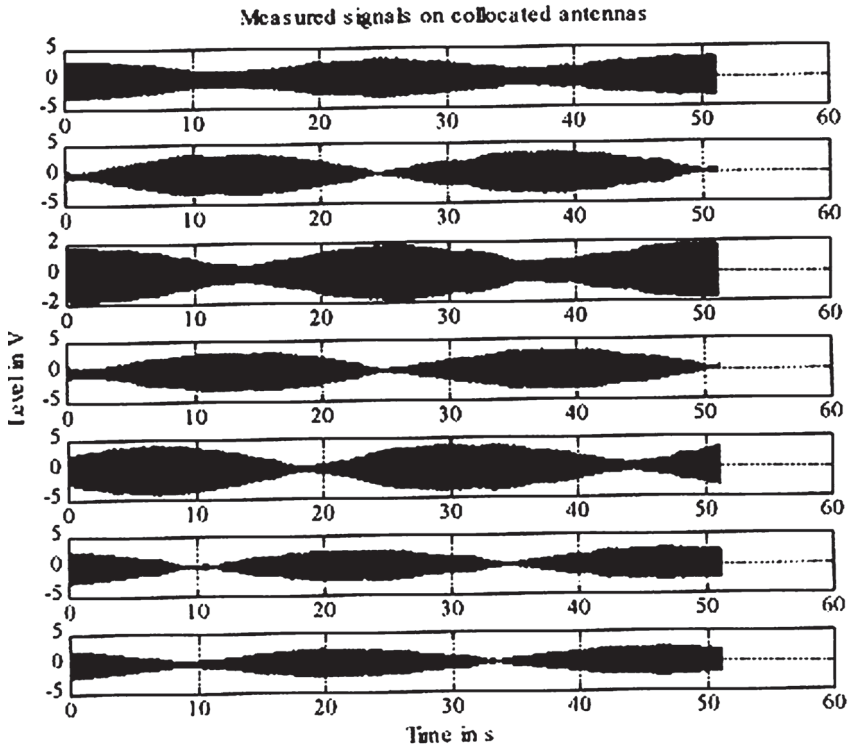


Figure 2. Experimental received signals

relative to the link under test (DOA, Doppler and group delay spreads) according to the possible scenario: one hop, one reflecting layer and to complementary modes O and X. As plotted in Fig. 3, the calculated signal envelopes appear to be quite similar to the experiments. Such observations underline the reliability of the proposed model.

Acquisitions in a 3 kHz "large" bandwidth give additional information by the means of time-frequency analysis. Actually, this bi-dimensional representation of the fading gives a simple method to jointly identify group delays and differential Doppler shifts. The instantaneous power spectral density (PSD) contains interference patterns, the geometric characteristics of which are in relation which the parameters of the propagation. As an example, large band acquisitions on the EW (East West oriented) and NS (North South oriented) loops are analysed on Fig. 4. Deep temporal fading and moderate frequency selectivity are visible on the two time-frequency representations. However, the minima of the PSD appear at different instants and frequencies for these 2 channels, indicating an effective decorrelation though the two sensors are collocated.

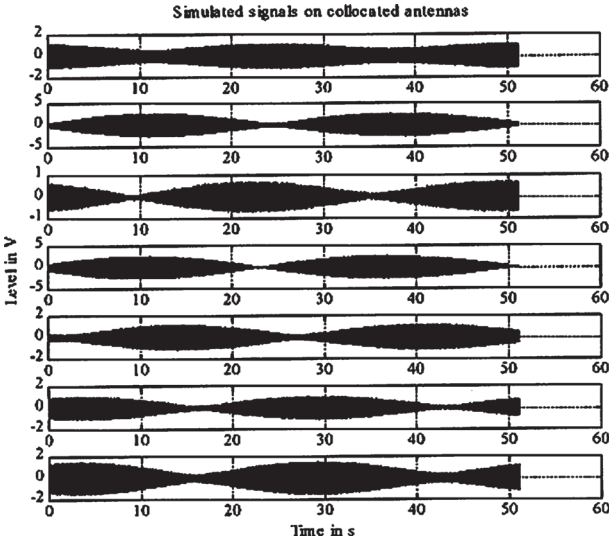


Figure 3. Simulated received signals

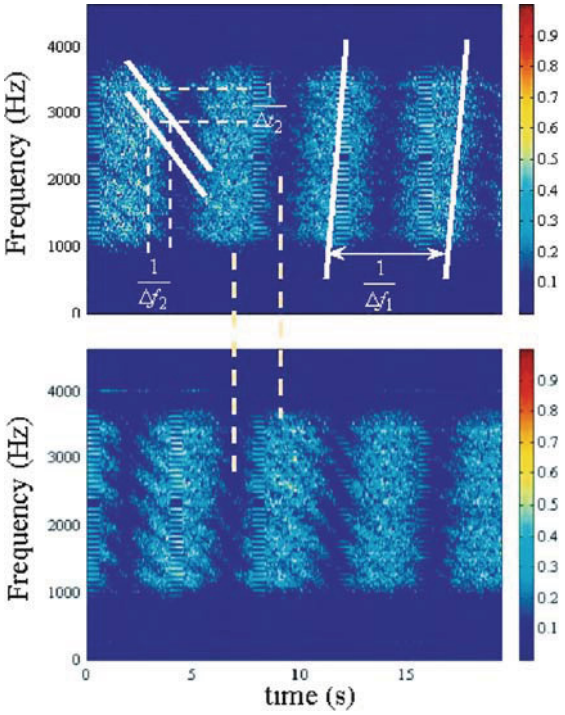


Figure 4. Time-frequency representations: EW loop (top) and NS (bottom) loop acquisitions (Bisiaux, 2001)

Synthetic Comparison of the Different Models of HF Signal

In numerous previous works (Watterson et al., 1970 Vogler et al., 1988), the dispersion bandwidth values were not estimated. The definitions of so called narrow or wide bandwidth were therefore ambiguous. According to the LOCAPI forecast software (Brousseau, 1999), it has been possible to determine such values.

For each ionospheric path, the associated signal appears to be stationary on bandwidth of at least 40 kHz. Table 1 sums up (not exhaustively) some useful signal models.

A SIMO SYSTEM OF DIGITAL TRANSMISSION THROUGH THE IONOSPHERIC CHANNEL

A unidirectional system of digital transmission for HF (3–30 MHz) applications has been developed, with the aim of increasing significantly the data transfer rate compared to standard modems. The original device of collocated antennas is a part of a multi-channel receiving system. Such a SIMO structure is represented in Fig. 5. In most of the experiments which are mentioned further, a maximum of four antennas where involved in the reception.

As indicated in the previous section, this device exploits the differences in the incoming wave polarisations to provide acquisitions with a relatively low level of correlation: array processing techniques are efficient though the system does not resort to spatial diversity. Thus, the setup is realized in a limited place.

Transmitter

The heart of the transmitter is a fully flexible modulator generating a waveform with adjustable parameters: symbol duration, roll-off factor, type of constellation restricted to the mono carrier case. The output of the modulator is connected to an amplifier with a maximum transmitted power of 200W.

The transmitter feeds an 8 m high delta antenna with a radiation pattern, which maximizes the power associated with the sky wave as suitable for a transhorizon radio link.

Receiving System

Hardware

The receiving system associates the following elements:

- the original collocated sensor device that induces an effective decorrelation of the acquisitions due to the polarisation sensitivity of the array
- a maximum of eight coherent receivers delivering baseband output signals with a maximum bandwidth of 12 kHz
- a synchronous eight channels analogue to digital converter providing samples at a rate of 24 ksamples/s with a resolution of 14 bits.

Table 1. HF signals models

Main steps	Type of measurements	Theoretical studies	Type of models	References
The first statistical HF model	Measurements in the mid latitudes	Statistical studies of the received signals	SISO, mid latitudes with limits	Watterson et al. (1970)
Extension of the statistical models	Measurements in several areas over the world	Identification of some characteristics of the propagation from the data	Application of the Watterson model to several areas in the world, variations around this model	Wagner et al. (1985) Le Roux et al. (1987) Vogler et al. (1988)
From statistics to physical approach	Extension of the measurements all over the world, determination of the number of modes, the corresponding Doppler shift and group delay	Wideband models Studies of the effects of the ionospheric profiles Introduction to the study of antenna and polarisation effects	Introduction of physical models SISO with application to wideband systems and introduction to physical data obtained from the electron profile	Vogler et al. (1988) Bertel et al. (1989)
From physical measurements to physical model	Wideband measurements	Determination of the dispersion bandwidth taking antenna effects into account Introduction of digital modulations in the models	SIMO vectorial Models	Bertel et al. (2000) Erhel et al. (2004) Le Roux et al. (2000)

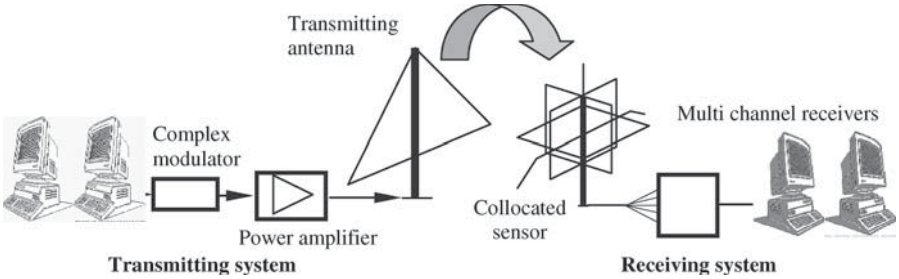


Figure 5. Description of the communication system

Depending on the type of experiment, these samples are considered as input of a real time computation implemented in a digital signal processor or are stored on the hard disk of a PC for further exploitation.

Signal processing

A spatiotemporal equalization is implemented, for each of the $NC = 4$ receiving channels, as a FIR filter involving NR delayed samples. At a previous stage of this project, the least mean squares (LMS) algorithm was performed; it requires the transmission of training sequences (Perrine et al., 2004). Investigations have been made to remove this constraint with a blind algorithm to increase (slightly) the actual data transfer rate and to authorize the connection of asynchronous subscribers. Therefore, a global estimation of the vector, containing the $NC \times NR$ taps coefficients, is carried out using a blind cost function CMA (Constant Modulus Algorithm) (Godard, 1980).

Considering that the mean square error (MSE) remains higher with that solution, an alternative consists in the use of the CMA criterion during the convergence step (avoiding training sequences) and then a switch to the LMS algorithm in the decision directed mode, which reduces the steady value of the MSE.

The synchronization techniques are classically based on a maximum of likelihood approach operating on the equalized samples (Mengali et al., 1997).

Experimental Results

An operational 780 km range radio link is set up between Valensole (French Southern Alps) and Rennes (France). The carrier frequencies are in the 8–10 MHz band. In the experiment presented herein, the transmitted waveform is a QAM 16. Its Nyquist envelope has a roll-off factor equal to 0.2. The symbol duration and the bit transfer rate are respectively equal to 0.1 ms (bandwidth of 12 kHz) and 40 kbits/s. The average signal to noise ratio SNR for each channel equals 16 dB.

The 20 kbits transmitted file contains a 256×256 compressed still image corresponding to a global compression rate equal to 25.

Efficiency of the spatiotemporal equalization

The efficiency of the multi-channel processing can be evaluated by comparing the spectrum of the transmitted signal (Fig. 6a) in a bandwidth of 12 kHz (corresponding to a data transfer rate equal to 40 kbits/s), the spectrum of the received signal on the channel with the best SNR (Fig. 6b) and the spectrum at the equalizer output (Fig. 6c). In presence of multi paths, the dispersion of group delays is responsible for the distortion of the received spectrum which contains several minima (Fig. 6b) appearing with a period of approximately $\Delta f = 1400$ Hz. This observation can be interpreted as the reception of two incident signals separated by a differential group delay of $\Delta\tau_g = 1/\Delta f = 0.71\,ms$.

The spatiotemporal equalizer appears efficient since the output spectrum plotted in Fig. 6c is almost flat in the 12 kHz bandwidth, indicating that the frequency selectivity seen in Fig. 6b is corrected.

Quality of service

The benefit of the multi-channel processing is obvious in the comparison of restored images for different configurations of the receiving system. For a given transmitted

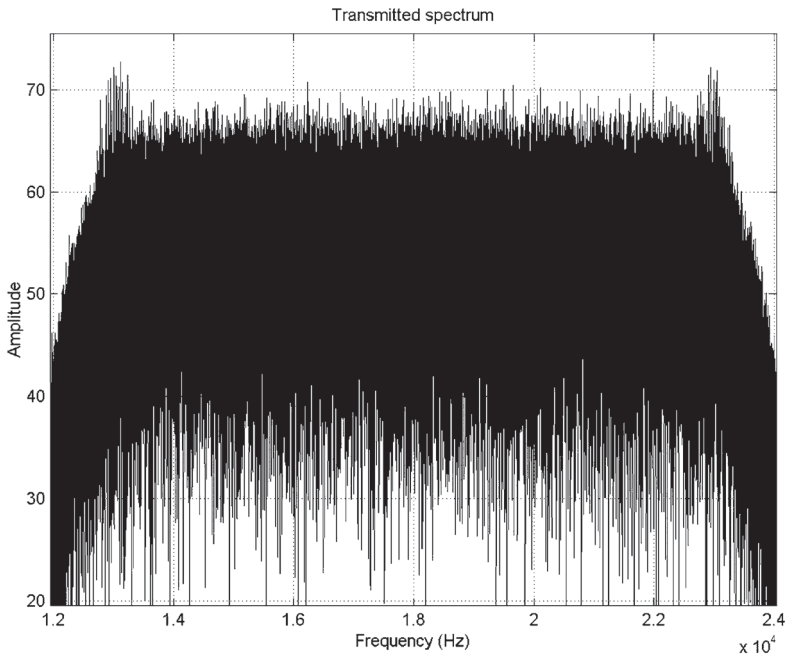


Figure 6a. Transmitted spectrum

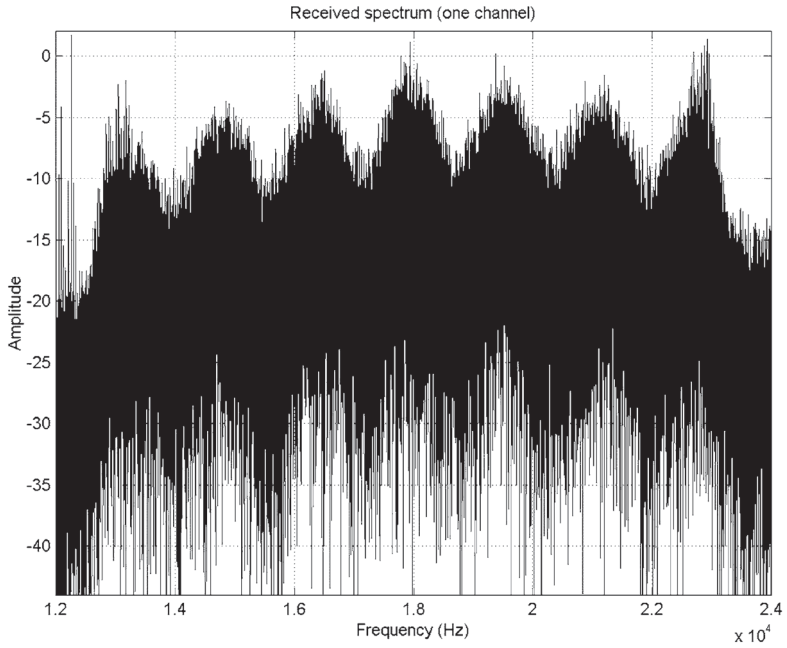


Figure 6b. Received spectrum

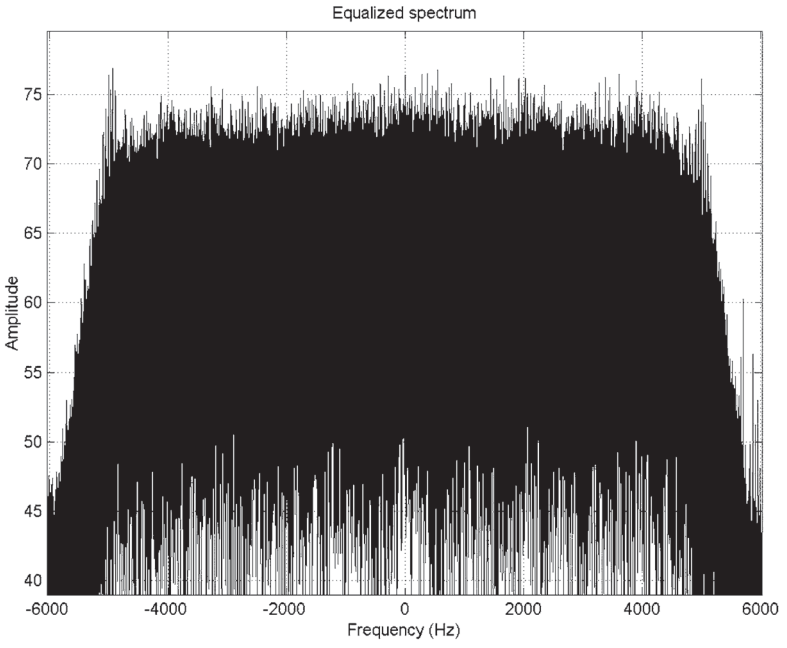


Figure 6c. Spectrum at the equalizer output (base band)

file, the visual quality is quantified by a measure of the BER on the decoded file and by the PSNR (Peak Signal to Noise Ratio) of the restored image. It appears from statistics based on a large volume of transmitted data that the minimum number of antennas should be, at least, four. However, these antennas have to be chosen to ensure uncorrelated signals (Perrine et al. 2004). This result is illustrated in Fig. 7, which shows the original image (Fig. 7a), and different restored images, the number of receiving channels being variable.

Figure 7a. Original "Lena" image

Figure 7b. Restored with a 1 channel processing, BER $= 4.710^{-1}$

Figure 7c. Restored with a 3 channels processing, BER $= 1.9510^{-1}$

Figure 7d. Restored with a 4 channels processing, BER = 1.7310^{-5}

Comparison with Existing Standards

The coherence bandwidth of the ionospheric channel is generally considered equal to 3 kHz. This numerical value is integrated in recent standards for HF modems like Mil-Std-188-110A for military purposes and in the design of commercial transceivers for amateur radio as well. Table 2 succinctly presents, in a non-exhaustive manner, a few HF modems with their characteristics in data bit rate and modulation techniques.

HF RADIO DIRECTION FINDING OPERATING ON A HETEROGENEOUS ARRAY

The challenge for radio direction finding in the HF band is the estimation of two angles (azimuth and elevation) per DOA in a context of strong spatial correlation.

Table 2. Some HF modems

Company	Norm	Bit rate	Modulation	Particularity	Reference
	military standard STANAG 5066 annexe G	9600 b/s	QAM 64	Only a simulation	Jorgenson et al. (1999)
Harris	military standard MIL-STD-188-110B	9600 b/s	QAM 64	With coding and interleaving	Nieto (2000)
Université de Rennes 1 (3 to 12 kHz)		9600 b/s to 40 kb/s	QAM 16 QAM 64	Use of collocated or multiple antennas for the receiving system, no interleaving	Erhel et al. (2001) Perrine et al. (2004)
DRM (6 KHz)	DRM	26 kb/s possible and 12 kb/s effective	COFDM +QAM		www.drm.org

Numerous methods (Capon, Esprit, Maximum Likelihood, Weighted Subspace Fitting, etc.) have been developed which are presented like "high resolution" techniques, in the sense that the minimum resolvable difference between two angles of arrival is much less than the width of the main directional lobe of the conventional beam former. Among this list, the MUSIC algorithm (MUltiple SIgnal Classification) became very popular by presenting several advantages listed in Erhel et al. (2004).

As indicated before, the heterogeneous array authorizes to consider the incoming polarisation as an additional factor for the discrimination of the incident signals. Therefore, a specific derivation of the MUSIC algorithm for such an array is proposed in this section.

Expression of the Observations

A heterogeneous array is made up of sensors, which are different from one another. An a priori knowledge is supposed for their respective spatial responses denoted by $\{F_n(\theta)\}$, n = 1, ..., NC.

In this context, the linear model for the output signals of the heterogeneous array is expressed as:

$$\mathbf{X_h}(t) = \sum_{k=1}^{NS} \mathbf{a_h}(\theta_k) s_{rk}(t) + \mathbf{N_h}(t) \tag{11}$$

The received signal $s_{rk}(t)$ is related to the k mode or path. The components of the steering-vectors $\mathbf{a_h}(\theta_k)$ combine the spatial responses and the exponentials which represent the phases $\varphi_n(\theta_k)$ calculated with respect to the array geometry:

$$\mathbf{a_h}(\theta_k) = (F_1(\theta_k)e^{j\varphi_1(\theta_k)}, F_2(\theta_k)e^{j\varphi_2(\theta_k)}, \ldots, F_{NC}(\theta_k)e^{j\varphi_{NC}(\theta_k)})^T \tag{12}$$

Denoting by $\mathbf{F}(\theta) = (F_1(\theta), \ldots\ldots, F_{NC}(\theta))^T$ the vector of the antenna responses for the DOA θ, the whole steering vector is expressed as:

$$\mathbf{a_h}(\theta) = \mathbf{F}(\theta) \otimes \mathbf{a}(\theta) \tag{13}$$

where $\otimes$ represents the Schur-Hadamard product for two matrices. The term $\mathbf{a}(\theta)$ represent the array steering vector considering only the spatial diversity (geometrical phases only).

It can be noticed that $\mathbf{a_h}(\theta)$ does not have a constant norm; this remark have to be taken into account when applying the MUSIC algorithm on this particular type of array.

MUSIC Algorithm

For applications in the HF band, two possible types of polarisation (O and X) are expected at the exit of an ionospheric radio link. Consequently, for a given

DOA θ, two steering-vectors are defined and attached to the corresponding incident modes by:

$$\mathbf{a}_{\mathbf{h}P}(\theta) = (F_{1P}(\theta)e^{j\varphi_1(\theta)}, F_{2P}(\theta)e^{j\varphi_2(\theta)}, \ldots, F_{\mathrm{NCP}}(\theta)e^{j\varphi_{NC}(\theta)})^T \quad (14)$$

P representing the polarisation type O or X.

MUSIC uses the eigen decomposition of the data covariance matrix $\mathbf{R}_{\mathbf{xxh}} = E[\mathbf{X}_\mathbf{h}(t)\mathbf{X}_\mathbf{h}(t)^H]$. Its computation is based on the orthogonality between an incident steering vector and the noise subspace spanned by the eigenvectors of **Rxxh** associated with its smaller eigen values.

The implementation includes the following steps:

- estimation of the covariance matrix:

$$\hat{\mathrm{R}}_{\mathbf{xx_h}} = \frac{1}{N_{ech}} \sum_{n=1}^{N_{ech}} \mathbf{X}_\mathbf{h}(n)\mathbf{X}_\mathbf{h}(n)^H \text{ with } N_{ech}\text{snapshots of data} \quad (15)$$

- Eigen decomposition of the covariance matrix and estimation of the number of sources NSE based on the dispersion of the eigen values
- computation for all the potential DOA θ, of the angular function (pseudo-spectrum) evaluating the norm of the projection of the steering-vector in the noise subspace. As the projected vector should have a constant norm, the calculus involves the vectors:

$$\mathbf{b}_{\mathbf{h}P}(\theta) = \frac{\mathbf{a}_{\mathbf{h}P}(\theta)}{|\mathbf{a}_{\mathbf{h}P}(\theta)|}, \quad \mathrm{P} = \mathrm{O} \text{ or } \mathrm{X} \quad (16)$$

The directions of arrival being estimated for both polarisations, the two sets of steering-vectors are projected in the noise subspace. Consequently, in this original version of the algorithm, two pseudo-spectra are computed according to the following expression:

$$\mathrm{PSSP}_P(\theta) = \frac{1}{\sum_{m=NSE+1}^{NC} |\mathbf{v}_m^T.\mathbf{b}_{\mathbf{h}P}(\theta)|^2} \quad (17)$$

The estimation of the DOA with two different array manifolds illustrates the polarisation sensitivity of the heterogeneous array.

Experimental Results

Three types of heterogeneous array have been set up for measurements of direction finding. The first one is the original device made up of collocated antennas and presented in section 2: its main advantage is a set up in a reduced size

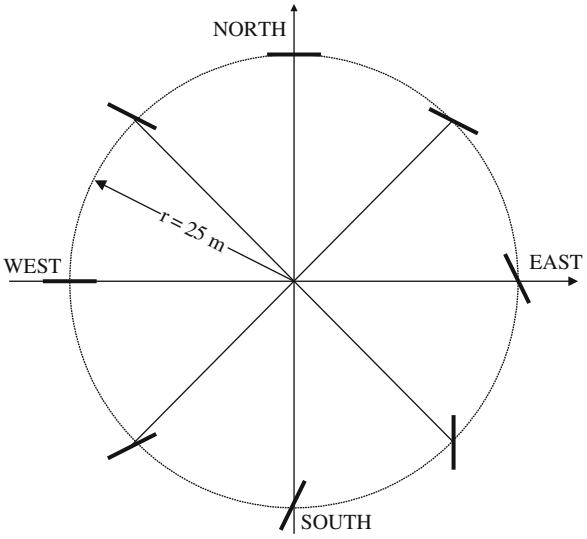


Figure 8. Array #2 (Erhel et al., 2004)

(volume of $2\,m^3$). (Fig. 8) The second one contains eight active loop antennas with different orientations and equally spaced on a horizontal circle with a 25 m radius. It combines diversity of the spatial responses and space diversity. The third array (Fig. 9) contains three groups of sensors set up with a moderate spacing (less than 10 meters) along the vertical direction and one horizontal axis. Each group contains three collocated antennas: two vertical crossed and one horizontal loop antennas. It appears as a relatively compact heterogeneous array.

Theses arrays are connected to a multi channel receiving system the outputs of which are acquired for the computation of the MUSIC algorithm. For

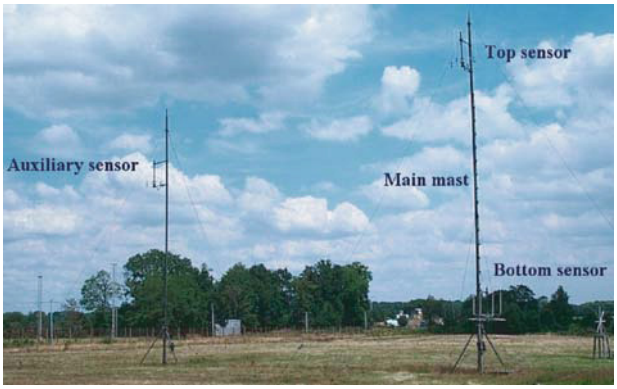


Figure 9. Array #3 (Le Bouter, 2004)

a given transmitter located in Southeast of France (geographical azimuth $A_{zo} = 132°$, distance 780 km, carrier frequency $f_0 = 6.735$ MHz), examples of experimental pseudo spectra, involving short acquisitions, are plotted on Figures 10 to 12.

Whatever the array, according to those rough results, the mean angular estimations are close one to each other: for example, the mean values of the azimuth, resulting from an averaging on 70 instantaneous estimations, are equal to 131°, 134° and 133° as the geographical azimuth of the transmitting site is equal to 132°. However, the contrast and resolution appear optimal for array #2, decreases slightly for array #3 and more seriously for array #1. This observation can be interpreted by the positive role of space diversity in arrays #2 and #3.

For the same transmitter, a long-term experiment has been running on array #1 with DOA estimations every 2 minutes during a period of 3H20. The angular estimations are plotted in Fig. 13 with a small circle for O modes and an "x" letter for X modes. Two complementary modes are detected and measured during the acquisition.

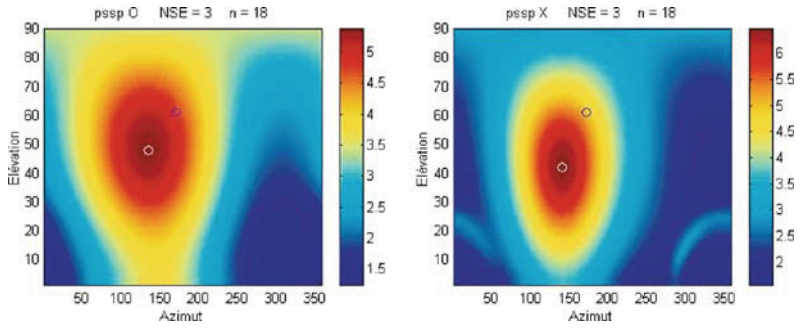


Figure 10. Experimental pseudo-spectra (O and X), array #1

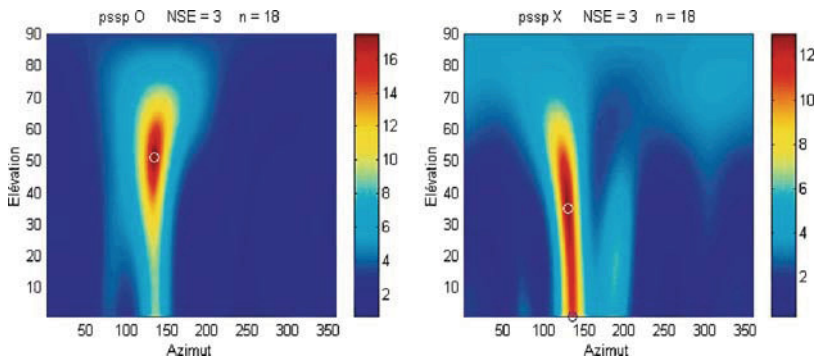


Figure 11. Experimental pseudo-spectra (O and X), array #2

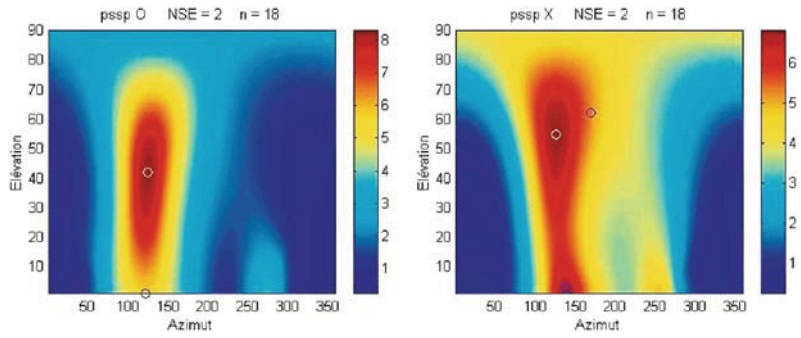


Figure 12. Experimental pseudo-spectra (O and X), array #3

The dispersion on the azimuth appears moderate with a mean value close to the reference (132°). The elevation estimation indicates the presence of two different paths at the beginning of the experiment, converging in a unique path at instant referenced in Fig. 13 by the snapshot 60.

Another operational system (Marie et al., 2005) is based on a heterogeneous array of differently oriented whip antennas distributed along a circle with a small radius (Fig. 14). The results of direction finding, coupled with the PRIME (Prediction and Regional Ionospheric Modelling over Europe) model of the ionosphere Bradley

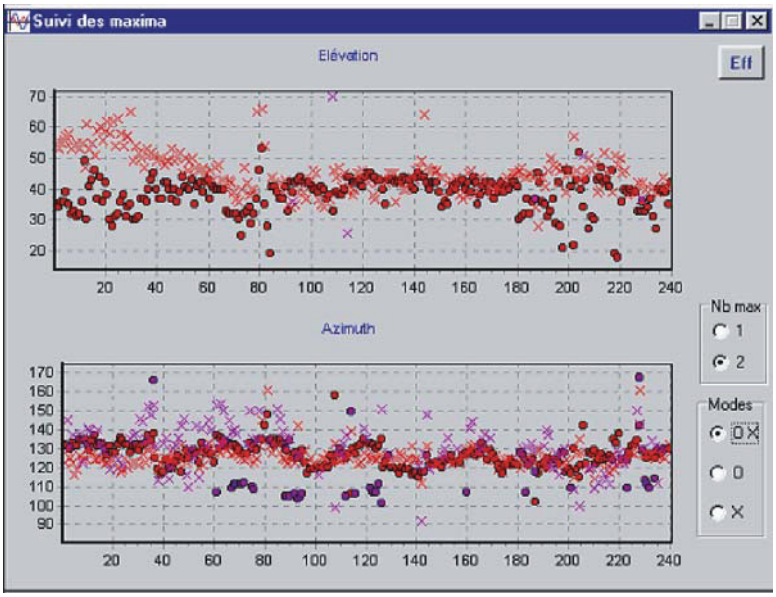


Figure 13. Angular estimations

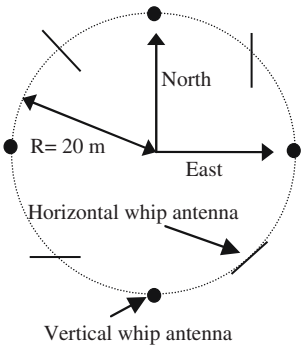


Figure 14. heterogeneous array of whips

1999), are the input data of a single site localization (SSL). The tests carried out localizations of HF transmitters located in Europe and North Africa at a maximum distance of 2000 km. More than 100 experiments, involving 35 military beacons or broadcast transmitters, were conducted during a period of 3 weeks, at various times in the day (and during the night) and for different expected azimuths. An example of pseudo spectra (for each O and X mode) is plotted in Fig. 15 and the corresponding SSL, estimated with a very good accuracy, is presented in Fig. 16.

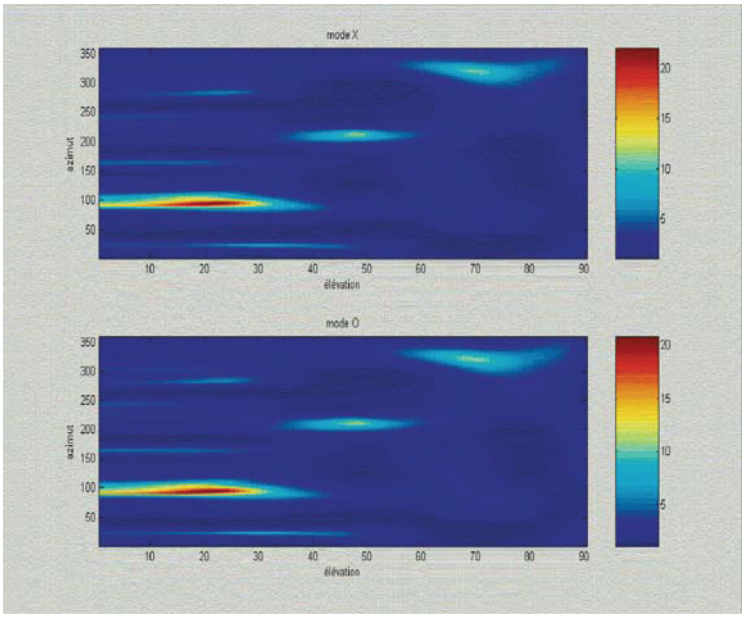


Figure 15. Estimated pseudo spectra O and X from a transmitter located at Tiganesti (Romania)

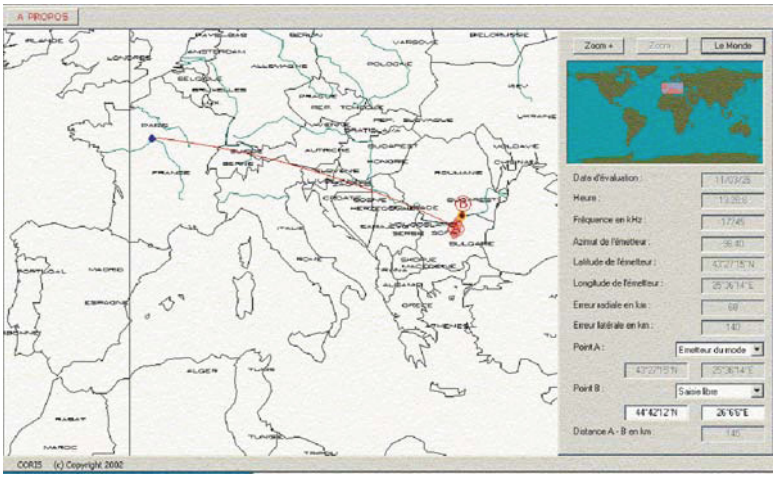


Figure 16. SSL estimation of the transmitter located at Tiganesti (Romania)

CONCLUSION

The presented model of HF signals and its applications underline the great interest to include antenna effects in the conception of systems. Both polarisation and antenna diversities have been considered in addition to – or to replace – space diversity in HF receiving systems. Several examples illustrate how these effects could be used to improve direction finding or the capability of transmitting systems. New simulators can also be developed with the proposed equations as basis.

For DF applications, the heterogeneous array induces a separated research of DOA for the two expected polarization types. This technique improves the angular resolution in general and authorizes the set up of arrays with a reduced aperture.

For transmission systems, the polarization sensitivity induces a significant decorrelation of the acquisitions so that an efficient spatio temporal equalization balances the distortion induced by the propagation in an extended bandwidth up to 12 kHz. This technical solution largely improves the performances of HF modems "on the shelf" since the data transfer rate attains 40 kbits/s. Associated with a robust joint source-channel coding, it permits to consider the feasibility of videoconferencing through the ionospheric channel in a next future.

REFERENCES

Bertel, L., Rojas-Varela, J., Cole, D., Gourvez, P.: Polarisation and ground effects on H.F. receiving antenna patterns, Annales des Télécommunications, vol. 44, n°7–8, pp 413–427 (1989)

Bertel, L., Marie, F., Lemur, D.: Model of narrow band H. F. ionospheric channel including both propagation and antenna effects. Antenna and Propagation Conference (ICAP), Davos, Suisse (2000)

Bisiaux, A.: Définition, conception et réalisation d'un modem vectoriel large bande utilisable dans la gamme Hautes Fréquences, Ph.D. thesis, Université de Rennes 1 (2001)

Bradley, P.: PRIME (Prediction and Retrospective Ionospheric Modelling over Europe), Action 238, Final Report, Rutherford Appleton Laboratory, Chilton Didcot, UK, European Cooperation in the field of Scientific and Technical Research (COST) (1999)

Brousseau, C., Parion, P., Bertel, L.: Possible Use of the LOCAPI Ionospheric Software to Digital Communications, Physics and Chemistry of the Earth, Part C, Vol. 24, n° 4, ISSN 1464–1917 (1999)

Budden, K.G.: The theory of the limiting polarisation of radio waves reflected from the ionosphere, Proc. Royal Soc vol. 215, n° 1121, pp 215–233 (1952)

Davies, K.: Ionospheric radio, Peter Peregrinus Ltd (1990)

Erhel, Y., Bisiaux, A., Lemur, D., Bertel, L.: Mise en œuvre de la séparation de source sur un ensemble d'antennes colocalisées: application à l'augmentation du débit numérique en transmission H. F. (3–30 MHz), Colloque GRETSI 2001, Toulouse (2001)

Erhel, Y., Lemur, D., Bertel, L., Marie, F.: H.F. radio direction finding operating on a heterogeneous array: principles and experimental validation, Radio-Science, vol. 39, n° 1, pp 1003–1; 1003–14 (2004)

Godard, D.N.: Self-recovering equalization and carrier tracking in two-dimensional data communications systems, IEEE transactions on communications, vol. 28, n° 11, pp 1867–1875 (1980)

Jorgenson, M.B., Johnson, R., Moreland, K.W.: Beyond 9600 bps at H.F., RTO (NATO) Symposium on Tactical Mobile Communications, Lillehammer (1999)

Le Bouter, G.: Conception, réalisation et tests d'un réseau de capteurs constitués d'antennes colocalisées dans la gamme hautes fréquences, Ph.D. thesis, Université de Rennes 1 (2004)

Le Meins, C., Erhel, Y., Bertel, L., Marie, F.: Source separation operation on a set of collocated antennas: theory and application in the H.F. band (3–30 MHz), Antennas Applications Symposium (IEEE), Monticello (USA) (1999)

Le Roux, Y., Savidan, G., Du Chaffaut, G., Gourvez, P., Jolivet, J.P.: A combined evaluation and simulation system of the HF channel, IEE Pub 274 on Antenna and Propagation, York, UK (1987)

Le Roux, Y., Bertel, L., Lassudrie-Duchesne, P.: Requirements for future models and simulators of the HF transmission channel, eighth international conference on H.F. Radio Systems and Techniques, IEE, University of Surrey (2000)

Marie, F., Erhel, Y., Danion, C.: An operational H.F. system for single site location, Proceedings of "European Conference on Propagation & Systems ECPS05; Brest; P2–18; 15–18 (2005)

Mengali, U., D'Andrea, A.: Synchronisation techniques for digital receivers, Kluwer Academic/ Plenum Publishers (1997)

Nieto, J.W.: Performance testing of Mil-Std-188-110B high data rate at HF waveforms, 8th International Conference on H.F. radio systems and techniques, Guildford (2000)

Perrine, C., Erhel, Y., Lemur, D., Bertel, L., Bourdillon, A.: A way to increase the bit rate in ionospheric radio links, Annals of Geophysics, vol. 47, n° 2–3, pp 1145–1160 (2004)

Ratcliffe, J.A.: The Magneto-ionic Theory and its Application to the Ionosphere, Cambridge University Press (1962)

Rojas-Varela, J.: Antennes filtre de polarisation dans la bande H.F., Ph.D. thesis, Université de Rennes 1 (1987)

Vogler, L., Hoffmeyer, J., Lemmon, J., Nensenbergs, M.: Progress and remaining in the developpement of a wideband HF channel model and simulation AGARD conference, Proceeding 442, Paris (1988)

Wagner, L.S., Golstein, J.A.: High resolution probing of the HF ionospheric skywave channel: F2 layer results, Radio SCI., 20, n° 3, pp 287–302 (1985)

Watterson, C.C., Juroshek, J.R., Bensema, W.D.: Experimental confirmation of an HF channel model, IEEE Trans. Commun. Technol., vol. COM-18, pp 792–803 (1970)

CHAPTER 3.3

SHORT-TERM *fo*F2 FORECAST: PRESENT DAY STATE OF ART

A.V. MIKHAILOV, V.H. DEPUEV AND A.H. DEPUEVA

Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation, Troitsk, Moscow Region 142190, Russia

Abstract: An analysis of the F2-layer short-term forecast problem has been done. Both objective and methodological problems prevent us from a deliberate F2-layer forecast issuing at present. An empirical approach based on statistical methods may be recommended for practical use. A forecast method based on a new aeronomic index (a proxy) AI has been proposed and tested over selected 64 severe storm events. The method provides an acceptable prediction accuracy both for strongly disturbed and quiet conditions. The problems with the prediction of the F2-layer quiet-time disturbances as well as some other unsolved problems are discussed

INTRODUCTION

Short-term (1–24 h in advance) ionospheric F2-layer forecast is still an unsolved and very challenging problem despite long history (e.g. Anderson, 1928; Hafstad and Tuve, 1929; Appleton and Ingram, 1935; Kirby et al., 1935) and many attempts undertaken. A good review of the topic at the level of 1995 was made by Wilkinson (1995). The problems with the ionosphere prediction are due to objective reasons. Physical mechanisms forming both negative and positive F2-layer disturbances are well-established by now. They are related to global thermospheric circulation, neutral composition and temperature, electric fields and plasmaspheric flux changes. The list of all pertinent processes may be found in Rishbeth (1991) and Prölss (1995). The problem is in intensity of each particular process contributing to a particular ionospheric storm formation. The Earth's upper atmosphere is an open system with many uncontrolled inputs forcing it both from above and below. If solar EUV radiation, magnetospheric electric fields, particle precipitation (impact from above) can be controlled to some extent, the intensity of internal gravity waves, dynamo and tropospheric electric fields, planetary waves (impact from below) are uncontrolled in principle. Depending on prehistory and current state

J. Lilensten (ed.), Space Weather, 169–184.

of the magnetosphere and thermosphere the reaction will be different to the same impact from above, but no thermosphere and magnetosphere monitoring is made at present and is not expected in an observable future. Thus the intensity of each particular process controlling the F2-region: magnetospheric electric fields, zones and characteristics of particle precipitation producing high-latitude Joule heating, global thermospheric circulation resulting in neutral composition and temperature variations, the internal gravity waves dissipation in the 100–120 km height range producing via eddy diffusion changes in neutral composition in the whole thermosphere above, planetary waves etc., is known pretty poor for each particular geomagnetic storm. This fact was also stressed by Wang et al. (2001) who tested a first-principle model TING. Therefore, there is not much hope at present to obtain a "deliberate" forecast of the F2-region using so called first principle or physical models.

An attempt to apply modern 1-3D physical models of the F2-region to predict even the simplest quiet time NmF2 and hmF2 daily variations gave overall unsatisfactory results in some cases (Anderson et al., 1998). A similar comparison by Fuller-Rowell et al. (2000) for disturbed conditions has demonstrated more "visual" success of the model predictions than quantitative one; correlation coefficients between model and observations are typically 0.3–0.65, depending on how the data are selected and smoothed. Negative F2-layer storm effects which are the most crucial for HF radio-wave communication cannot be satisfactory modelled without special fitting of aeronomic parameters for each particular ionospheric storm (e.g. Richards et al., 1989, 1994; Buonsanto, 1999). But it should be stressed that physical modelling is the only way to understand the mechanisms of the ionosphere formation under various geophysical conditions and its role hardly can be overestimated. Thus, theoretical modelling may be considered as a powerful tool for physical analyses rather than practical applications.

Therefore, an empirical approach to the F2-layer short-term prediction based on statistical methods (Zevakina, 1990; Wu and Wilkinson, 1995; Muhtarov et al., 1998; Kutiev et al., 1999; Marin et al., 2000; Kutiev and Muhtarov, 2001; Stanislawska and Zbyszynski, 2001, 2002; Araujo-Pradere et al., 2002, 2003; Liu et al., 2005; Tsagouri and Belehaki, 2006), or a neural network approach (Altinay et al., 1997; Cander et al., 1998; Cander and Mihajlovic, 1998; Tulunay et al., 2000; Francis et al., 2000, 2001; Wintoft and Cander, 2000; Chan and Cannon, 2002; McKinnell and Poole, 2004; Tulunay et al., 2000, 2004) which can provide an acceptable accuracy may be recommended for practical use.

Speaking about *fo*F2 short-term (1–24 h) forecast we mainly mean strongly disturbed geophysical conditions (magnetic storms). Such events are relatively rare to occur, but they are the most interesting, challenging and important from practical point of view. Depending on the intensity of geomagnetic storm and latitude of observation, electron concentration in the F2-layer maximum, NmF2 may drop by an order of magnitude (negative storm effect) compared to quiet time pre-storm conditions and this is crucial for HF radio communication. Positive F2-layer storm effects, on one hand, are less impressive (usual NmF2 increase is less than factor of 2),

on the other hand, NmF2 increase only broadens the HF working band, so they are not very important for HF communication. Prediction of quiet time *fo*F2 variations usually is not a problem and can be done with good accuracy (mean relative deviation MRD $\leq$ 10–15%) using empirical methods. The only problem with such predictions is related to so called quiet time F2-layer disturbances (Mikhailov et al., 2004) which occur under quiet geomagnetic conditions and presumably reflect the impact from below, no precursor has been established yet.

PROBLEMS WITH THE EMPIRICAL APPROACH

The empirical approach to ionospheric forecast is widely used in practice. However, there are problems with this approach as well. There is no an efficient geophysical index to predict the ionospheric storm onset, its magnitude and duration. The correlation coefficients of δfoF2 (relative *fo*F2 deviation from monthly median) with currently available planetary indices are not very high, being latitudinal dependent. Some estimates from Zevakina et al. (1990) are given in Table 1.

The best correlation is seen to provide AE and Bz IMF indices, but they are not predictable at present. Only daily Ap indices (which may be converted to Kp) are predicted currently up to 3 days in advance. Time weighed accumulation indices such as ap(τ) proposed by Wrenn (1987), Wrenn et al. (1987) seem to increase the correlation with δfoF2, but the improvement is not significantly larger than for instantaneous indices (aa, ap, Kp, Dst) – correlation coefficients $r < 0.7$. So a conclusion was made that time–weighted accumulation indices might have limited use in a forecasting environment (Wu and Wilkinson, 1995). However, the very idea of using the time accumulation impact is correct and fruitful. For instance, the empirical thermospheric models of MSIS series (e.g. Hedin, 1987; Picone et al., 2002) also use time accumulation 3-hour ap indices. A detail description of available solar, ionospheric and geomagnetic indices may be found in Perrone and de Franceschi (1998).

Next step in this direction was made by Araujo-Pradere et al. (2002, 2003) who proposed a correction model STORM based on a new index – the integral of 3-hour ap index over the previous 33 hours weighted by a filter obtained by the method of singular value decomposition. A non-linear relationship of δfoF2 with this index $\delta fo\mathrm{F2} = a_0 + a_1\mathrm{X}(t_0) + a_2\mathrm{X}^2(t_0) + a_3\mathrm{X}^3(t_0)$, where $\mathrm{X}(t_0) = \int \mathrm{F}(\tau)\mathrm{P}(t_0 - \tau)\mathrm{d}\tau$ and $\mathrm{F}(\tau)$ is the filter weighting function of the ap index over the 33 previous hours is used to correct monthly median *fo*F2 values. This new STORM model has been included to the latest version of IRI, IRI2000 (Bilitza, 2001). As the IRI2000 model may be considered as the international standard, an additional analysis of this new

Table 1. Correlation coefficients of δfoF2 with some planetary indices

Index	AE	Bz IMF	Dst	Kp
Corr. coeff.	0.86–0.52	0.86–0.69	0.71–0.46	0.77–0.33

inclusion may be interesting. Relationship between the proposed time accumulation index and δfoF2 for severe storms observed at Slough over the 1949–1996 period is given in Fig. 1 for different seasons. The correlation coefficients are seen to be rather small for such disturbed conditions. For a comparison the analysis has been repeated for the same storm periods using ap(τ) indices by Wrenn (1987), Wrenn et al. (1987). The correlation coefficients turned out to be larger: -0.494 for April, -0.346 for September, -0.575 for July, and -0.342 for November. So, the newly proposed and accepted by IRI2000 index is hardly better than ap(τ), both using the same idea of time weighed accumulation.

So, available indices of geomagnetic activity either direct or transformed do not provide high enough correlation with δfoF2. Partly this is due to the following: (a) during severe geomagnetic storms magnetometric stations are out of the auroral zone thus underestimating index values; (b) high latitude energy deposition (heating) is not uniform in longitude while global indices do not reflect this; (c) indices of geomagnetic activity (due to the very method of their generation) are 'blind' depending only on UT, while the ionospheric storm onset depends on LT, season,

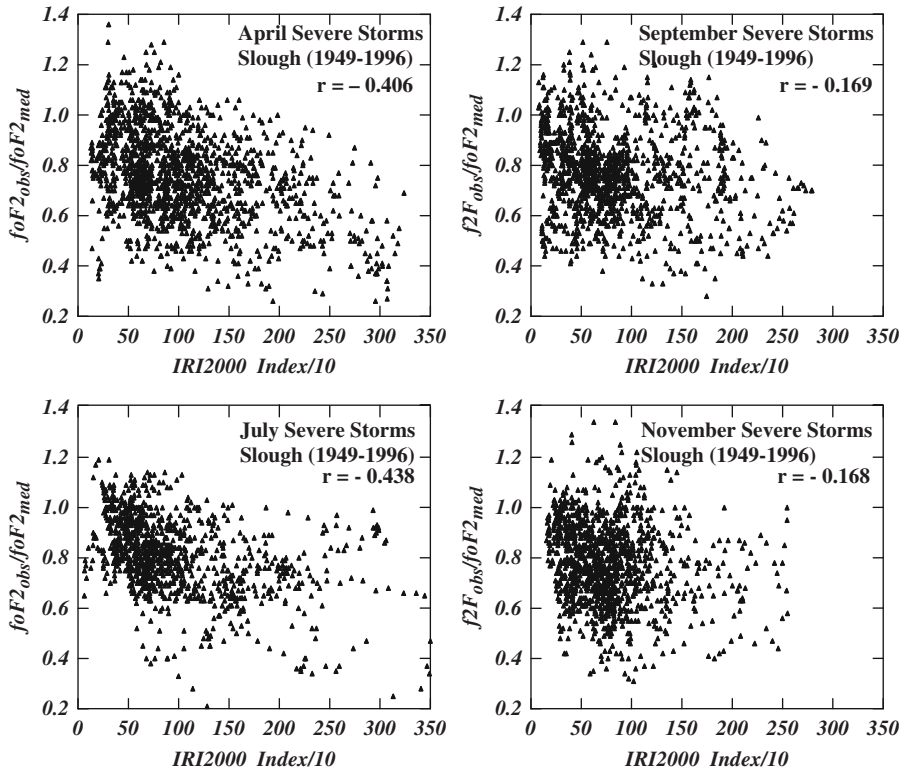


Figure 1. Relationship of δfoF2 with the IRI2000 index for selected severe storms observed at Slough over the 1949–1996 period

latitude and prehistory (state of the magnetosphere and thermosphere). The latter results in large scatter in delay between geomagnetic and ionospheric storms onset obtained by different authors: 0–6 h for positive disturbances (Zevakina and Kiseleva, 1978), 12 h (Wrenn et al., 1987), 15 h (Wu and Wilkinson, 1995), 6–12 h (Forbes et al., 2000), 16–18 h (Kutiev and Muhtarov, 2001); 8–20 h depending on season (Pant and Sridharan, 2001), 3–20 h depending on LT sector (Tsagouri and Belehaki, 2006), no time delay is considered by Araujo-Pradere et al. (2002).

Despite all the problems with geomagnetic indices, they are widely used in the ionosphere forecasting and this is due at least to the following reasons. Only geomagnetic indices (aa, ap, kp) are available for the whole period of ionospheric observations and this is important for the forecast methods development. Only daily Ap index is predicted currently up to 3 days in advance and prediction of a controlling index is necessary for any forecast method functioning. The proposed method for *fo*F2 short-term prediction (Section 3) is also based on 3-hour ap time accumulation indices.

STATISTICAL AND NEURAL NETWORKS METHOD RESULTS

A statistical approach to ionospheric forecast is based on *fo*F2 or δfoF2 regressions with various geophysical indices, the regression coefficients being specified over previous training period. In the utmost case only previous observations of the predicted parameter are used for training the method, for instance, the autocovariance method used by Stanislawska and Zbyszynski, (2001, 2002). A widely used way to demonstrate the merits of the method is to compare it with the median forecast. A 29% gain over climatology was obtained by Kutiev and Muhtarov (2001). A 34% gain in the Northern and 20% in the Southern Hemispheres has exhibited the STORM model by Araujo-Pradere et al. (2003). The best results were obtained for summer (up to 50%) but no improvement in winter. A 44% gain was obtained by Tsagouri and Belehaki (2006) over 15 impulsive storm events.

A neural networks approach is a new and perhaps promising direction in the ionospheric forecasting. The prediction results depend on the architecture of the network designed and the list of the input parameters used by the authors. In fact the neural approach may be considered as a more sophisticated version of the multi-regressional methods and the obtained results demonstrate the same merits and drawbacks as the statistical methods. A 45% accuracy improvement compared to the persistence prediction was obtained for a 1-day ahead forecast and about 42% improvement for 1-hour ahead prediction (Francis et al., 2000). Up to 50% accuracy gain over the reference persistence prediction was obtained by Chan and Cannon (2002) for a 1-hour ahead *fo*F2 forecast.

In relation with this it worth mentioning that 1-hour ahead *fo*F2 forecast (also Altinay et al., 1997; Stanislawska and Zbyszynski, 2001; Tulunay et al., 2004) hardly can demonstrate the merits of a method. The e-fold characteristic time of the NmF2 variations is ≥ 1.5 h, therefore a 1-hour ahead *fo*F2 forecast can be done with an acceptable accuracy even during strongly disturbed conditions (see Section 3.2)

if current *fo*F2 observations are included to the list of input parameters. This is valid for any prediction method in which the idea of persistence is used.

A neural networks approach was used by Wintoft and Cander (2000) to predict specially selected F2-layer storms. The observed *fo*F2 decreases (negative F2-layer storm effect) turned out to be much larger than the predicted ones. Their results are a good illustration of the general problem related to the empirical (statistical) approach for the ionospheric forecasting. Severe storms are rare events and practically there is no chance for them to occur during the training period when it is relatively short. When the training period is long (some years or even some solar cycles) such outstanding events are just lost in the sea of quiet time and slightly disturbed conditions after a statistical treatment. An extrapolation to large Ap index values observed during severe storms turns out inefficient and this results in underestimated *fo*F2 decrease. Therefore, separate methods are required for prediction of quiet and moderately disturbed conditions and severe storm periods.

FORECAST METHOD DESCRIPTION

After thorough and critical analysis of various empirical approaches to *fo*F2 short-term forecast a new statistical method has been developed and tested. The proposed method is supposed to be used in practice, so it should be based on available in near real time input information. Depending on conditions different methods may be used. If current *fo*F2 hourly (better 10–15-minute) observations are available from an ionosonde station, this allows one to obtain higher *fo*F2 prediction accuracy for the station in question and the area nearby inside the spatial correlation radius. When current *fo*F2 observations are absent due to any reason (station has closed or never worked in the area), the forecast can be made as well but with lower accuracy. The proposed method is based on the statistical approach, so it is not free from all earlier mentioned drawbacks of this approach and special efforts were undertaken to minimise their effect. To obtain a satisfactory prediction accuracy for disturbed conditions was the main concern of our development.

The Idea of the Method

Unlike all earlier discussed approaches we use a semi-empirical one, which combines both theoretical and empirical elements in the prediction scheme. Theory of ionospheric F2-layer gives the NmF2 and hmF2 dependencies on main aeronomic parameters, neutral composition and temperature being the most important. Although such dependencies are known long ago (e.g. Rishbeth and Barron, 1960; Ivanov-Kholodny and Mikhailov, 1976) and they were also confirmed by direct observations (Prölss, 1980), they were considered that time in a deterministic sense rather than in a statistical one. On one hand such relationships are approximate, on the other hand the required thermospheric parameters are not known with a sufficient accuracy for each particular geomagnetic storm – all this gave in general unsatisfactory testing results and the idea had not got further development

that time. Next step in this direction was made by Shubin and Anakuliev (1995) who used an analytical expression resulted from the continuity equation solution for electron concentration in the F2-region. They find relative deviations δfoF2 to monthly median empirical model IRI-90, necessary thermospheric parameters being taken from MSIS-86. A comparison with observed foF2 for many storm periods has shown that such approach has sense and provides good statistical results in various geophysical conditions.

The newly proposed method also uses a relationship of NmF2 with main aeronomic parameters responsible for the F2-layer formation. This expression may be considered as a new index AI (aeronomic index or a proxy) instead of solar and geomagnetic indices usually used for the ionospheric forecast. Coming from a well-known expressions by Rishbeth and Barron (1960) for a steady-state daytime mid-latitude F2-layer maximum:

$$\beta_m H^2/D_m = 0.6;\ N_m F2 = 0.75 q_m/\beta_m \tag{1}$$

where $H = kT/mg$ – scale height for neutral atomic oxygen, q_m – photoionization rate, β_m – linear loss coefficient ($\gamma_1[N_2] + \gamma_2[O_2]$), and D_m – ambipolar diffusion coefficient (all parameters are given at the height of F2-layer maximum). In case of an isothermal thermosphere with temperature T it is possible (Mikhailov et al., 1995, see also Ivanov-Kholodny and Mikhailov, 1976) to write down an expression for NmF2:

$$N_m F2 = \frac{0.75 q_1 H^{2/3}}{\beta_1^{2/3} (0.6 D_1)^{1/3}} \tag{2}$$

where all aeronomic parameters are specified now at a fixed height h_1 (say, 300 km). After substitution of $q_1 \sim [O]_1$, $D_1 \sim T^{1/2}/[O]_1$, $H \sim T$ and keeping in mind that $fo\text{F2} \sim (\text{NmF2})^{1/2}$, we obtain the final expression:

$$fo\text{F2} \propto \frac{[O]_1^{2/3} T^{1/4}}{\beta_1^{1/3}} \tag{3}$$

In our prediction method we work with relative deviations $\delta fo\text{F2} = fo\text{F2}/fo\text{F2}_{\text{med}}$, where $fo\text{F2}_{\text{med}}$ is a 28-day running median obtained over the preceding period. Such 28-day median rather than monthly one is used for the following reasons. On one hand a 28-day running median looks more natural as this period equals to one solar rotation, on the other hand, this saves us from large and unreal disturbance effects in the beginning and in the end of a month as well as at the junction of two months especially during the equinoctial periods when changes in the thermosphere and ionosphere are very fast. The advantage of using running foF2 median for F2-layer disturbance analyses was stressed long ago (e.g. Mednikova, 1957). An expression similar to (3) should be written down for a 'median' day as well, which is selected from 28 previous ones: this day should demonstrate the least sum deviation from the 28-day foF2 running median. The

'median' day is not necessary a quiet one as this depends on geomagnetic conditions for the preceding 28-day period.

Then the aeronomic index AI may be written as

$$AI = \left(\frac{[O]_1}{[O]_{1med}}\right)^{2/3} \left(\frac{\beta_{1med}}{\beta_1}\right)^{1/3} \left(\frac{T}{T_{med}}\right)^{1/4} \tag{4}$$

Neutral composition and temperature at $h_1 = 300$ km can be taken from any thermospheric model, for instance, MSIS-86 (Hedin, 1987) or from the latest one NRLMSISE-00 (Picone et al., 2002). Rate constants γ_1 and γ_2 for the ion-molecule reactions of O^+ with N_2 and O_2 are taken from Hierl et al. (1997).

Thermospheric models need their own input indices: 3-hour ap for current and some previous period and $F_{10.7}$ for the previous day and an average over 81 day centred to the day in question. Necessary information can be found in Internet:

http://www.geomag.bgs.ac.uk/gifs/apindex.html	- estimated 3-hour ap for the previous period;
http://www.sel.noaa.gov/forecast.html	- a 3-day forecast of daily Ap;
http://www.sel.noaa.gov/forecast.html	- daily $F_{10.7}$ for current day + a 3-day forecast;
http://www.sel.noaa.gov/ftpdir/indices/DSD.txt	- daily $F_{10.7}$ for 30 previous days;
http://www.sel.noaa.gov/ftpdir/latest/45DF.txt	- a forecast of daily $F_{10.7}$ for 45 days.

Indices AI calculated by this way are used both for training the method and for prediction. Unlike global direct solar and geomagnetic indices which exhibit only UT dependence, the proposed index AI, in principle, should demonstrate (via thermospheric parameters variations) the dependence on UT, LT, latitude and longitude, season, level of solar activity etc. Detailed analysis is needed in future to reveal real possibilities of this index for the ionospheric forecasting.

It is known that δNmF2 exhibits a pretty good inter-hour correlation within a day. During daytime hours the e-fold time of NmF2 changes with respect to recombination is about 1.5 hours. But daytime F2-region is strongly controlled by thermosphere (neutral composition, temperature) and the e-fold time for these parameters is longer than 1.5 hours. So the time interval of temporal correlation may be up to 3–6 hours depending on geophysical conditions. During night-time hours the characteristic time with respect to the loss process is more than 10 hours due to low linear loss coefficient at the hmF2 height and the NmF2 inter-hour correlation normally is very good for night-time hours. Therefore, the regression for *fo*F2 prediction should include previous *fo*F2 observations. But this inter-hour correlation breaks down during storm periods decreasing to 1–3 hours and special methods are needed for such conditions.

The regression used in our prediction method is written as follows:

$$\delta fo\mathrm{F2}(UT+n) = C_0 + C_1 \delta fo\mathrm{F2}(UT) + C_2 AI(UT+n) \tag{5}$$

where δfoF2(UT) – the last observed deviation at UT moment, n – the lead time (1–24 h). Previous analysis has shown that no AI/*fo*F2 time shift is required unlike methods based on direct indices such as Ap (e.g. Marin et al., 2000). The unknown coefficients C_i are obtained over the 28-day training period using the least squares multi-regressional method.

Such approach is applied during quiet time and moderately disturbed conditions, the prediction accuracy being high enough (MRD $\leq$ 10–15% for all lead times). The method should be modified for disturbed periods. As it was mentioned earlier, this is due to the fact that severe storms are relatively rare (although they are the most important and interesting), and there is practically no chance for them to occur during the training 28-day period – according to Kutiev and Muhtarov (2001) the most probably ionosphere conditions correspond to kp $\approx 3_0$(ap = 15). This results in poor forecasts for strongly disturbed periods. The following has been done to overcome the problem at least to some extent. Specially selected strong disturbances observed at a given station over the whole available period of observations were grouped in 12-month bins and δfoF2 versus AI index regressions (second order polynomial) were obtained for each month. The thresholds for the ionospheric storm onset were specified for each month as well. When during the forecast we turn out in a storm area (the threshold has been exceeded), the method switches from expression (5) to a corresponding regression. It should be stressed that despite the fact that the correlation of δfoF2 with AI index for severe storms is pretty poor especially in winter, such regressions provide an acceptable prediction accuracy and their use is convenient in practice. Due to the earlier mentioned ionosphere inertia the basic method (5) is used for lead times $\leq$3–4 hours even under disturbed conditions as this provides better prediction accuracy than with the regressions.

An example of forecast calculations for different lead times is given in Fig. 2 for a severe storm observed at Slough on Jun 4–6, 1991. This period was also tested by Wintoft and Cander (2000, their Fig. 11) using a neural networks method and their results are given in Fig. 2 for a comparison.

Testing Results

Testing of the proposed prediction method has been done using all available Slough/Chilton (1949–2004) *fo*F2 observations for strongly disturbed periods. We tried to select isolated storms and analysed the mostly disturbed 24-hour periods in the course of a storm. Such periods were chosen using ap(τ) indices by Wrenn (1987) as an indicator of the disturbance magnitude. Each storm period has been run in a routine mode without any special fitting of the controlling parameters. Mean relative deviations (MRD in %) and standard deviations (SD in MHz) were

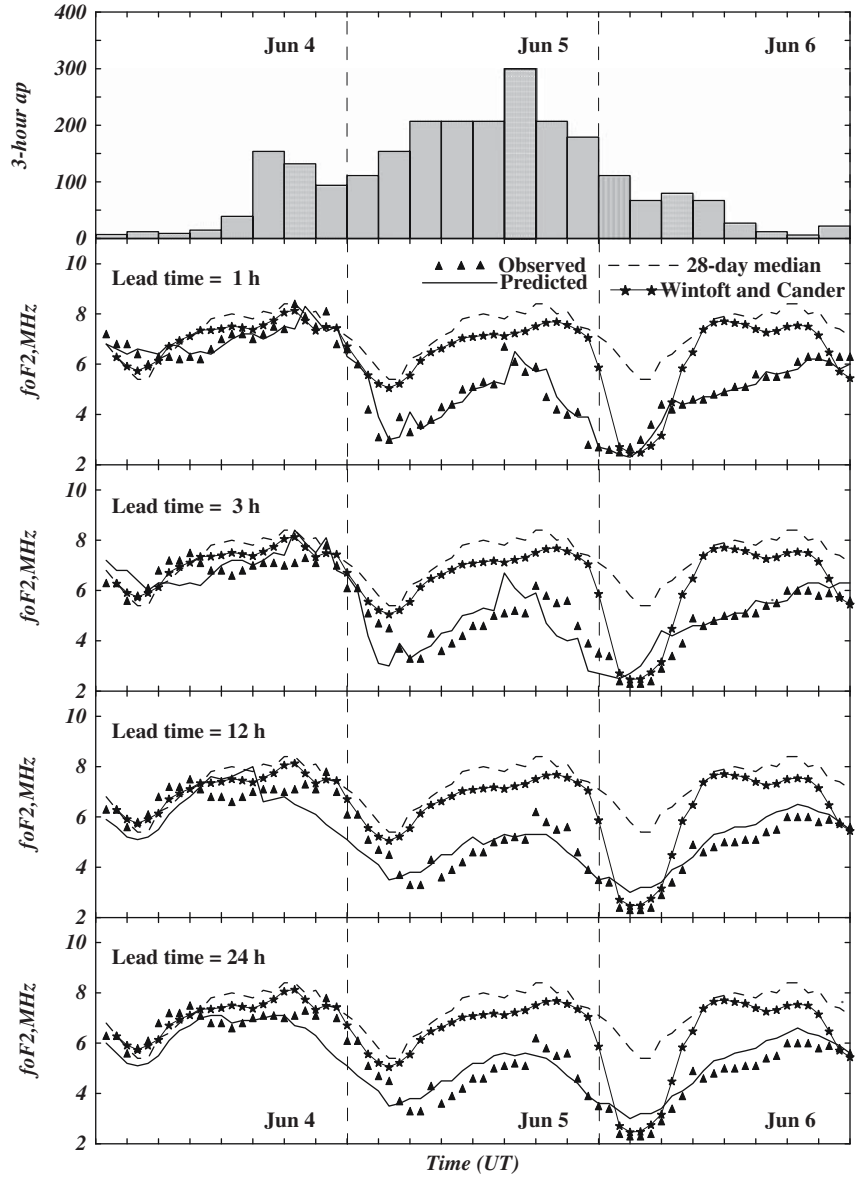


Figure 2. Observed at Slough and predicted for different lead times *fo*F2 variations for a severe geomagnetic storm on Jun 4–6, 1991. Running median and neural networks forecast by Wintoft and Cander (2000) are given for a comparison

calculated for lead times n = (1–24) h for the same 24-hour disturbed period. The results are given in Table 2 for three seasons. The basic version of the method with training over previous 28 days without using the storm regressions has been

Table 2. Testing results of the proposed method applied to selected strong storms observed at Slough (Chilton) for three seasons (N-number of storms considered). Relative mean deviation δfoF2 (in %) and standard deviations SD (in MHz) of predicted for different lead times foF2 with respect to observed values are given. Second line (italic) values obtained without using storm regressions (see text)

Season	Character	Lead time (hours)					
		1	2	3	6	12	24
	MRD (%)	6.1	10.4	13.4	16.6	16.7	16.6
Summer					*18.4*	*20.6*	*21.7*
N = 22	SD (MHz)	0.37	0.57	0.67	0.66	0.63	0.62
					0.77	*0.73*	*0.64*
	MRD (%)	7.8	13.2	17.0	23.5	23.8	21.8
Equinox					*25.5*	*27.8*	*28.7*
N = 21	SD (MHz)	0.49	0.78	0.96	0.99	0.98	0.92
					1.17	*1.12*	*0.91*
	MRD (%)	8.6	14.8	19.7	22.3	22.4	21.6
Winter					*26.9*	*27.6*	*26.9*
N = 21	SD (MHz)	0.49	0.80	1.02	1.05	1.05	1.02
					1.30	*1.19*	*1.19*

also tested and results are given in Table 2 for a comparison by italic (results for $n \leq 3$–4 h are the same and not repeated).

Similar estimations for the same storms have been obtained (Table 3) using a 28-day running median, the IRI2000 model (Bilitza, 2001) with a correction for storm periods and the model by Shubin and Anakuliev (1995). The last two models are not related to current *fo*F2 observations and can be applied to any UT moment providing the input information is available.

The proposed method is seen to provide an acceptable prediction accuracy with MRD ranging from 6 to 24% depending on lead time and season. The results were obtained for severe storm periods while in quiet and moderately disturbed conditions typical MRD $\leq$ 10–15% for all lead times. The best prediction accuracy is in summer when F2-region is mainly controlled by photochemical (local) processes and the worst in winter when dynamical processes (mainly global thermospheric circulation) dominate. During equinoctial months summer/winter transitions result in large day-to-day variations of thermospheric parameters (Mikhailov and Schlegel, 2001) and

Table 3. Testing results for the same periods as in Table 2 but for three forecasting methods (models) which can be applied for any lead time (see text). MRD (in %) and SD (in MHz, values in brackets) are given

Season	28-day median	IRI-2000	Shubin's model
Summer (N = 22)	42.6 (0.79)	20.4 (0.78)	19.4 (0.70)
Equinox (N = 21)	49.1 (1.08)	30.5 (1.07)	29.0 (1.10)
Winter (N = 21)	39.2 (1.32)	35.2 (1.23)	22.5 (1.00)

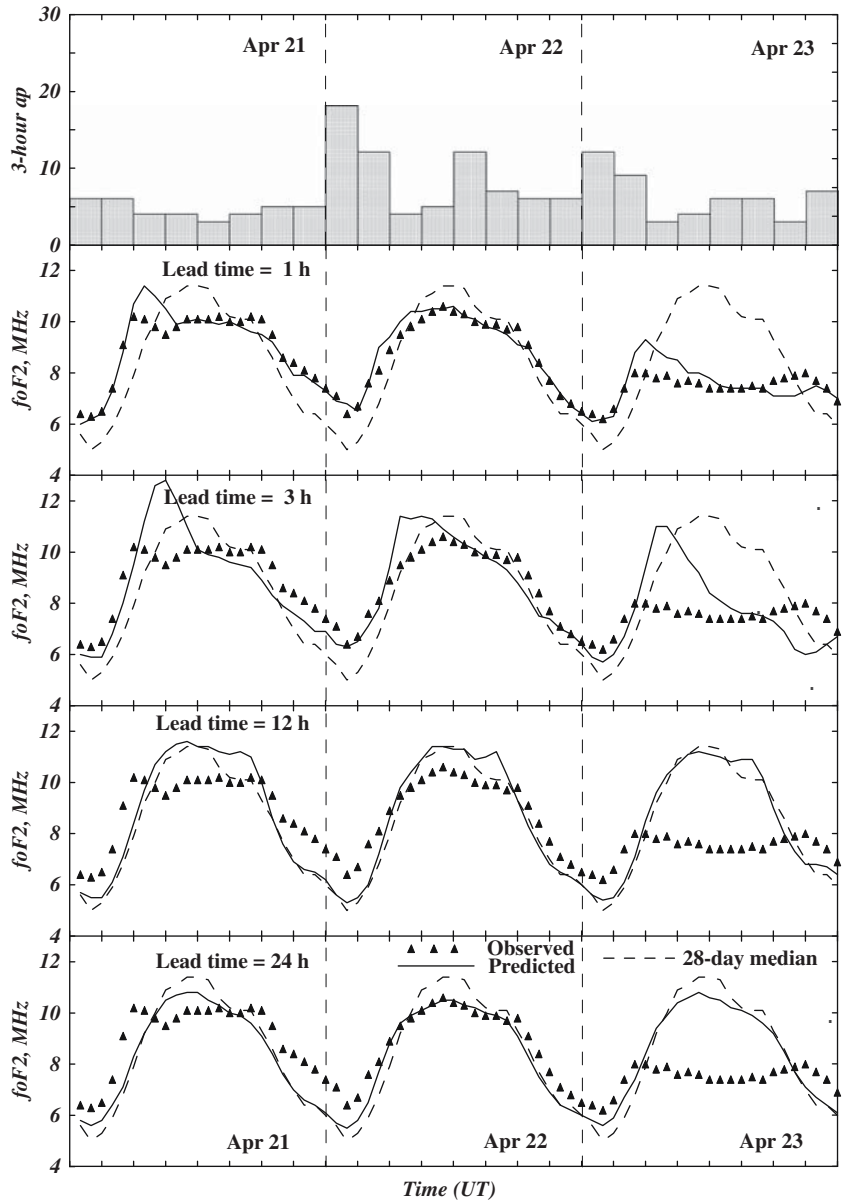


Figure 3. Observed at Moscow and predicted *fo*F2 variations for a quiet-time disturbance event on Apr 23, 1980. The method is seen to be inefficient in such conditions due to lack of an precursor

this decreases the prediction accuracy. The method without special allowance for storm conditions gives worse results for large lead times (italic in Table 2), but it can be used with a success for $n \leq 3$–4 h even in disturbed conditions. A comparison with the 28-day median results demonstrates an obvious accuracy improvement for all seasons, the gain depending on the lead time (Table 3).

The IRI2000 and Shubin's model provide less accurate forecast, but it should be kept in mind that both models are not linked to any current *fo*F2 observations and, in principle, can be used globally at any point and this is a great merit of the two models. Both models provide close results in summer and equinox, but the Shubin's model is more efficient in winter and this is very important keeping in mind the IRI2000 problems for winter season when no quantitative improvement over median forecast can be demonstrated (Araujo-Pradere et al., 2002), the latter conclusion being confirmed by our results (Table 3).

In principle, storm regressions similar to those used for Slough can be obtained for any ionosonde station which has been working (or worked in the past) for 2–3 solar cycles to have sufficient amount of observations. A forecast in this case can be done for any UT moment providing input ap and $F_{10.7}$ indices are available. The expected prediction accuracy will be close to those given in Table 2 for $n > 6$ h.

Among the problems with the *fo*F2 short-term forecast quiet time F2-layer disturbances (Mikhailov and Schlegel, 2001; Mikhailov et al., 2004) occurring under quiet geomagnetic conditions should be considered as a serious one. Presumably such disturbances are due to an impact from below but no precursor for their occurrence has been revealed yet. Any prediction method will be inefficient in such cases unless a precursor is found. An example is given in Fig. 3 for a quiet-time disturbance observed at Moscow on Apr 23, 1980.

CONCLUSIONS

An analysis of the F2-layer short-term forecast problem shows that acceptable from practical point of view solutions can be found on the way of an empirical approach. Independently on the method used there is a problem of an efficient geophysical index(es) for ionospheric F2-layer storms forecast. Some indices (e.g. AE or Bz IMF) seem to be more efficient than Ap, but only daily Ap is predicted 1–3 days in advance at present. The problem of ionospheric storm onset, its magnitude and duration has not been solved yet, therefore these very important characteristics cannot be predicted with a sufficient accuracy. Available real time *fo*F2 observations may help solve this problem. Positive F2-layer storm effect which is due to an interplay of thermospheric wind and neutral composition variations during daytime and plasmaspheric flux variations during nighttime hours cannot be predicted yet for a particular storm event.

Additional efforts are needed to reveal a precursor for quiet time F2-layer disturbances both negative and positive, their magnitude being comparable to usual F2-layer storm effects related to a moderate geomagnetic activity.

ACKNOWLEDGEMENTS

This work was in part supported by the Russian foundation for Basic Research under Grant 06-05-64227.

REFERENCES

Altinay, O., Tulunay, E., Tulunay, Y.: Forecasting of ionospheric critical frequency using neural networks, Geophys. Res. Lett. **24**, 1467–1470 (1997)

Anderson, C. N.: Correlation of long wave transatlantic radio transmission with other factors effected by solar activity, Proc. Inst. Radio Eng. **16**, 297–347 (1928)

Anderson, D.N., Buonsanto, M.J., Codrescu, M., Decker, D., Fesen, C.G., Fuller-Rowell, Reinisch B.W., Roble, R., Schunk, W., Sojka, J.J.: Intercomparison of physical models and observations of the ionosphere, J. Geophys. Res. **103**, 2179–2192 (1998)

Appleton, E.V., Ingram, L.J.: Magnetic storms and upper atmospheric ionization, Nature, **136**, 548–549 (1935)

Araujo-Pradere, E.A., Fuller-Rowell, T.J., Codrescu, M.V.: STORM: An empirical storm-time ionospheric correction model 1. Model description, Radio Sci. **37**, 1070, doi:10.1029/2001RS002467, (2002)

Araujo-Pradere, E.A., Fuller-Rowell, T.J.: Validation of the STORM response in IRI2000, J. Geophys. Res. **108**, A3, 1120, doi:10.1029/2002JA009720 (2003)

Bilitza, D.: International reference ionosphere 2000, Radio Sci. **36**(2), 261–275 (2001)

Buonsanto, M.J.: Ionospheric storms – a review, Space Sci. Rev. **88**, 563–601 (1999)

Cander, Lj.R., Mihajlovic, S.J.: Forecasting ionospheric structure during the great geomagnetic storms, J. Geophys. Res. **103**, 391–398 (1998)

Cander, Lj.R., Milosavljevic, M.M., Stankovic, S.S., Tomasevic, S.: Ionospheric forecasting technique by artificial neural network, Electron. Lett. **34**, 1573–1574 (1998)

Chan, A.H.Y., Cannon, P.S.: Nonlinear forecast of foF2: variation of model prediction accuracy over time, Ann. Geophysicae, **20**, 1031–1038 (2002)

Forbes, J.M., Palo, S.E., Zhang, X.: Variability of the ionosphere, J. Atmos. Solar-Terr. Phys. **62**, 685–693 (2000)

Francis, N.M., Cannon, P.S., Brown, A.G., Broomhead, D.S.: Nonlinear prediction of the ionospheric parameter foF2 on hourly, daily, and monthly timescales, J. Geophys. Res. **105**, 12839–12849 (2000)

Francis, N.M., Brown, A.G., Cannon, P.S., Broomhead, D.S.: Prediction of the hourly ionospheric parameter foF2 using a novel nonlinear interpolation technique to cope with missing data points, J. Geophys. Res. **106**, 30077–30083 (2001)

Fuller-Rowell, T.J., Codrescu, M.V., Wilkinson, P.: Quantitative modeling of the ionospheric response to geomagnetic activity, Ann. Geophysicae, **18**, 766–781 (2000)

Hafstad, L.R., Tuve, M.A.: Note on Kennely-Heaviside layer observations during a magnetic storm, Terr. Magn. Atmos. Electr. **34**, 39–43 (1929)

Hedin, A.E.: MSIS-86 thermospheric model, J. Geophys. Res. **92**, 4649–4662 (1987)

Hierl, P.M., Dotan, I., Seeley, J.V., Van Doran, J.M., Morris, R., Viggiano, A.A.: Rate coefficients for the reactions of O^+ with N_2 and O_2 as a function of temperature (300–1800 K), J. Chem. Phys. **106** (9), 3540–3544 (1997)

Ivanov-Kholodny, G.S., Mikhailov, A.V.: Relation of the parameters of the ionospheric F region to the parameters of the neutral atmosphere at a fixed height, Geomag. and Aeronom. **16**, 378–380 (1976)

Kirby, S.S., Gilliland, T.R., Judson, E.B., Smith, N.: The ionosphere, sunspots, and magnetic storms, Phys. Rev. **48**, 849 (1935)

Kutiev, I., Muhtarov, P.: Modelling of midlatitude F region response to geomagnetic activity, J. Geophys. Res. **106**, 15501–15509 (2001)

Kutiev, I., Muhtarov, P., Cander, L.R., Levy, M.F.: Short-term prediction of ionospheric parameters based on autocorelation analysis, Ann. Geofis. **42**, 121–127 (1999)

Liu, R., Xu, Z., Wu, J., Liu, S., Zhang, B., Wang, G.: Preliminary studies on ionospheric forecasting in China and its surrounding area, J. Atmos. Solar-Terr. Phys. **67**, 1129–1136 (2005)

Marin, D., Miro, G., Mikhailov, A.V.: A method for foF2 short-term prediction, Phys. Chem. Earth (C), **25**, 327–332 (2000)

Mednikova, N.B.: Mid-latitude ionospheric disturbances, in "Physics of solar corpuscular fluxes and their impact on the upper atmosphere of the Earth", Academic Press of USSR, 183–244 (1957)

McKinnell, L.-A., Poole, A.W.V.: Predicting the ionospheric F layer using neural networks, J. Geophys. Res. **109**, A08308, doi:10.1029/2004JA010445 (2004)

Mikhailov, A.V., Skoblin, M.G., Förster, M.: Day-time F2-layer positive storm effect at middle and lower latitudes, Ann. Geophysicae, **13**, 532–540 (1995)

Mikhailov, A.V., Schlegel, K.: Equinoctial transitions in the ionosphere and thermosphere, Ann. Geophysicae, **19**, 783–796 (2001)

Mikhailov, A.V., Depueva, A.Kh., Leschinskaya, T.Yu.: Morphology of quiet time F2-layer disturbances: High and lower latitudes, Int. J. Geomag. Aeronom, **5**, 1–14, GI1006, doi:10.1029/2003GI000058 (2004)

Muhtarov, P., Cander, L., Levy, M., Kutiev, I.: Application of the geomagnetically correlated statistical model to short-term forecast of foF2, Proc. of the 2nd COST 251 Workshop, 30–13 March 1998, Side, Turkey, 241–245 (1998)

Pant, T.K., Sridharan, R.: Seasonal dependence of the response of the low latitude thermosphere for external forcing, J. Atmos. Solar-Terr. Phys. **63**, 987–992 (2001)

Perrone, L., de Franceschi, G.: Solar, ionospheric and geomagnetic indices, Ann. Geofis. **41**(5), 843–855 (1998)

Picone, J.M., Hedin, A.E., Drob, D.P., Aikin, A.C.: NRLMSISE-00 empirical model of the atmosphere: Statistical comparison and scientific issues, J. Geophys. Res. **107**, A12, 1468, doi:10.1029/2002JA009430 (2002)

Prölss, G.W.: Magnetic storm associated perturbations of the upper atmosphere: recent results obtained by satellite-borne gas analyzers, Rev. Geophys. Space Phys. **18**, 183–202 (1980)

Prölss, G.W.: Ionospheric F-region storms, Handbook of Atmospheric Electrodynamics, Vol. 2 (ed. Volland), CRC Press/Boca Raton, pp. 195–248 (1995)

Richards, P.G., Torr, D.G., Buonsanto, M.J., Miller, K.L.: The behaviour of the electron density and temperature at Millstone Hill during the equinox transition study September 1984, J. Geophys. Res. **94**, 16969–16975 (1989)

Richards, P.G., Torr, D.G., Buonsanto, M.J., Sipler, D.: Ionospheric effects of the March 1990 magnetic storm: comparison of theory and measurements, J. Geophys. Res. **99**, 23359–23365 (1994)

Rishbeth, H.: F-region storms and thermospheric dynamics, J. Geomag. Geoelectr. **43** (Suppl.), 513–524 (1991)

Rishbeth, H., Barron, D.W.: Equilibrium electron distributions in the ionospheric F2-layer, J. Atmos. Terr. Phys. **18**, 234–252 (1960)

Stanislawska, I., Zbyszynski, Z.: Forecasting of the ionospheric quiet and disturbed foF2, Radio Sci. **36**, 1065–1071 (2001)

Stanislawska, I., Zbyszynski, Z.: Forecasting of ionospheric characteristics during quiet and disturbed conditions, Ann. Geophysics, **45**, 169–175 (2002)

Shubin, V.N., Anakuliev, S.K.: Ionospheric storm negative phase model at middle latitudes, Geomagnetism and Aeronomy, **35**, 363–369 (1995)

Tsagouri, I., Belehaki, A.: A new empirical model of middle latitude ionospheric response for space weather applications, Adv. Space Res. **37**(2), 420–425 (2006)

Tulunay, E., Özkaptan, C., Tulunay, Y.: Temporal and spatial forecasting of the foF2 values up to twenty four hours in advance, Phys. Chem. Earth (C), **25**, 281–285 (2000)

Tulunay, Y., Tulunay, E., Senalp, E.T.: The neural network technique-2: an ionospheric example illustrating its application, Adv. Space Res. **33**, 988–992 (2004)

Wang, W., Killeen, T.L., Burns, A.G., Reinisch, B.W.: A real-time model-observation comparison of F2 peak electron densities during the Upper Atmospheric Research Collaboratory campaign of October 1997, J. Geophys. Res. **106**, A10, 21077–21082 (2001)

Wilkinson, P.J.: Predictability of ionospheric variations for quiet and disturbed conditions, J. Atmos. Solar-Terr. Phys. **57**, 1469–1481 (1995)

Wintoft, P., Cander, L.R.: Twenty-four hour predictions of foF2 using time delay neural networks, Radio Sci. **35**, 395–408 (2000)

Wrenn, G.L.: Time-weighted accumulations ap(τ) and Kp(τ), J. Geophys. Res. **92**, 10125–10129 (1987)

Wrenn, G.L., Rodger, A.S., Rishbeth, H.: Geomagnetic Storms in the Antarctic F-region. I. Diurnal and seasonal patterns for main phase effects, J. Atmos. Terr. Phys. **49**, 901–913 (1987)

Wu, J., Wilkinson, P.J.: Time-weighted magnetic indices as predictors of ionospheric behaviour, J. Atmos. Terr. Phys. **57**, 1763–1770 (1995)

Zevakina, R.A., Zhulina, E.M., Nosova, G.N., Sergeenko, N.P.: Short-term prediction manual, Materials of the World Data Centre B, Moscow, pp. 71, 1990 (in Russian)

Zevakina, R.A., Kiseleva, M.V.: F2-region parameter variations during positive disturbances related to phenomena in the magnetosphere and interplanetary medium. In: The diagnostics and modelling of the ionospheric disturbances, Nauka, Moscow, 151–167, 1978 (in Russian)

CHAPTER 3.4

MANIFESTATION OF STRONG GEOMAGNETIC STORMS IN THE IONOSPHERE ABOVE EUROPE

D. BURESOVA[1], J. LASTOVICKA[1] AND G. DE FRANCESCHI[2]

[1] *Institute of Atmospheric Physics, Prague, Czech Republic; e-mail: buresd@ufa.cas.cz*
[2] *Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy*

Abstract: The solar wind effects on Earth environment are studied for their basic science value as well as for their crucial practical impact on human technological systems. Increased dissipation of solar wind energy in the near-Earth environment is a significant source of consequent perturbations in the upper atmosphere and ionosphere. This chapter addresses the ionospheric manifestation of geomagnetic storms induced by solar wind. Changes in the electron density distribution at the ionospheric F region heights above Europe during strong-to-severe geomagnetic storms, which occurred over present solar cycle, have been analysed. As for the seasonal preference, during storm main phase only negative phases dominate in summer, while during winter occurrence of both negative and positive phases is probable. Enhancements of electron density have been sometimes observed several hours before the onset of geomagnetic storm. Also the existence of few-hours-long periods during storm main phase, when the deviation of the electron density from median was insignificant, has been observed. Independent of the sign of the storm effect on F2 region ionisation, the effect on electron density at the F1 region heights at European higher middle latitudes has been found negative, if any at all. The F1 region response to magnetic disturbances also shows substantial summer/winter asymmetry. The stormy high latitude F region is most variable compared with middle and lower middle latitudes, being strongly influenced by magnetospheric processes, in particular, strong electric fields, which are usually present during geomagnetic storms. Several specific features of the storm-time high latitude ionosphere will briefly be mentioned including behaviour of ionospheric scintillations. The comparative analysis illustrates that the improved IRI-2001 model with the activated STORM option provides better description of the ionisation distribution above Europe under geomagnetic storm conditions. Nevertheless, our results show that model not always estimates correctly the storm phase and the magnitude of the effects on F region electron density

Keywords: ionosphere, geomagnetic storm, space weather

J. Lilensten (ed.), Space Weather, 185–202.

INTRODUCTION

Earth's upper thermosphere and ionosphere over middle latitudes have been studied extensively under geomagnetically quiet and disturbed conditions for several decades. Ionospheric parameters exhibit variations on a wide range of time-scales, ranging from long-term changes down to time-scales of days, hours or minutes. Forbes et al. (2000), Rishbeth and Mendillo (2001), Rishbeth (2006) reduced possible sources of ionospheric F region variability to three main categories: solar (associated with solar photon radiation), geomagnetic and meteorological. Their results applied to a 34 years dataset of measured parameters obtained from different ionospheric stations as well as analysis of ionospheric variability over Slough for low and high solar activity led to general conclusion that the average percentage standard deviation Σ(NmF2) is 20% by day and 33% at night. It is well known that the major cause of ionospheric variability is geomagnetic storms. Geomagnetic storms are created by a variety of large disturbances originating from the Sun. Ionospheric storms result from large energy inputs to the upper atmosphere associated with geomagnetic storms. Particularly severe geomagnetic storms create complicated changes in the complex morphology of the electric fields, temperature, winds and composition and affect all ionospheric parameters. Adverse stormy conditions can cause disruption of satellite operation, navigation, and degradation of radio communications, leading to significant economic losses. Current understanding of the response of the ionosphere to geomagnetic storms has been obtained through different observations, modelling and theoretical studies. Several outstanding reviews on ionospheric reaction to geomagnetic storm-induced disturbances have been published in the last decade (e.g., Rishbeth, 1998; Buonsanto, 1999; Danilov, 2001; Ondoh and Marubashi, 2001; Lastovicka, 2002, Prölss, 2004).

Geomagnetic activity effects on the ionosphere over mid-latitudes are caused by the magnetospheric ring current as well as the out-flying effects of the polar electrojects. Fuller-Rowell et al. (1994), Buonsanto el al. (1999) and Prölss (2004) suggested that the high latitude energy input launches large-scale equatorward travelling atmospheric disturbances (TADs) that precede global storm-related equatorward meridional circulation. It was also shown by those authors that TAD can penetrate all the way to the equator and into the opposite hemisphere and even drive the poleward wind for a couple of hours. The equatorward winds cause additional upward shift of the ionospheric layers. Also changes in the neutral thermospheric composition (increase in molecular nitrogen (N_2) concentration and decrease in atomic oxygen (O) concentration) under disturbed conditions are not longer restricted in the polar upper atmosphere and are transported toward middle latitudes (Prölss, 2004). The state of ionospheric ionisation can be affected by equatorward winds directly by elevating it to higher altitudes where the chemical loss is lower and indirectly by causing changes of the neutral composition (Balan et al., 2004). Also storm-time high-latitude electric fields can penetrate promptly equatorward and undergo modification by disturbed time dynamo. Field-aligned currents cause additional disturbances.

Geomagnetic activity at high latitudes clearly differs from that at low latitudes. The disturbances at high latitudes are considerably more intense than at lower

latitudes. In vicinity of auroral oval the ionosphere is directly connected with interplanetary space by means of the geomagnetic field. Here the ionosphere is particularly sensible to external perturbations, coming from the Sun especially around a maximum of solar activity. High latitude ionosphere may become highly turbulent showing, among others, the presence of small-scale (from centimetres to meters) structures or irregularities embedded in the large-scale (tens of kilometres) ambient ionosphere. These irregularities produce short-term phase and amplitude fluctuations of radio waves, which pass through them, commonly called Amplitude and Phase Ionospheric Scintillations (e.g., Basu et al., 2002). Scintillations affect the reliability of GPS navigational systems and satellite communications and their experimental observations are needed to test the reliability of existing forecasting models as well as to build a sufficiently large data collection useful for statistical studies to be used, in turn, for creating new models. At the same time, such continuous and systematic monitoring could give an important contribution to the Space Weather activities.

Due to differences in physical mechanisms responsible for the changes in ionisation, the effects on the ionospheric F region electron density under geomagnetic storm conditions are different from those in the lower ionosphere. Effects of geomagnetic storms on the lower ionosphere, middle atmosphere and troposphere were summarised by Lastovicka (1996). Depending on storm onset time, location and season the F region response to the direct and indirect storm-induced disturbances can exhibit positive (increase in electron density) and negative (decrease in electron density) effects. Also storm effects on the electron density at different altitudes within the F region can be different (Buresova et al., 2002). Most important for radio communications and certainly best studied are changes in ionisation in the ionospheric F2 region, however, in particular those near the peak of electron density. According with scenario based on an intimate coupling between thermospheric and ionospheric storms, negative effects on the F region peak electron density over mid-latitudes are caused by changes in the neutral gas composition and positive effects are predominantly caused by TADs and thermospheric wind circulation (Prölss, 2004; Rishbeth, 1998). Travelling atmospheric disturbances are considered to be a key element for explaining an occurrence of short duration ionospheric storm positive phases. Longer lasting positive storms could be attributed to a series of rapidly successive travelling atmospheric disturbances or a modification of the global thermospheric wind circulation. Concerning the latter, summer or winter type determines, if the regular (solar-induced) and storm-induced meridional winds coincide or have opposite direction. In the case of the winter type circulation from about the vernal equinox to the autumnal equinox (October–March) during the daytime the circulation is poleward and it hinder the storm-induced circulation from expanding toward middle latitudes. The equatorward wind at middle latitudes is weakened and an additional component of the upward vertical wind appears. In the F2 region it frequently leads to an increase in the electron density and in the higher altitudes of the F1 region it should lead to the depletions of the O/N_2 ratio and thus to a decrease in ionisation. In the case of the summer type circulation from about the autumnal equinox to the vernal equinox (April–September) the two circulations coincide and the gas with decreased O/N_2 ratio is moved from the high

latitudes toward middle latitudes. This compensates the effect of the upward wind increase both in the F2 region and F1 region. Under these conditions the F1 region electron density at the higher middle latitudes remains without significant changes. During geomagnetic storm the F1 region sometimes becomes more important for the ionospheric radio wave communications, particularly under so-called G-conditions, when the F2 region is substantially reduced and is screened by the F1 region, which then serves also as the radio wave reflecting layer. This was the main reason why the effects on the F1 region ionisation induced by geomagnetic storm had been studied in COST 271 (Bencze et al., 2004). Recent progress and outstanding questions about changes in ionospheric F1 region ionisation under storm conditions have been discussed by Buresova et al. (2002), Mikhailov and Schlegel (2003) and Buresova (2005).

Nowadays large-scale numerical simulations of the ionospheric response to storm induced disturbances show that there is an increasing understanding of the storm scenarios/mechanisms and influences of storm onset time, intensity and season on the consequent changes in the ionosphere (e.g., Fuller-Rowell et al., 1994; Araujo-Pradere et al., 2002; Araujo-Pradere and Fuller-Rowell, 2002). Nevertheless, some features of this phenomenon are still not clear and hardly predictable. It was noticed by Codrescu et al. (1997) that one of the major limitations to upper atmosphere modelling and forecasting is the accuracy of the determination of the high-latitude forcing (e.g., interhemispheric penetration of the high-latitude electric fields and auroral precipitation). Furthermore the authors also pointed out that these influences could not be uniquely separated from the effects of neutral composition and ionospheric layer heights. Szuszczewicz (1998) pointed out that the agreement between observations and model-generated values tends to be more qualitative than quantitative, and the quantitative tests tend to be insufficiently reliable. Fuller-Rowell et al. (2000) showed how difficult is to model the storm effects on the ionosphere above single station. One of the main conclusions of the study was that current ability to predict ionospheric response to storm-induced disturbances is significantly lower than recent knowledge of the physical processes.

In the present study we report observational results of strong-to-severe ionospheric storms that occurred in the period 1995–2005 over European region. We examine these storms in some details, focusing on when they occur, on ionospheric height profile of their effects, similarities and unexpected differences in their morphology. Presented results show that there are still problems unsolved, like occasional enhancements of F2 region peak electron density before the onset of geomagnetic storms, or forecasting an appearance of positive and negative phases within stormy period over middle latitudes.

DATA SET AND ANALYSIS METHOD

Ionospheric F region response to geomagnetic storms was studied analysing changes in the state of ionisation at the peak of electron density as well as at different bottomside F region altitudes. The created database incorporates

65 strong-to-severe geomagnetic storms for 11 years-long period from 1995 to 2005. The study was carried out using electron density N(h) profiles available in the World Data Centre for Solar-Terrestrial Physics at Chilton (WDC1), http://www.ukssdc.ac.uk/wdcc1/ionosondes/secure/iono_data.shtml and in the COST 271 database http://www.wdc.rl.ac.uk/cgi-bin/digisondes/cost_database.pl. Parameters used were mainly hourly interval resolution values of the F2 region peak electron density NmF2 and electron densities N(e) at different F region altitudes obtained from electron density N(h) profiles for the initial, main and partly for recovery phases of the analysed geomagnetic storms including at least two days before the storm onset and NmF2 monthly medians. N(h) profile simulation has been performed by using online information available at the IRI web site http://nssdc.gsfc.nasa.gov/space/model/models/iri.html. Ionospheric stations included in the analysis are listed in Table 1. The stations cover geomagnetic latitude from 36.4 °N to 67.0 °N.

The onset of a geomagnetic storm, its intensity and different phases are defined using Dst index in the following way:

- Sudden storm commencement (SSC) has been chosen as an indicator of the storm onset. The storm main phase is defined by the decrease of Dst (decrease in magnetic field strength) and the subsequent recovery phase by its gradual reversion to quiet conditions.
- Strong storm conditions were defined when Dst ≤ -100 nT for at least four consecutive hours.
- Storm conditions when Dst < -50 nT.

The F region response to storm-induced disturbances is described in terms of deviations of N(e) from the quiet time median values, i.e. δN(e).

$$\delta \mathrm{N}(e) = \{\mathrm{N}(e)_m - \mathrm{N}(e)_{med}\}/\mathrm{N}(e)_{med},$$

where $\mathrm{N}(e)_m$ and $\mathrm{N}(e)_{med}$ are measured and monthly median values of F region electron concentration. Deviations less than 20% have not been taken into account.

Table 1. List of ionospheric stations involved in study

Name of the ionospheric station	Geographic latitude and longitude	Magnetic latitude and longitude
Tromso	69.70N, 19.0E	67.0N, 117.5E
Juliusruh	54.60N, 13.4E	54.3N, 99.7E
Chilton	51.6N, 358.7E	54.1N, 83.2E
Pruhonice	50.0N, 14.6E	49.7N, 98.5E
Rome	41.9N, 12.5E	42.3N, 93.2E
Ebro	40.8N, 0.5E	46.3N, 80.9E
Athens	38.0N, 23.6E	36.4N, 102.5E
El Arenosillo	37.1N, 353.2E	41.4N, 72.3E

OCCURRENCE FREQUENCY OF POSITIVE AND NEGATIVE PHASES OF IONOSPHERIC STORMS

Many years studies of geomagnetic storm effects on the ionosphere gave a typical course of the F region response described by Prölss (1995) and summarised by Rishbeth and Field (1997) with an initial phase with enhanced peak electron density NmF2 lasting a few hours after the geomagnetic storm onset (usually SSC). Subsequent main phase of the storm lasts a day or more. A recovery phase of the storm could last several days. According to long-term ionospheric observations above European middle latitudes, storm-induced variations of the F2 region ionisation during storm main phase often change from large enhancements (positive phase) to depletions (negative phase). Such a change of sign of the storm effect makes a systemic description and prediction of the disturbed ionosphere rather complicated. Strong longitudinal and latitudinal asymmetries or the completely different storm-induced disturbance behaviour of the ionospheric F2 region above two comparable locations are frequently observed (Prölss, 1995). Moreover, the distribution of storm effects may vary substantially from one event to another.

A statistical picture of the occurrence of negative and positive phases during analysed strong-to-severe geomagnetic storms main phase for the period from 1995 to 2005 for three European stations is given in Fig. 1. Our results show that the changeover from one type of the effects to the other is more common for winter than for summer, and the occurrence of such behaviour increases with decreasing latitude. During summer half of the year all three stations display more frequent appearance of only negative effect during the entire main phase of

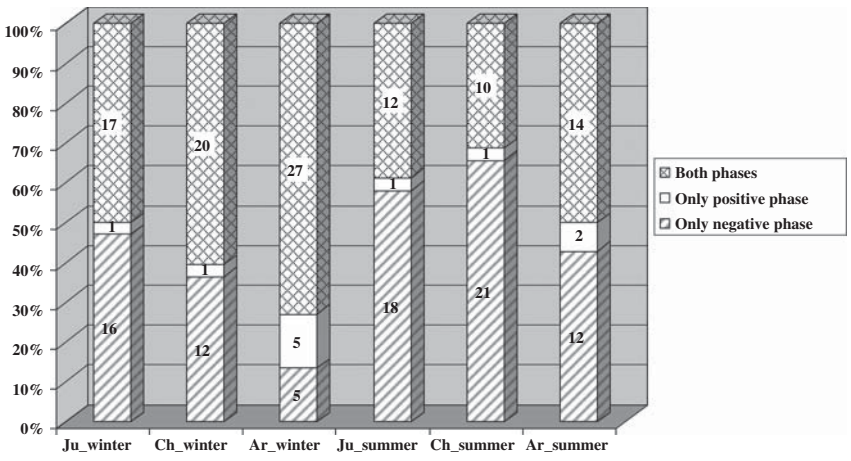


Figure 1. Occurrence of negative and positive phases during the geomagnetic storm main phase above three European stations Juliusruh, Chilton and El Arenosillo for winter and summer half of the year (according to the type of thermospheric circulation) for the period 1995–2005

the analysed storms. Considering both winter and summer periods, higher-middle latitude station Juliusruh shows much higher appearance of only negative phase (34 events) or both phases (29 events) during strong-to intense geomagnetic storm than the appearance of only positive phase (2 events). Summer-winter difference in storm phase appearance above Juliusruh is relatively small. Higher-middle latitude station Chilton shows a similar distribution of storm phases and more frequent appearance of negative phase in summer compared with Juliusruh. Lower-middle latitude station El Arenosillo exhibits a shift to more frequent appearance of the positive phase, especially during wintertime geomagnetic storms. Among storms with only positive phase, wintertime storms dominate.

In Fig. 2 alternation of positive and negative phases are shown for five European stations ordered from higher middle to lower-middle latitudes for February 1999 (left side panels) and October 2000 (right side panels) geomagnetic storms. During the main phase of the February 1999 geomagnetic storm El Arenosillo displayed prevailing positive effect on NmF2, while over higher-middle latitudes both negative and positive phases of the storm have been observed. Large latitudinal differences in ionospheric storm-induced disturbances are also well seen from plots representing the course of October 2000 storm effects on F2 region peak ionisation (right side panels of Fig. 2). The largest effect (positive) has been observed above Juliusruh on the storm onset day followed by both negative and positive

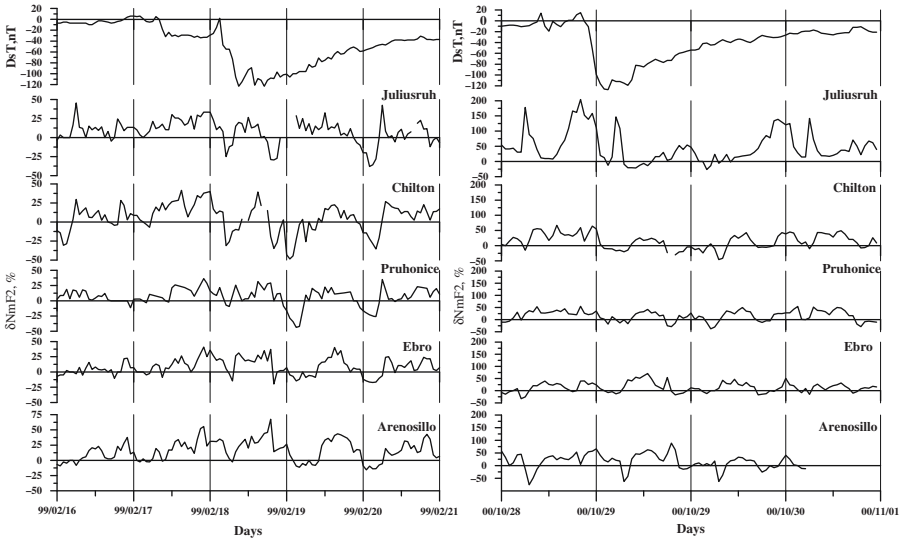


Figure 2. Effects of February 1999 (left side panels) and October 2000 (right side panels) geomagnetic storms on F2 layer peak electron density NmF2. Top plots panels illustrate hourly D_{st} variation for entire period analysed. Plots below display the F2 region response to storm-induced disturbances described in terms of deviations of NmF2 from the quiet time median values, i.e. δNmF2, above five European stations Jusliusruh, Chilton, Pruhonice, Ebro and El Arenosillo. Time is in UT

effects during the next day (storm maximum day). Compared with Juliusruh, Chilton and Pruhonice show insignificant deviation from monthly median during storm maximum day ($\delta < 25\%$), except midnight deviation for Chilton. Both positive and negative effects of larger magnitude were observed above Ebro and El Arenosillo.

TEC AND SCINTILLATIONS AT HIGH LATITUDES

Since 2003 the GPS Total Electron Content (TEC) and scintillations monitoring is performed in the Northern Europe at Ny-Ålesund (Svalbard, Norway; geographic coordinates 78.9 °N, 11.9 °E) in the frame of ISACCO (Ionospheric Scintillations Arctic Campaign Coordinated Observations) project (De Franceschi et al., 2003). A modified GPS receiver is used consisting of a NovAtel dual-frequency receiver with special firmware comprises the major component of a GPS signal monitor, specifically configured to measure amplitude and phase scintillation from the L1 frequency GPS signals, and ionospheric TEC from the L1 and L2 frequency GPS signals. Amplitude scintillation is monitored by computing the S4 index which is the standard deviation of the received power normalised by its mean value over 60 seconds interval. It is derived from the detrended received signal intensity. Phase scintillation computation is accomplished by monitoring the standard deviation σ_ϕ of the detrended carrier phase. It is computed over 1, 3, 10, 30 and 60-second intervals. TEC is computed every second from phase-smoothed L1 and L2 pseudorange differences. For detrending the 50 Hz raw phase and amplitude measurements, a high-pass six-order Butterworth filter and a high pass filter are used, respectively (Van Dierendonck et al., 1993 and references therein). A fixed choice of a 0.1 Hz 3-dB cut-off frequency for both phase and amplitude filtering is used.

Three events have been selected to show the manifestation of external disturbances on high latitude ionosphere as derived from the GPS Ny-Alesund station, i.e. 30 October 2003 between 21.00 and 21.59 UT, 20 November 2003 between 20.00 and 20.59 UT, and 15 May 2005 between 11.00 and 11.59 UT. An extremely large solar eruption, the biggest for decades, on 28 October 2003 caused an intense geomagnetic storm. A second solar eruption on 29 of October resulted in a re-intensification of the storm about a day later. On 20 November 2003, a Coronal Mass Ejection (CME) shock from the activity on 18 November produced severe geomagnetic conditions. On 15 May 2005, effects from the large, full halo CME from the 13 May M8.0 flare arrived at Earth, geomagnetic activity increased significantly and the field was at minor to severe storming. In a recent investigation by Mitchell et al. (2005), the 30 October scintillation event as seen by GPS Ny-Alesund station has been associated to ionospheric plasma structures convecting across the polar cap from the American sector in an antisunward flow consistent with the expected convection pattern with the southward IMF (Carlson, 1994; Coker et al., 2004). Moreover severe scintillations characterised by high values of σ_ϕ and S4 have been found coinciding with strong local horizontal gradients of TEC (Fig. 3 – top

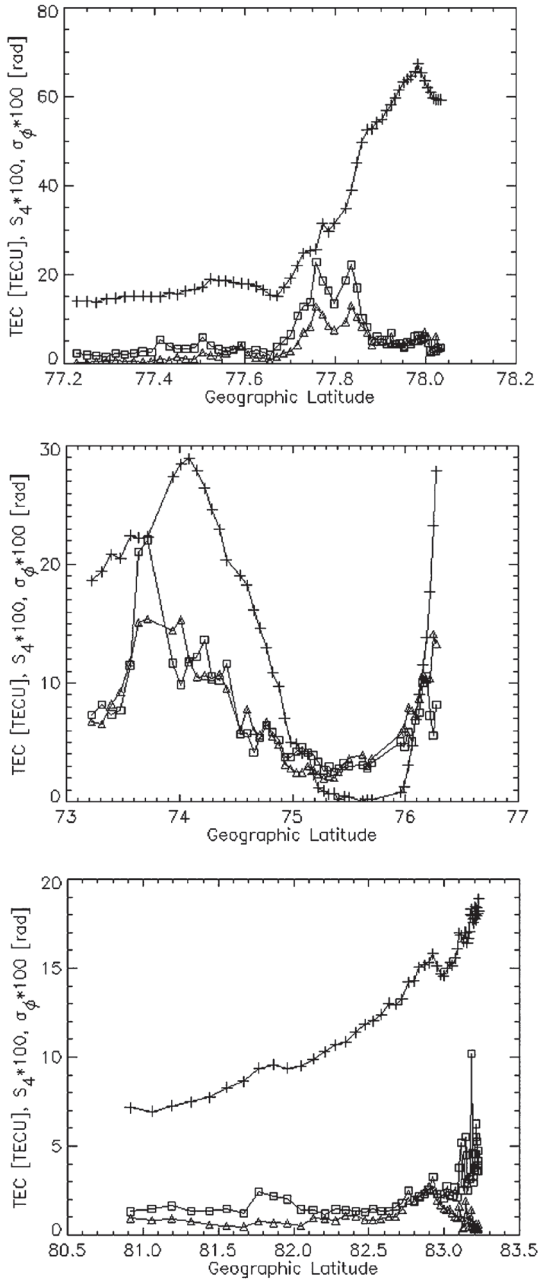


Figure 3. Equivalent vertical TEC (crosses), phase scintillation (squares) and amplitude scintillation (triangles) recorded by the Svalbard receiver for selected periods and satellites. From top to bottom: 30 October 2003 between 21.00 and 21.59 UT, PNR = 31 (from Mitchell et al., 2005); 20 November 2003 between 20.00 and 20.59 UT, PNR = 3; 15 May 2005 between 11.00 and 11.59 UT, PNR = 30

panel), a characteristic of the edge of polar-cap patches. In Fig. 3 the "equivalent vertical" components of TEC (crosses), σ_ϕ (squares) and S4 (triangles) are plotted for selected satellites that experienced scintillations during the three events as a function of the geographic latitude of the subionospheric point calculated for the assumed ionospheric altitude 350 km. A geometrical correction has been applied to the TEC, phase and amplitude data from the Svalbard receiver that allows the "equivalent vertical" components of TEC, σ_ϕ and S4 to be intercompared. In all cases only the signals coming from satellites with an elevation angle greater than 25° and with a time of lock greater than 240 seconds have been taken into account. The 20 November 2003 (Fig. 3 – middle panel) and 15 May 2005 (Fig. 3 – bottom panel) events do not show strong TEC gradients as in the case of 30th of October. The regions of both phase and amplitude scintillation occur on the increasing TEC values. In the case of 15th of May it can been noted that: (*i*) only the phase scintillation increases with TEC and (*ii*) the σ_ϕ values are generally lower than for other two events.

IRI SIMULATION OF IONOSPHERIC STORMS

The well-known empirical International Reference Ionosphere model, IRI, is being improved and updated continuously after evaluation of new results at the annual workshops (Rawer, 1994, Araujo-Pradere et al., 2002; Bilitza, 2001, 2003). The recently updated IRI-2001 version contains a geomagnetic activity dependence option based on Time Empirical Ionospheric Correction Model (STORM) (Araujo-Pradere et al., 2002). The model was designed on the basis of the study of the consistent and repeatable storm-time ionospheric response characteristics. STORM design includes seasonal dependence in the migration of the composition bulge by the global wind field and a non-linear dependence on the integrated time history of the geomagnetic activity index A_p. Quantitative validation studies of the STORM model (Araujo-Pradere and Fuller-Rowell, 2002; Araujo-Pradere et al., 2004a and 2004b; Buresova et al., 2004) showed that the new option is an improvement on the previous IRI-95 version, which had no geomagnetic activity dependence, nevertheless some recent comparative analyses pointed out areas where improvement is still needed in the empirical description. Buresova et al. (2002) have analysed the F-region electron density distribution over Europe under the both geomagnetically quiet and disturbed conditions. Comparison of the measured values with those model-predicted showed that the IRI does not always represent correctly the actual N(h) profile. It is particularly valid during strong-to-intense geomagnetic storms. Figure 4 shows an example illustrating medium degree of coincidence between model and observations.

Araujo-Pradere and Fuller-Rowell (2004) evaluated the quality of the storm-time correction by comparing the model with the observed storm effects for 15 ionospheric stations during all the significant geomagnetic events in 2000 and 2001. They pointed out two main areas, where challenges remain for the empirical storm-time correction model: the initial rapid increase in electron density at the storm

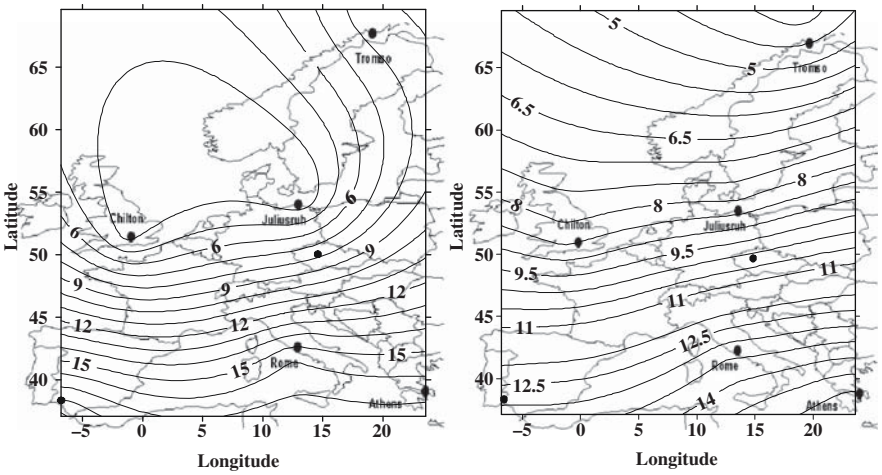


Figure 4. Observed (left side panel) and IRI-2001generated (right side panel) NmF2 over Europe, for storm maximum day on October 29, 2003 at 12:00 UT

onset, and the regional dependence of the storm negative phase where one longitude sector is hit harder than another sector. In both cases, additional information is still required.

A PRE-STORM ENHANCEMENT OF NMF2

The analysis of the NmF2 course over European middle latitudes under pre-storm and storm conditions showed that some events are preceded by an increase of NmF2, which occurs several hours before the storm onset. We found 15 such storms among 65 analysed events. Typical examples of pre-storm enhancements of NmF2 are shown in Fig. 5 as a substantial and several hours lasting increase of NmF2 not associated with any corresponding change in D_{st}. In the case of October 2003 event (left side panel), larger pre-storm enhancement of NmF2 has been observed for higher-middle latitude stations Juliusruh and Chilton than for lower-middle latitude station Arenosillo, while during the day prior to May 1997 strong geomagnetic storm (right side panel) Juliusruh showed weaker pre-storm enhancement comparing with Chilton and Arenosillo. The enhancement of the F region peak electron density prior to October 2003 event onset has also been reported by Kane (2005). He pointed out a large long-lasting positive deviation on 28 October, a day before the SSC of geomagnetic super-storm above several ionospheric stations located at different longitudes and latitudes of both hemispheres. As for the IRI-2001 simulation, it does not recognise the pre-storm enhancement (Fig. 6). Therefore the coincidence between model simulations and observations is poor, worse than for the storm itself (Fig. 4 versus Fig. 6).

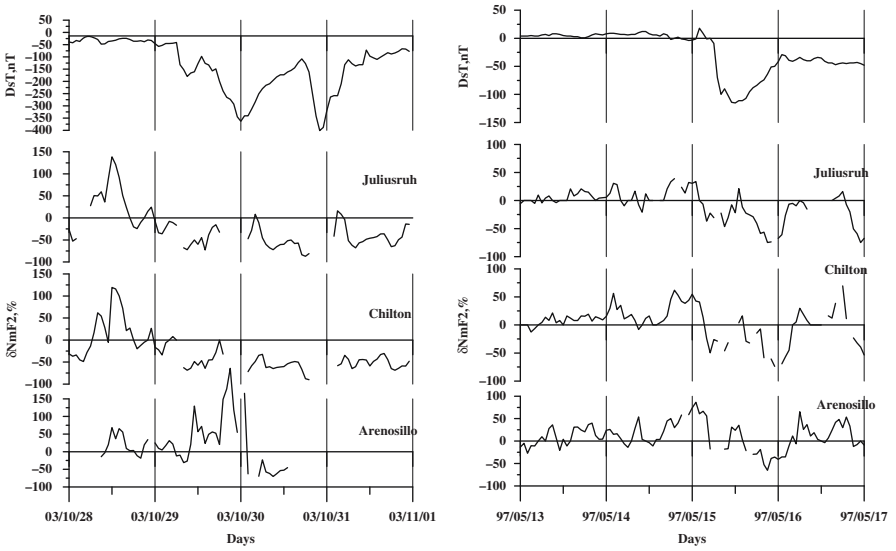


Figure 5. Effects of October 2003 (left side panel) and May 1997 (right side panel) geomagnetic storms on F2 layer peak electron density NmF2. Top plots illustrate hourly D_{st} variation for entire stormy periods including day before the storm onset. Plots below display the F2 region response to storm-induced disturbances described in terms of deviations of NmF2 from the quiet time median values, i.e. δNmF2, including changes of NmF2 for at least one day prior to storm onset

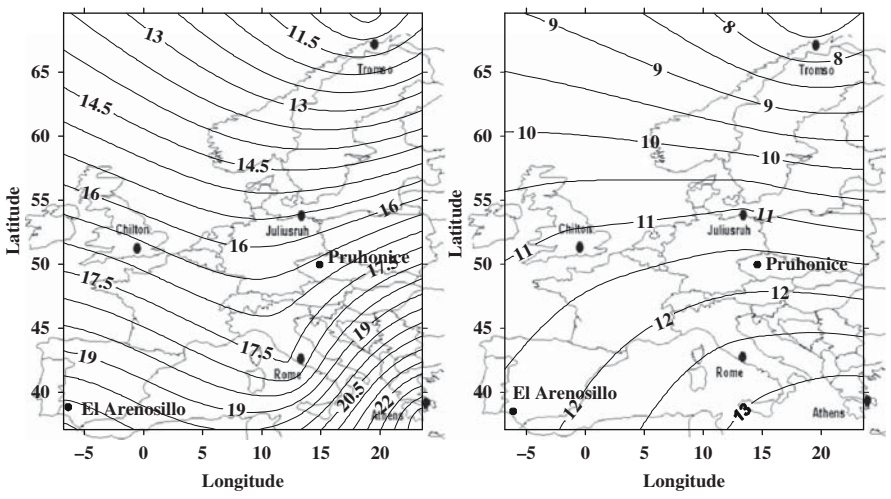


Figure 6. Observed (left side panel) and IRI-2001generated (right side panel) NmF2 over Europe, October 28, 2003 at 12:00 UT – positive "quiet" disturbance before the beginning of the storm – worse IRI-observation agreement

HEIGHT PROFILE OF STORM EFFECT UP TO PEAK OF THE F2 REGION

Inspection of a couple of events showed that in the case of positive storm the maximum effect could be observed below the peak of the F2 region (decrease of the electron density at about the upper boundary of the F1 region) accompanied by the considerably smaller effect of different sign (increase of the electron density) at heights of the peak of the F2 region, as Fig. 7 illustrates for Chilton and for El Arenosillo during storm of November 1998. The height profile of the storm effects indicates that on 13 November 1998 a maximum for Chilton was below hmF2. Compared with Chilton, El Arenosillo shows smaller decrease of electron density below 220 km and larger opposite storm effect on NmF2. During the negative storm maximum day at 12 August 2000 both Chilton and El Arenosillo display the magnitude of the storm effects continuously increasing with altitude. Broader investigations of this phenomenon are underway.

STORM EFFECTS IN THE F1 REGION

There are a couple of physical processes, which could contribute to the observed effect of geomagnetic storms on the state of ionisation in the bottomside F region at middle latitudes. According to current theories, the main physical mechanisms controlling the F1 region response to geomagnetic storm are photochemical processes and the proportion of atomic and molecular ions, which is related with neutral composition (O, O_2, N_2) seasonal and storm time variation (Buresova et al., 2002; Mikhailov and Schlegel, 2003). Ionospheric F1 region is the transition region, where the ion composition is changing with height very rapidly. The transition height between the region dominated by molecular ions (NO^+ and O_2^+) and the region where the atomic ions O^+ dominate (the transition height between the α-type and β-type recombination) lies at about 160–200 km. Figure 8 illustrates the annual noontime course of O and N_2 concentration at the height of 190 km for 1998 for Chilton calculated by online computation through the MSIS 90 website. The ratio $n(O)/n(N_2) < 1$ during summer, May to August, means that the transition height between the α-type and β-type recombination is located above 190 km. In April and September (still the summer half of the year according to the type of thermospheric circulation) the ratio $n(O)/n(N_2)$ indicates that the transition height is at about 190 km. On the other hand, the ratio $n(O)/n(N_2)$ indicates the winter transition height (approximately from October to March – winter type of the thermospheric circulations) to be well below 190 km. Buresova and Lastovicka (2001) analysed effects of a few geomagnetic storms in electron density at F1 heights during daytime, based on data of selected European ionosondes. Their results suggested that the F1 region response to geomagnetic storms exhibits systematic seasonal behaviour and depends partly on latitude. Figure 9 shows the penetration of the storm effects into F1 region for higher-middle latitude station Chilton and lower-middle latitude station Ebro. Comparing with Ebro, Chilton displays much stronger effect during

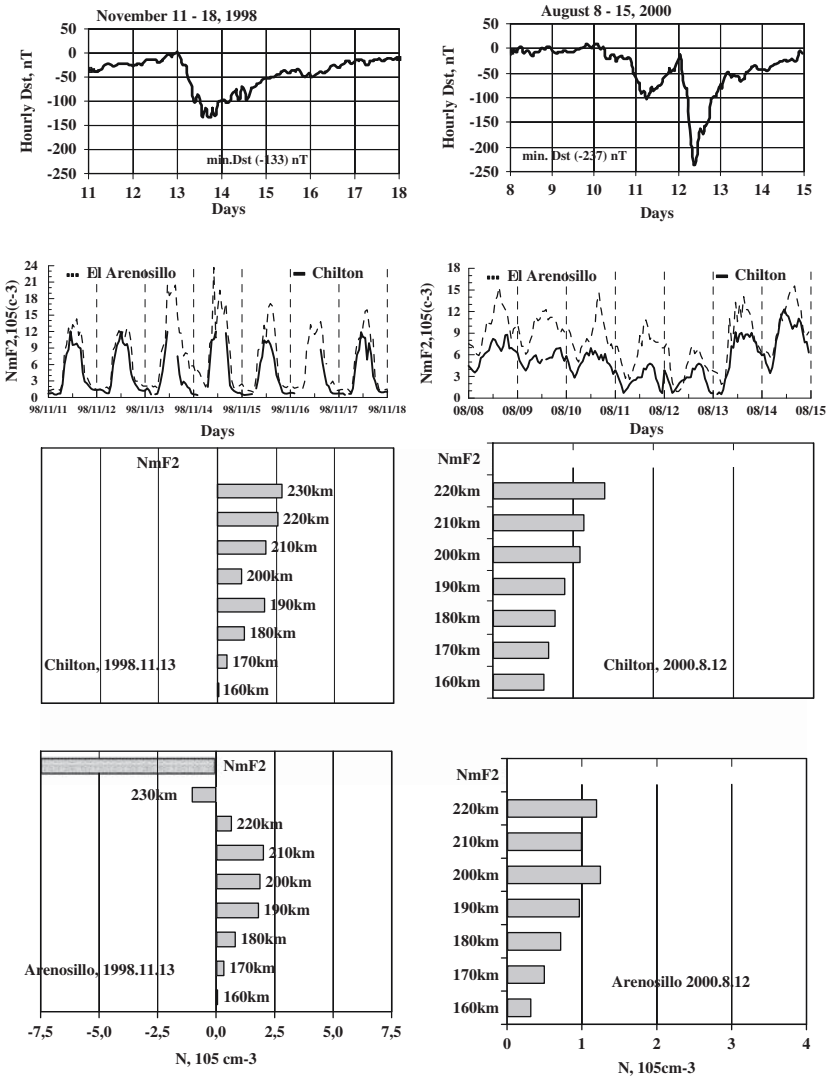


Figure 7. Geomagnetic storms of November 1998 (left side panels) and August 2000 (right side panels). Hourly D_{st} indices (top panels) and courses of NmF2 (the other two panels) for Chilton and El Arenosillo for both events. Changes in the ionospheric bottomside F region ionisation at every ten kilometres of altitude (differences between mean electron density of the pre-storm quiet days and the electron density during storm maximum day at 11:00–13:00 UT) above Chilton and El Arenosillo

fall and winter than during spring and summer. Another important finding is that the pattern of the response of the F1 region at European higher middle latitudes, which is a decrease in electron density, does not depend on the type of response of the F2 region or on solar activity (Buresova et al., 2002). The magnitude of the

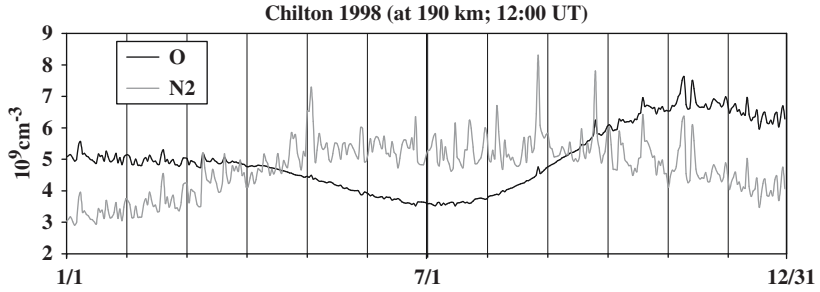


Figure 8. The annual noontime course of O and N_2 concentration at the height of 190 km for 1998 for Chilton

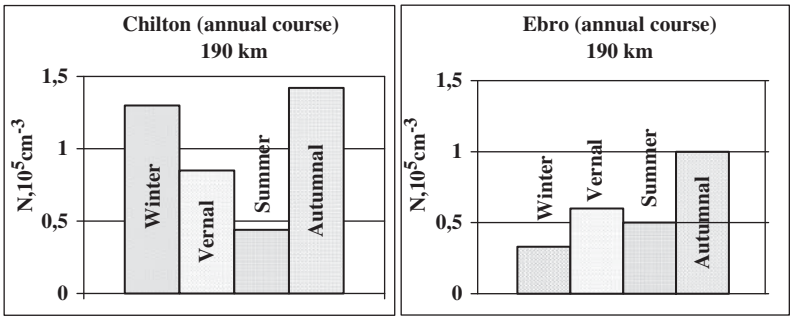


Figure 9. The seasonal dependence of the strong-to severe geomagnetic storm effects on electron density (difference between mean electron density of the two pre-storm quiet days and the electron density during storm main phase at 11:00–13:00 UT) at the height of 190 km for Chilton (left side panel) and Ebro (right side panel). Five events for winter and summer each and thirteen events for spring and autumn each have been involved into analysis

storm effects in F1 region increases with altitude. Mikhailov and Schlegel (2003) have reported similar results.

CONCLUSIONS

According to present-day understanding, the storm effects in the mid-latitude F2 region ionosphere are predominantly attributed to ionospheric response to storm effects in the thermosphere, and the effects in the lower ionosphere are predominantly caused by the storm-associated precipitating energetic particles.

A statistical picture of the strong-to-severe geomagnetic storms effects on NmF2 during storm main phase for the period from 1995 to 2005 shows that above European middle latitudes the changeover from positive to negative phase of the ionospheric storm is rather common and appears more frequently during winter than

during summer half of the year. The occurrence of such behaviour increases with decreasing latitude. Appearance of only negative effect during the entire main phase is more frequent during summer half of the year and at higher middle latitudes. Summer–winter difference in rare only positive storm phase appearance is relatively small and seems to be more pronounced for lower-middle latitudes.

In spite of many years of investigations of effects of geomagnetic storms on the F region ionosphere, there are still many open questions, like the pre-storm enhancements of *fo*F2, storm effects on the bottomside F region (they are relatively well-known and understood in the F2-region maximum), or model (IRI) reproducibility of the observed geomagnetic storm-related effects. We are able to predict appearance of ionospheric storms based on geomagnetic storm predictions, but we cannot predict reliably phase (positive or negative) of the storm.

Polar observations are scarce. The ISACCO project could offer a good opportunity to the users/scientific community by contributing to the ground based monitoring of ionospheric scintillations during storms, when the ionosphere deviates considerably from the average behaviour represented in empirical model. These observations can serve for improving the understanding of the physics of ionospheric irregularities since their behaviour is unpredictable with current ionospheric models albeit their effect on radio communications can be quite destructiv.

ACKNOWLEDGEMENTS

This work has been supported by Grant 1QS300120506 of the Academy of Sciences of the Czech Republic. One of the authors thanks the PNRA (National Antarctic Research Program-Italy), the Polarnet project (CNR-Italy), the NOAA/SEC (Boulder, CO, USA).

REFERENCES

Araujo-Pradere, E.A., Fuller-Rowell, T.J., Codrescu, M.V.: STORM: An empirical storm-time ionospheric correction model −1. Model description, Radio Sci. **37** (5), doi: 10.1029/2001RS002467 (2002)

Araujo-Pradere, E.A., Fuller-Rowell, T.J.: Storm: An empirical storm-time ionospheric correction model – 2. Validation, Radio Sci. **37** (5), 1071 (2002)

Araujo-Pradere, E.A., Fuller-Rowell, T.J.: Time Empirical Ionospheric Correction Model (STORM) response in IRI 2000 and challenges for empirical modelling in the future, Radio Sci. **39** (1), RS1S24, doi: 10.1029/2002RS002805 (2004)

Araujo-Pradere, E.A., Fuller-Rowell, T.J, Codrescu, M.V. Anghel, A.: Evaluation and prospects for storm-time corrections in the International Reference Ionosphere, Adv. Space Res, **33**, 908–909 (2004a)

Araujo-Pradere, E.A., Fuller-Rowell, T.J., Bilitza, D.: Ionospheric variability for quiet and disturbed conditions, Adv. Space Res, **34**, 1914–1921 (2004b)

Balan, N., Kawamura, S., Nakamura, T., Yamamoto, M., Fukao, S., Igarashi, K., Muruyama, T., Shiokawa, K., Otsuka, Y., Ogawa, T., Alleyne, H., Watanabe, S., Murayama, Y.: Simultaneous mesosphere/lower thermosphere and thermospheric F region observations during geomagnetic storms, J. Geophys. Res. **109**, A04308, doi:10.1029/203JA009982 (2004)

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** (16), 1745–1754 (2002)

Bencze, P., Buresova, D., Lastovicka, J., Marcz, F.: Behaviour of the F1-region, and Es and spread-F phenomena at European middle latitudes, particularly under geomagnetic storm conditions. Annals Geophys. COST 271 Action Final Report, **47** (2/3), 1131–1143 (2004)

Bilitza, D.: International Reference Ionosphere 2000, Radio Sci. **36** (2), 261–275 (2001)

Bilitza, D.: International reference ionosphere 2002: Examples of improvements and new features, Adv. Space Res, **31** (3), 757–767 (2003)

Buonsanto, M.J.: Inospheric storms – A review, Space Sci. Revs., **88**, 563–601 (1999)

Buresova, D., Lastovicka, J.: Changes in the F1 region electron density during geomagnetic storms at low solar activity, J. Atmos. Solar-Terr. Phys. **63**, 537–544 (2001)

Buresova, D., Lastovicka, J., Altadill, D., Miro, G.: Daytime electron density at the F1 region in Europe during geomagnetic storms, Ann. Geophysicae, **20**, 1007–1021 (2002)

Buresova, D., Lastovicka, J., Altadill, D., Miro, G.: Predicted and measure botomside F-region electron density and variability of the D_1 parameter under quiet and disturbed conditions over Europe, Adv. Space Res, **34**, 1973–1981 (2004)

Buresova, D.: Effects of geomagnetic storms on the botomside ionospheric F region. Adv. Space Res, **35**, 429–439 (2005)

Carlson, H.C.: The dark polar ionosphere: Progress and future challenges, Radio Sci. **29**, 157–166 (1994)

Codrescu, M.V., Fuller-Rowell, T.J., Kutiev, I.S.: Modelling the F layer during specific geomagnetic storms, J. Geophys. Res. **102**, 14 315–14 329 (1997)

Coker, C., Bust, G.S., Doe, R.A., Gaussiran, T.L., II: Highlatitude plasma structure and scintillation, Radio Sci. **39**, RS1S15, doi:10.1029/2002RS002833 (2004)

Danilov, A. D.: F2-region response to geomagnetic disturbances, J. Atmos. Solar-Terr. Phys. **63**(5), 441–449 (2001)

De Franceschi, G., Romano, V., Alfonsi, L., Pezzopane, M., Zolesi, B.: ISACCO (Ionospheric Scintillations Arctic Campaign Coordinated Observations) project at Ny-A° lesund, in: proceedings of "Atmospheric Remote Sensing using Satellite Navigation Systems, Special Symposium of the URSI Joint Working Group FG", Matera, Italy (Ottobre 2003)

Forbes, J. M., Scott, E. P., Zhang, X.: Variability of the ionosphere, J. Atmos. Solar-Terr. Phys. **62**, 685–693 (2000)

Fuller-Rowell, T.J., Codrescu, M.V., Moffett, R.J., Quegan, S.: Response of the thermosphere and ionosphere to geomagnetic storms, J. Geophys. Res. **99**, 3893–3914 (1994)

Fuller-Rowell, T.J., Codrescu, M.C., Wilkinson, P.,: Quantitative modelling of the ionospheric response to geomagnetic activity, Ann. Geophysicae, **18**, 766–781 (2000)

Kane, R.P.: Ionospheric foF2 anomalies during some intense geomagnetic storms, Ann. Geophysicae, **23**, 2487–2499 (2005)

Lastovicka, J.: Effects of geomagnetic storms in the lower ionosphere, middle atmosphere and troposphere, J. Atmos. Solar-Terr. Phys. **58**(7), 831–843 (1996)

Lastovicka, J.: Monitoring and forecasting of ionospheric space weather – effects of geomagnetic storms. J. Atmos. Solar-Terr. Phys. **63**, 697–705 (2002)

Mikhailov, A.V., Schlegel, K.: Geomagnetic storm effects at F1-layer heights from incoherent scatter observations, Ann. Geophysicae, **21**, 583–596 (2003)

Mitchell, C.N., Alfonsi, L., De Franceschi, G., Lester, M., Romano, V., Wernik, A.W.: GPS TEC and scintillation measurements from the polar ionosphere during the October 2003 storm, Geophys. Res. Lett. **32**, L12S03, doi:10.1029/2004GL021644 (2005)

Ondoh, T., Marubashi, K.: Science of space environment, Ohmsha, Ltd, Tokyo, Japan (2001)

Prölss, G.W.: Ionospheric F-region storms. In: Volland, H. (ed.), Handbook of Atmospheric Electrodynamics 2, 195–248. CRC Press, Boca Raton, FL (1995)

Prölss, G.W.: Physics of the Earth's space environment, Springer-Verlag Berlin Heidelberg, Germany (2004)

Rawer, K.: Problems arising in empirical modelling of the terrestrial ionosphere, Adv. In Space Res. **14** (12), 12(7)–12(16) (1994)

Rishbeth, H., Field, P.R.: Latitude and solar-cycle patterns in the response of the ionosphere F2-layer to geomagnetic activity, Adv. Space Res. **20** (9), 1689–1692 (1997)

Rishbeth, H.: How the thermospheric circulation affects the ionospheric F2 layer, J. Atmos. Solar-Terr. Phys. **60**, 1385–1402 (1998)

Rishbeth, H., Mendillo, M.: Patterns of F2-layer variability, J. Atmos. Solar-Terr. Phys. **63**, 1661–1680 (2001)

Rishbeth, H.: F-region links with the lower atmosphere?, J. Atmos. Solar-Terr. Phys. **68**, 469–478 (2006)

Szuszczewicz, E.P., Lester, M., Wilkinson, P., Blanchard, P., Abdu, M., Hanbaba, R., Igarashi, K., Pulinets, S., Reddy, B.M.: A comparative study of global ionospheric response to intense magnetic storm conditions, J. Geophys. Res. **103** (A6), 11,665–11,684 (1998)

Van Dierendonck, A. J., Klobuchar, J., Hua, Q.: Ionospheric scintillation monitoring using commercial single frequency C/A code receivers, in ION GPS-93 Proceedings: Sixth International Technical Meeting of the Satellite Division of the Institute of Navigation, pp. 1333–1342, Inst. of Navig., Salt Lake City, Utah (1993)

CHAPTER 3.5

EFFECTS OF SCINTILLATIONS IN GNSS OPERATION

Y. BÉNIGUEL AND J.-P. ADAM
IEEA, Courbevoie, France

INTRODUCTION

At altitudes above about 80 km, molecular and atomic constituents of the Earth's atmosphere are energized and ionized by solar ultraviolet (UV) radiation and additionally by energetic electrons of solar and magnetospheric origin in particular at high latitudes. The resulting plasma whose density peaks around 300 km is a dispersive and due to the geomagnetic field also an anisotropic propagation medium for radio waves. The plasma interacts with radio waves to degrade transionospheric propagation at the VHF up to the C-band frequency range.

Whereas medium scale variations in time and space such as Traveling Ionospheric Disturbances (TID's) mainly impact reference networks, local small scale irregularities may cause radio scintillations inducing severe signal degradation and even loss of lock at any user. Amplitude scintillations and phase fluctuations are produced by refractive and diffractive scatter by ionospheric plasma-density irregularities, especially at equatorial and auroral-to-polar latitudes.

The study of this ionosphere variability is one of the aims of the space weather studies for navigation systems. They divide into two kinds of studies: the Total Electron Content (TEC) variability and the scintillations (cf Fig. 1). The TEC is defined as the sum of electrons along a line, usually considered at the zenith of one observation point. Magnetic storms and traveling disturbances may greatly enhance the TEC variability and affect consequently the navigation systems. Intensity of magnetic storms is predominant at high latitudes whereas traveling disturbances mainly occur at mid-latitudes.

Geographically, the two areas most affected by scintillations are equatorial (+/− 30 degrees magnetic latitude) and auroral (around the poles) regions. In solar active years – during the peak of the 11-year solar cycle, both, the diurnal scintillations

J. Lilensten (ed.), Space Weather, 203–216.

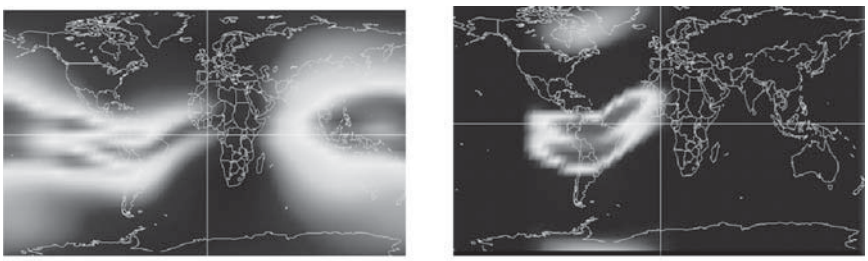


Figure 1. TEC and scintillations maps

(occurring in the equatorial region after sunset) as well as the ionospheric-storm induced scintillations can cause significant link outages leading to a degradation of navigation tasks.

Scintillations at Equatorial Regions

A typical example of scintillations is presented on Fig. 2. for L1 frequency, both for phase and scintillations. These data have been recorded at Ascencion Island.

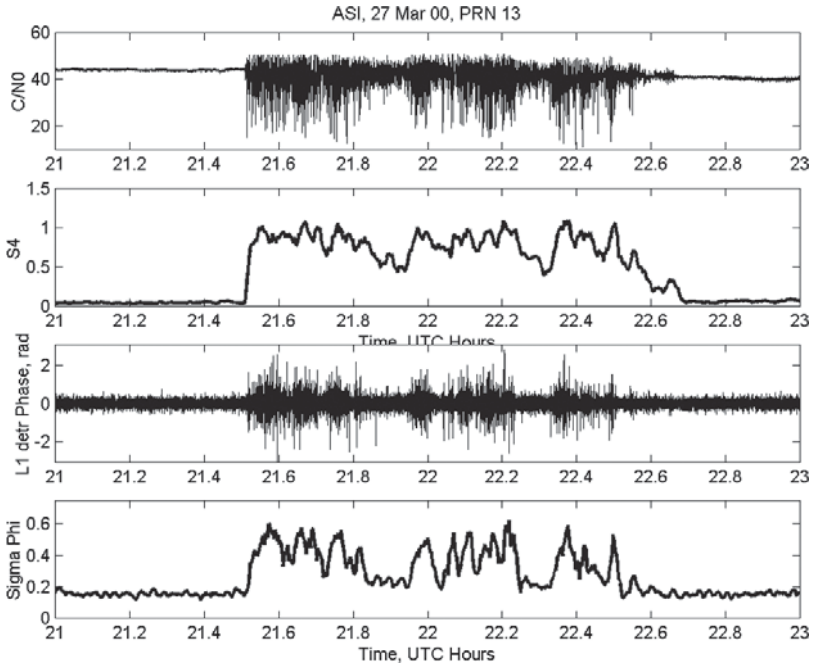


Figure 2. Scintillations recorded at Ascencion Island (Courtesy K. Groves, AFRL)

The scintillations start after sunset. They last in that case approximately 1 hour. The intensity and phase scintillations are correlated.

The Fig. 3 shows the percentage occurrence of amplitude scintillations in the equatorial zone. The fades amplitude is characterized by the scintillation index S4 which corresponds to the standard deviation of the intensity fluctuations. Its value is between 0 (no scintillation) and 1 (saturated).

As can be seen from the measurements, the scintillations activity increases with the solar activity (peak value in year 2000). There are typically two periods of the year of higher activity in spring and fall and as shown on the previous figure, the scintillations appear in the post sunset hours and last a few hours.

Polar regions

The polar region extends approximately from magnetic latitude 55° up to the pole. The amplitude scintillations are much lower than in the equatorial case. Typical peak values of the scintillation index are equal to 0.2.

Systematic studies of the radio scintillations can help to avoid problems in communication with satellites. Since transionospheric propagation errors due to ionospheric scintillations are a major source of errors in space based communication and navigation, the prediction of ionospheric scintillations is a promising way to reduce the impact of scintillations on operational systems.

Several studies have been already performed showing clear evidences of space weather – induced adverse effects on the Earth's ionosphere-plasmasphere system. Such effects can ultimately cause various types of problems such as range errors,

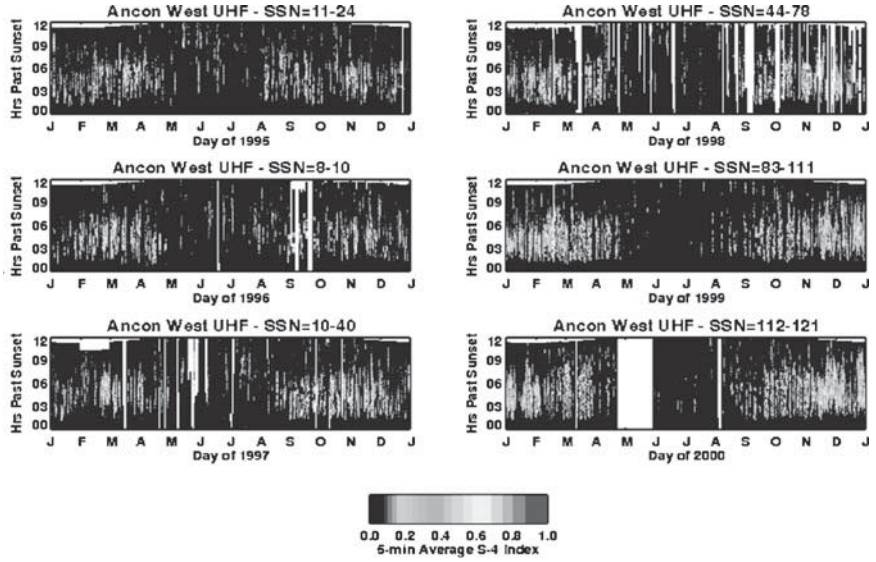


Figure 3. Scintillations dependency on the season and on the solar activity (Courtesy K. Groves, AFRL)

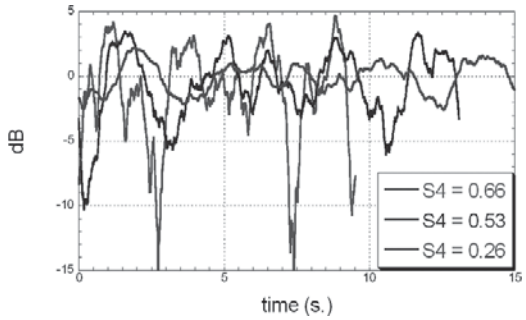


Figure 4. Time series obtained with the GISM scintillation model

rapid phase and amplitude fluctuations (radio scintillations) of satellite signals, leading to pronounced signal degradation and correspondingly to a degradation of the system performance.

Models are needed in particular to forecast the behavior of modeled ionospheric parameters some hours ahead to ensure high reliability of the Galileo system.

One of these models, the GISM model (ITU model) that has been developed at IEEA (Béniguel 2002), allows obtaining the different scintillation parameters, including the generation of time series for Test Cases. Such an example of a time series is presented on Figure 4 for the fades amplitude. The slope of the intensity variations can be calculated from these values. The ability of the system to cope with such fluctuations is dependent on this slope variation.

As GISM simulations show, the peak to peak dynamics of signal strength may be greater than 20 dB in a strong fluctuations case.

DATA ANALYSIS

The scintillations are mainly characterized by the S4 and sigma phi parameters which are the standard deviations of the intensity and phase of the signals received. Additional parameters such as the fades amplitudes and durations and their related probabilities are also of interest for system studies. The extent of the scintillation region is of particular interest regarding the probability that a user link can be affected by scintillations. The knowledge of this value is also one objective of the scintillations measurements campaigns.

All these parameters highly depend on the space weather geophysical characteristics: the solar spot number, the magnetic activity (specialy at high latitudes), the latitude, the season and the local time.

The results presented below have been deduced from measurements of GPS signals recorded in Douala, Cameroon and in São Jose dos Campos, Brazil, in the low latitude region. The data in Douala were recorded in 2004 using a GPS

scintillation receiver installed by ESA/ESTEC (Van Dierendonck and Arbesser-Rastburg 2004; Adam et al.). The data obtained in São Jose dos Campos were recorded in 2002 by INPE, Brazil. The solar flux numbers were respectively equal to 100 in Douala (2004) and 190 in Brazil (2002) in relation with the solar cycle. Both receivers record the data at a 50 Hz sampling frequency. In Douala both intensity and phase of signal received have been analyzed. In São Jose dos Campos, only the intensity has been analyzed.

From 50 Hz raw data, this receiver computes the amplitude and phase scintillation indices S4 and σ_ϕ, which are the standard deviations of the intensity and phase of the received signals. In addition to the GPS satellites, the receiver installed in Douala is able to track the SBAS signal of the EGNOS geostationary satellite PRN 131. It extracts scintillation parameters from this GPS like signal. This is of particular interest due to the fact that only the ionosphere motion modifies the received signal. This satellite is seen with a relatively low elevation angle and the power received is reduced as compared to the GPS satellites.

SPATIAL EXTENT

Since the receiver provides information about the satellite position (azimuth and elevation angles), an analysis of the spatial behaviour of the scintillation activity is possible. Figures 5 and 6 apply to data recorded in Douala. We have considered an arbitrary day: 2004-05-11, between 19 and 20. Figure 5 presents the tracks of the visible satellites. Figure 6 shows the corresponding S4 values recorded.

For this period, the GEO satellite link recorded a scintillation activity (Fig. 6a). The GPS satellites PRN 28 and PRN 7 coming near the angular position of the GEO satellite also experienced high scintillation activity. In addition, PRN 29 and PRN 26 were at intermediate distance from PRN 131 and were also affected by moderate scintillation activity. On the contrary, PRN 24, PRN 10 and PRN 5 weren't affected by scintillation.

The duration of scintillations is typically of the order of 1 hour. Calculations of correlation distances have been performed separately and are presented on Fig. 6b.

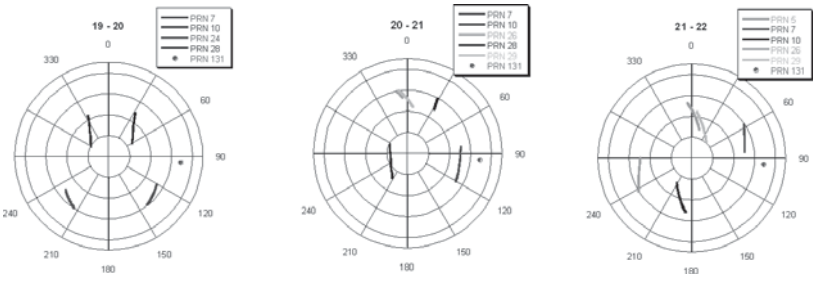


Figure 5. Visible GPS satellites on 2004-05-11. Elevation vs. azimuth is represented for 3 hours between 19 and 22. All Satellites over 30° were taken into account

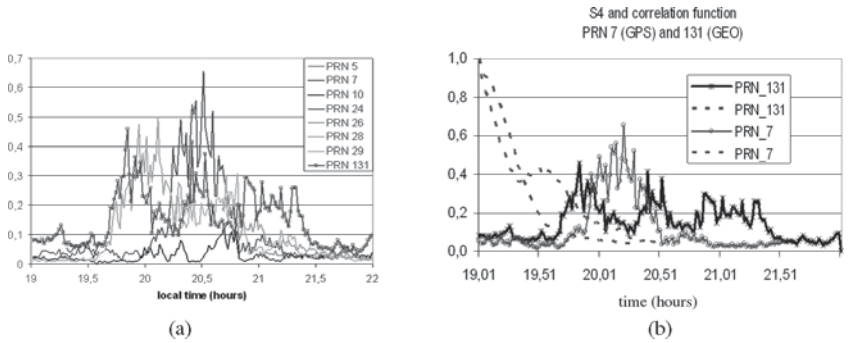


Figure 6. S4 recorded for each satellite visible on 2004-05-11 between 19 and 22 (a) and correlation function (b)

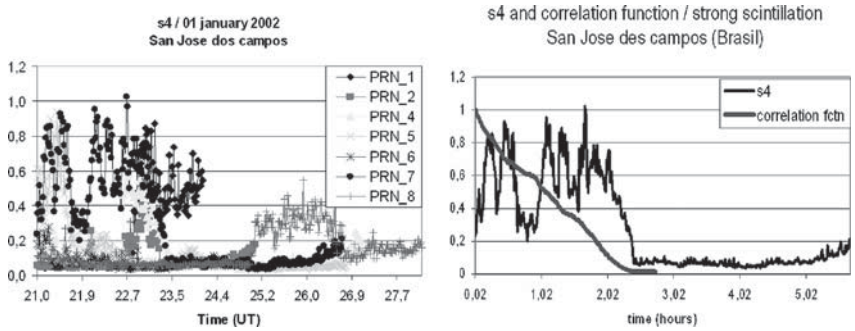


Figure 7. S4 recorded for each satellite visible on 2002-01-01 and correlation function

The correlation time which can be deduced is approximately 15 minutes (about 4° of earth rotation). At the altitude of the F-layer, set to 300 km for this example, it would correspond to an inhomogeneities region of about 450 km of diameter. For an observation point on the earth surface the separation angle is 72° for the worst geometrical location.

The same calculations were made from data at São Jose dos Campos (flux number 190). The results obtained are presented on Figure 7. The average extent is about 40 minutes in that case.

PROBABILITY OF SIMULTANEOUS FADING

Figure 8 presents the probability of simultaneous fading, given the value of S4 and given the number of satellites affected. This probability drops quickly with the number of satellites affected and increases with the flux number. All satellites were used for this calculation.

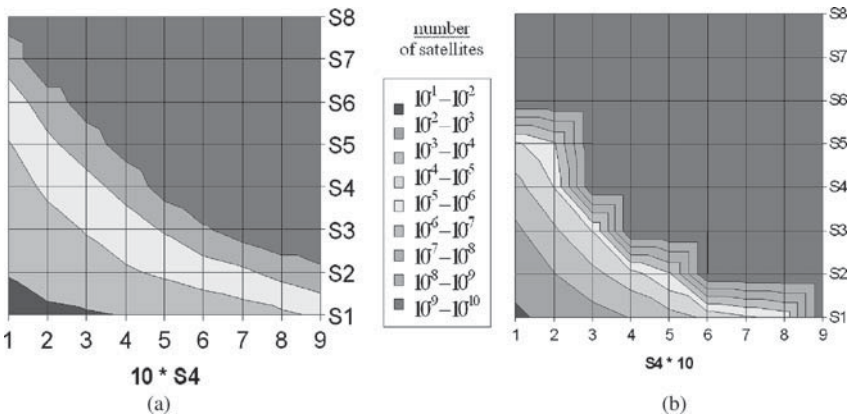


Figure 8. Probability of simultaneous fading (log scale). The left plot corresponds to the data recorded in Douala with a flux number equal to 100. The right plot corresponds to the data recorded in São Jose dos Campos with a flux number equal to 190. The vertical scale is the number of satellites simultaneously affected from 1 to 8

LOCAL TIME DEPENDENCY

One of the advantages of the geostationary satellite observed from Douala is its fixed geometry. This allows the analysis of the temporal variation in the scintillation behavior. Figure 9 presents the amplitude scintillation recorded during a whole day. As expected from the results of other measurements campaigns, scintillation occurs only during nighttime, essentially during the post sunset period between 19 and 24 LT.

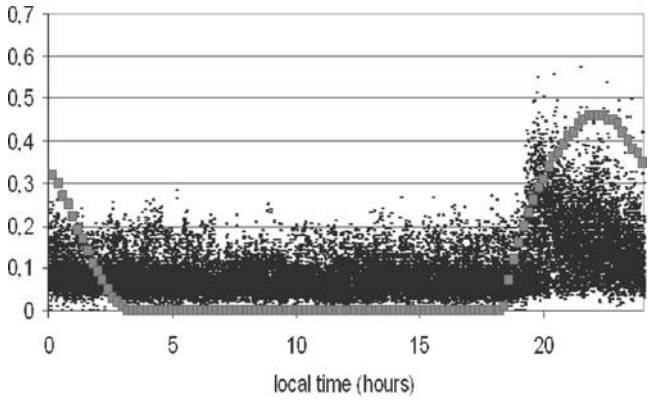


Figure 9. S4 vs. local time for the GEO satellite during the first 20 days of measurements compared with GISM prediction (red squares)

SEASONAL DEPENDENCY

An analysis of the long-term temporal variation of scintillation has been done with the data recorded in Douala. The mean value of S4 over the nighttime period is chosen as an indicator of the scintillation activity for a particular day. Figure 10 presents the evolution of this parameter over the 150 days of observation. It appears that after June the scintillation activity seems to be quieter. The same observation was made in South America.

Figures 11 presents the highest values obtained each day for the worst link for the GPS and the GEO satellites. In the case of the GEO, due to the fact that the C/N value is lower (cf Fig. 16), loss of locks occur for lower values of S4 than for the GPS. This is the reason why the values recorded for the GEO are lower than for the GPS links. In all cases, S4 has been calculated on 1 mn samples. The mean and max values over all GPS links are presented on Fig. 10 and Fig. 11 for the amplitude (S4) and on Fig. 12 for the phase (sigma phi). The peak values for sigma phi in radians are in the same range than S4. The phase was not measured for the GEO.

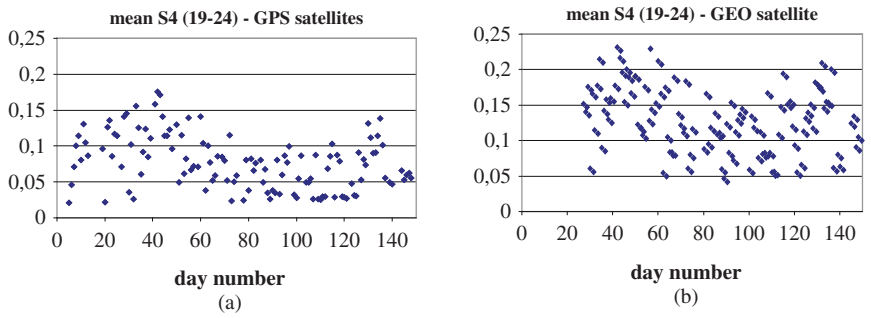


Figure 10. Mean value of S4 calculated from the S4 values of all GPS satellites with elevation angles greater than 30° over the nighttime period (a) and for the GEO satellite (b)

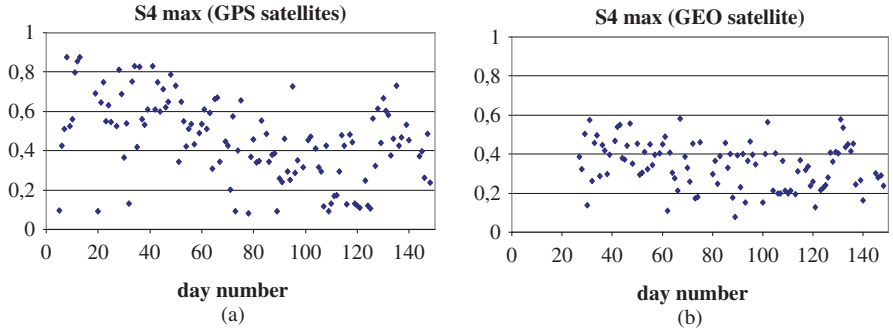


Figure 11. Highest values recorded one day for S4 on all the links of the GPS constellation for the nighttime period (a) and for the GEO satellite (b)

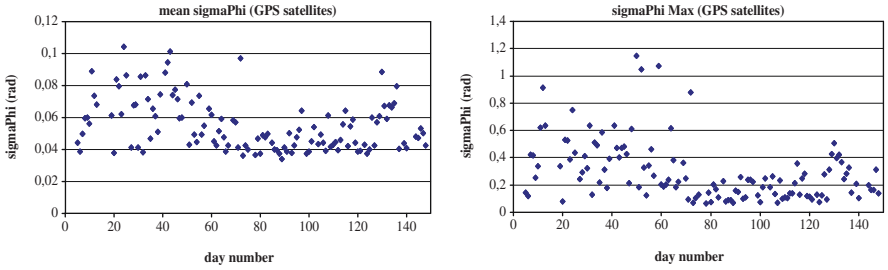


Figure 12. Mean and highest values recorded one day for σ_ϕ on all the links of the GPS constellation

LOSS OF LOCK

The receiver provides the lock time. This value indicates how long the receiver has been locked to the carrier phase of the GPS signal. This also indicates the time of the last loss of lock and it can be used to detect this failure.

The GEO link uses a GPS like signal with an L1 carrier. That is the reason why we have only considered the loss of lock of L1. Figure 13 presents the value of S4 before the loss of lock. It is possible to estimate the probability of having a loss of lock and a given value of S4. In addition, the frequency of occurrence of S4 may also be evaluated from the samples. Therefore we can calculate the probability of loss of lock vs. the value of S4. This result is presented on Fig. 14.

As a comparison, the Fig. 15 presents the values obtained with GISM for a typical receiver. The same behaviour is exhibited. The differences of levels are related to different values of the receiver parameters and to the fact that curves presented on Fig. 13 have plotted independently of the value of C/N and correspond consequently to an average value of this ratio. The GISM model allows setting any value to these parameters.

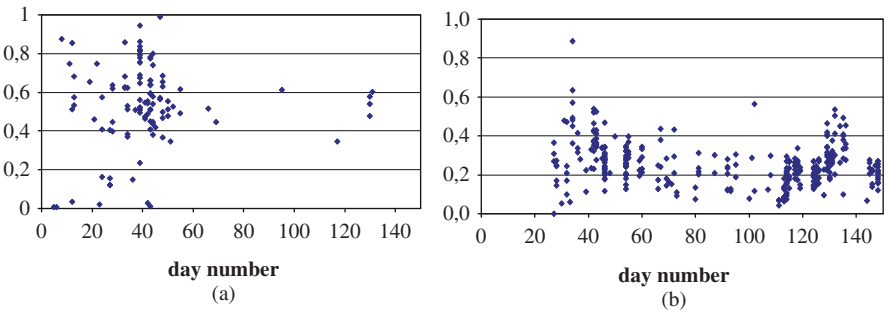


Figure 13. S4 before loss of lock of L1 carrier for all GPS satellites over 30° (a) and for the GEO (b). For each day, there may have several losses of lock. Only nighttime (19–24 LT) losses of lock were considered. The difference of S4 levels for GPS and GEO satellite is related to the signal level, which is significantly lower for the GEO

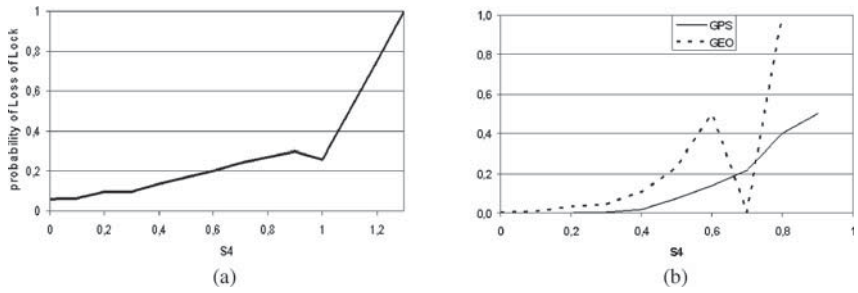


Figure 14. Probability of loss of lock, given the value of S4 in São Jose dos Campos (a) and Douala (b). For the GEO, for S4 greater than 0.6, there are not enough loss of lock occurrences to get correct statistics. This explains the discontinuity in the curve

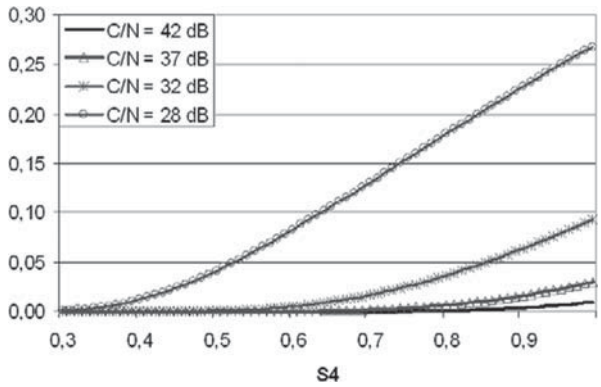


Figure 15. Probability of Loss of Lock for a typical receiver obtained with GISM scintillation model

In the GISM model, loss of lock is evaluated through the standard thermal noise tracking error for the PLL (Conker et al. 2003):

$$\sigma_{\Phi\mathrm{T}}^2 = \frac{\mathrm{B}_n}{(c/n_0)Is}\left[1 + \frac{1}{2\eta(c/n_0)Is}\right]$$

where Bn is the receiver bandwidth, and η is the predetection time. For airborne GPS receiver, B$n = 10$ Hz and $\eta = 10$ ms. *Is* is the scintillation intensity. Its mean value is 1 and it has a Nakagami distribution characterized by S4.

This relation expresses the thermal noise as a decreasing function of the scintillation intensity. As a result, if $\sigma_{\Phi\mathrm{T}}$ is above the 15° threshold then *Is* is below a value computed using this relation. As *Is* distribution is known for a given S4, the probability of occurrence of "*Is* < threshold" can be evaluated. The result is the probability of Loss of Lock. Figure 14 presents this probability versus S4 at given values of the C/N0. It can be noticed that links with high C/N0 are quite robust. On the contrary, links with low values of C/N0 are likely to be lost.

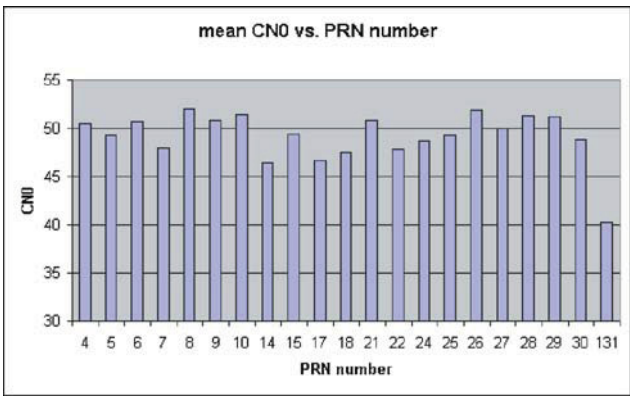


Figure 16. Mean C/N0 for each satellite PRN number. PRN 131 corresponds to the GEO satellite. The GPS satellites taken into account are all seen with an elevation angle greater than 30°. The elevation angle for the GEO satellite is 28°

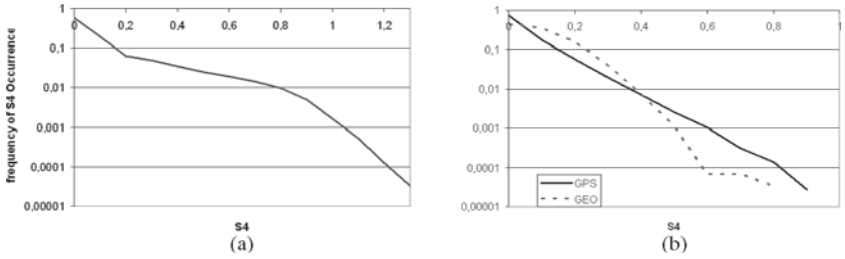


Figure 17. Frequency of occurrence of S4 in São Jose dos Campos (a) and Douala (b). Ten S4 intervals of equal width were considered. Each sample is counted in one of these intervals

There are more losses of lock on the GEO link than on the GPS links (Fig. 13). For the SBAS signal, the loss of lock appears at a lower value of S4. This is due to the lower signal power provided by the GEO satellite. The receiver also provides the C/N0 value of the link. As can be noticed on Fig. 16, the C/N0 value is 10 dB lower for the GEO satellite link.

The frequency of occurrence of S4 is presented on Fig. 17 for the two data sets. It exhibits a Log normal distribution. The red line (São Jose dos Campos) and dashed line (GEO satellite) have been plotted with a reduced data set on the contrary to the blue curve (GPS satellites in Douala).

POSITIONING ERROR

To analyze the effect of scintillation on the positioning error, we have simulated the receiver behavior affected by scintillation characterized by S4. According to

(Conker et al. 2003), in presence of scintillation, the tracking variance for a DLL (in C/A code chips squared) may be expressed as:

$$\sigma_\tau^2 = \frac{\mathrm{B}_n \mathrm{d}}{2(c/n_0)(1-\mathrm{S}_4^2)}\left[1+\frac{1}{\eta(c/n_0)(1-2\mathrm{S}_4^2)}\right]$$

Where Bn is the one-sided noise bandwidth (typical value is 0.1 Hz) and d is the correlator spacing in C/A code chips (typical value is 1 to 0.1). η is the predetection time. The chip length is about 293 m.

To evaluate the positioning error, the following steps were performed for each tracked satellite:

- S4 is measured.
- σ_τ is deduced from S4.
- assuming a gaussian distribution characterized by σ_τ, a range error is computed.
- a Yuma file is used to evaluate the satellite position in order to fill the navigation equations.

The navigation equations are solved with these range errors to compute a positioning error.

Figure 18 presents the results of this simulation. In that example, the scintillation effects aren't significant. The mean value of S4 shows that the scintillation activity was weak. As a result, the number of tracked satellites was always high enough to mitigate the range error.

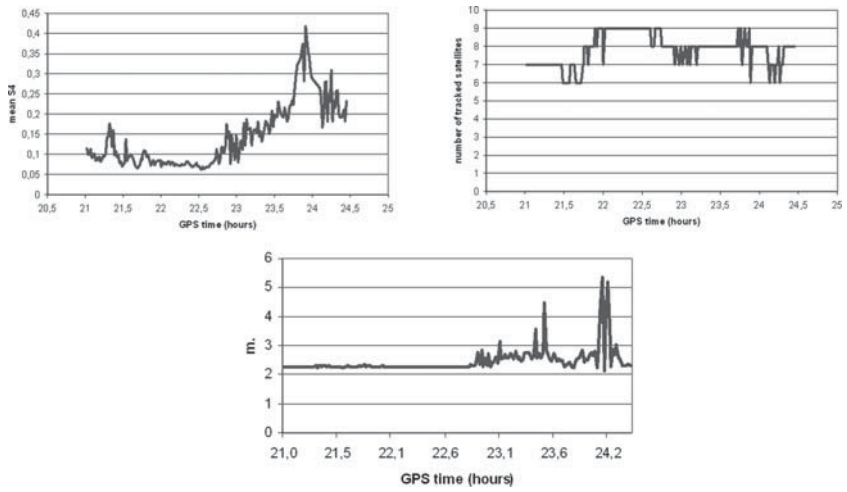


Figure 18. The first curve represents the mean value of S4 (all visible GPS satellites), it shows the scintillation activity. The second presents the number of tracked satellites. The last one corresponds to the evaluated positioning error

CONCLUSION

The characteristics of signal scintillations, due to the propagation through ionosphere inhomogeneities, have been presented. This as been illustrated by measurement results of two measurements campaigns: in Douala, Cameroon starting beginning of year 2004 with an average solar flux number equal to 110 and in São Jose dos Campos, Brazil, during the first 2 weeks of January 2002 with a solar flux number equal to 190.

One interesting feature of the measurement campaign in Douala was the fact that we had an additional link with a GEO satellite. This link is not moving with respect to the ground station and is therefore only affected by the motion of the ionosphere. It is seen from the observation point with an elevation angle equal to 28° which is in principal large enough to get rid of multipath effects.

Results have been presented for the local time dependency, the probability of loss of lock, the positioning error due to scintillations, the S4 distribution, the spatial extent and the probability of the simultaneous loss of lock.

The measurements show the usual dependency to local time: the scintillations appear at post sunset hours and may last a few hours. Modeling with GISM model provides concurrently a reasonable agreement. The probability of loss of lock has been calculated, but for an average value of C/N. This point should be investigated in more detail in the future but a rough estimate has been obtained. The simultaneous probability of loss of lock, which highly depends on the solar flux number has also been estimated.

The positioning error has been calculated for a case of medium values of the scintillation ratio. The errors obtained are a few meters in that case. It increases very rapidly with S4.

We have shown that the S4 probability seems to converge towards a log normal distribution. Three curves have been plotted (GEO satellite, GPS low latitudes in Brazil and GPS low latitudes in Africa) and increasing the data base shows this tendency. Finally we have calculated the spatial extent of the inhomogeneities region, depending on the solar flux number. Values of hundreds of km have been estimated.

Another measurement campaign in the frame of the ESA/ESTEC PRIS project (Prediction of Ionospheric Scintillation) is currently running on a larger basis, with receivers all over the globe. This will allow to enlarge the data base which has been initiated.

REFERENCES

Adam, J-P., Béniguel, Y., De Paula, E., Arbesser-Rastburg, B.: "Analysis of scintillation data recorded at low latitudes" to appear in Radio Science, (2007)

Béniguel, Y.: "Global Ionospheric Propagation Model (GIM): a propagation model for scintillations of transmitted signals", Radio Science, **37**, (3), (May 2002)

Conker, R.S., El-Arini, M.B., Hegarty, C.J., Hsiao, T.: "Modeling the Effects of Ionospheric Scintillation on GPS/SBAS Availability", Radio Science, (January/February 2003)

Van Dierendonck, A.J., Arbesser-Rastburg, B.: "Measuring Ionospheric Scintillation In The Equatorial Region Over Africa, Including Measurements From Sbas Geostationary Satellite Signals", Beacon Symposium, Trieste, Italy (October 2004)

CHAPTER 4.0

RADIATION ENVIRONMENT OF THE EARTH–SPACECRAFT AND AIRCRAFT ENVIRONMENT

IOANNIS A. DAGLIS

Institute for Space Applications and Remote Sensing, National Observatory of Athens
Metaxa & Vas. Pavlou Str., Penteli, 15236 Athens, Greece. Tel +30-210-8109185
Fax +30-210-6138343. http://www.space.noa.gr. daglis@space.noa.gr

INTRODUCTION

The Sun–Earth space environment and in particular the geospace environment is a complex multi-component multi-scale physical system and a challenging object of scientific desire. Moreover, it is also a system of crucial practical importance. The effects of disturbances in the geospace environment on human beings in space, on satellite operations, on the electrical power grid, upon communication links, and probably even on climate make understanding of Sun–Earth coupling vital. Space weather is crucial from a pragmatic point of view, just as space plasma physics is important from a basic science point of view.

The papers of this chapter cover three central topics of space weather: modeling and reproduction of radiation belt dynamics; radiation effects on spacecraft and suitable countermeasures; aircrew radiation exposure in aviation altitudes.

In the first paper Sebastien Bourdarie and co-workers present and discuss several approaches to the problem of accurate and reliable description of the radiation belts at all points in space and under all conditions. The energetic electron and proton radiation environment is one of the most important regions of geospace, exhibiting very complex dynamical behavior that defines its weather-like features. This population is a most important part of the chain that interconnects the Sun and interplanetary space with the terrestrial magnetosphere, ionosphere, and atmosphere.

While the average conditions of the geospace radiation environment are fairly well understood and described, the dynamics during magnetic storms is relatively unexplored. Radiation belt dynamics can be addressed theoretically, empirically, and through model/data assimilation. Bourdarie et al. discuss the virtues and deficiencies

J. Lilensten (ed.), Space Weather, 217–218.

of the theoretical and the empirical approach, and suggest that a way to overcome the intrinsic limitations of these approaches is to use data assimilation techniques, ranging from simple (such as direct insertion) to extremely complex (such as Kalman filters).

In the second paper, Wolfgang Keil reviews the procedures used to evaluate spacecraft behaviour in orbit due to particle radiation effects. Keil discusses predicted and manifested effects for selected cases, and presents the possible countermeasures that can be implemented against space radiation effects. There are a number of countermeasures that are implemented by design, while countermeasures by operational measures are still in development.

Finally, Peter Beck presents in the extensive third paper an overview of measurements by TEPC (tissue equivalent proportional counter) dosimeters during quiet solar conditions and focuses on dose results using the EURADOS In-Flight Radiation Data Base. The third paper furthermore describes TEPC measurement campaigns during extreme solar conditions, which are not part of the EURODOS report, as well as radiation monitoring regulations and procedures.

The actual session of the second European Space Weather Week on the Radiation Environment of the Earth was complemented by a further three papers ("Space radiation effects on aircraft" – Bryn Jones, "Towards operational use of radiation belt modeling" – Daniel Heynderickx, "Near- and mid-term needs for a radiation belt modeling upgrade" – Richard Horne and Sebastien Bourdarie), which however could not be delivered for this volume, because of other heavy commitments of the authors.

CHAPTER 4.1

COMPLEMENTARITY OF MEASUREMENTS AND MODELS IN REPRODUCING EARTH'S RADIATION BELT DYNAMICS

S. BOURDARIE[1], V. MAGET[1], R. FRIEDEL[2], D. BOSCHER[1], A. SICARD[3] AND D. LAZARO[1]

[1] *ONERA/DESP*
[2] *LANL*
[3] *CNES*

Abstract: The harsh radiation environment in the inner magnetosphere up to geosynchronous orbit is of major concern to an ever increasing amount of space hardware. While the average or quiescent conditions of the energetic particle population are fairly well characterized, the dynamics during magnetic storms are severely under-sampled. The description of the energetic electron and proton radiation environment at all points in space, which can provide reliable environmental data for locations of satellites that do not carry any energetic particle instrumentation is not trivial. The following approaches, theoretical, empirical, and model/data assimilation are examined in this paper

INTRODUCTION

While the detailed global knowledge of the dynamic behavior of energetic particles is of obvious importance to spacecraft and or humans operating in that environment, the understanding of the inner magnetospheric particle population dynamics is also an important scientific question. Many recent studies of geomagnetic storms [Baker et al., 1997; Reeves et al., 1998a, b; Li et al., 1998; Baker et al., 1998a, b, c; Friedel et al., 1999; Li et al., 1999; Kanekal et al., 2000, Reeves et al., 2003, Green and Kivelson, 2004] have investigated the dynamical behavior of relativistic electrons in the Earth's magnetosphere.

Observations at geosynchronous orbit have revealed that relativistic electron flux increases occur several days after the onset of a magnetic storm, and that there is a poor correlation between the size and/or duration of the increase with other storm-time indicators such as Dst, Kp. On the other hand, correlation between solar wind speed have been well established since the 1960s [Paulikas and Blake, 1976].

J. Lilensten (ed.), Space Weather, 219–229.

In addition, the orientation of the IMF may also be an important factor [Paulikas and Blake, 1976; Blake et al., 1997].

Theoretical models have proposed mechanisms such as radial diffusion, wave-particle interaction and impulsive shock acceleration as underlying processes. Some of the more detailed suggestions include ULF wave heating [Liu et al., 1999], simple diffusion and heating by conservation of the first adiabatic invariant [Hilmer et al., 2000], acceleration by substorms [Ingraham et al., 2000] and energy diffusion by whistler mode chorus waves [Horne and Thorne, 1998; Summers et al., 2002; Horne et al., 2003, 2005, Meredith et al., 2002a, 2002b, 2003]. Currently, the situation remains unclear as to whether a given mechanism or several processes operate simultaneously and as to the nature of the relationship of these on external factors.

Observational studies use data from single or multiple points in space and are unable to rigorously distinguish between possible acceleration processes. These studies are limited due to poor coverage of measurements [Friedel et al., 1999]. To have a good representation of what really happens during magnetically active periods, several spacecraft have to be at the "right location" at the "right time" which is rarely the case.

A global model of relativistic electron and proton acceleration is needed in order to increase measurement coverage. With the use of physical models of the radiation belts it is possible to "physically interpolate" between measurements (particularly in L) as well as extrapolate taking into account the limited channel number and energy range of instruments. Combining both measurements and theoretical model is a good way to increase the spatial and temporal resolution of measurements to produce time-dependent maps of the radiation belts, both for now-casting capabilities and as a tool for further research into mechanisms and causes.

In the following sections pros and cons for each approach used (theoretical, empirical, and model/data assimilation) to describe the radiation belt dynamics will be presented.

THEORETICAL APPROACH

One option used to describe the Earth's radiation belt dynamics is with diffusion theory. Here a complex Boltzmann equation is used, where the electric and magnetic fields are supposed to be known everywhere in the magnetosphere. Then the Boltzmann equation is written in the action-angle phase space:

$$\frac{\partial f}{\partial t} + \frac{\mathrm{d}\vec{J}}{\mathrm{d}t}\frac{\partial f}{\partial \vec{J}} + \frac{\mathrm{d}\vec{\varphi}}{\mathrm{d}t}\frac{\partial f}{\partial \vec{\varphi}} = \left(\frac{\partial f}{\partial t}\right)_{colision}$$

where f is the distribution function, $\vec{J}$ the action variables and $\vec{\varphi}$ the associated canonical angle variables (the phases). In such models all physical processes are expressed and can be monitored by one or more input variables. For example, field fluctuations can be driven using a magnetospheric index. Taking into account

particle interaction and a complex Boltzmann equation allows the description of hot plasma in the radiation belts. Formulations of the diffusion theory for the radiation belts start in the late sixties [Fälthammar, 1965].

Salammbô, developed at ONERA-DESP starting in 1990 [Beutier and Boscher, 1995, Beutier et al., 1995, Bourdarie et al., 1997, Boscher et al., 1998], is the most advanced diffusion code in this field. Other similar codes do exist [Brautigam and Albert, 2000, Miyoshi et al., 2003, Li et al., 2001, Shprits et al., 2005]. Two particle populations are taken into account separately, i.e. protons and electrons with respective energy range 10 keV–300 MeV and 10 keV–10 MeV. Schematic views of the Salammbô electron and proton model are given in Fig. 1.

One of the strengths of the Salammbô code are the limited input data needed (is shown in the light grey ellipses in Fig. 1). Data for particles fluxes at an outer boundary or a pre-defined boundary condition, Kp and Dst – this is sufficient to reproduce all the important dynamic features observed in the real magnetosphere at the storm time scale. Kp and Dst in the model are a proxy for several processes – Kp scales the diffusion coefficients for radial diffusion by magnetic and electric perturbation fields, the magnetopause position as a loss mechanism, the location of the plasmapause for wave activity and the magnitude of the wave activity itself, Dst scales the drift shell position for day side losses, Kp and Dst scales the magnetic cut-off. To reproduce slow solar cycle variations the Ap and $F_{10.7}$ indexes are used to scale atmospheric density modulations, which directly impact proton losses. The currently used boundary fluxes in the electron model derive from LANL geosynchronous observations, and in the proton model (for Solar Energetic Particle events) from NOAA-GOES observations. This limits the valid range of the model to inward of geosynchronous orbit.

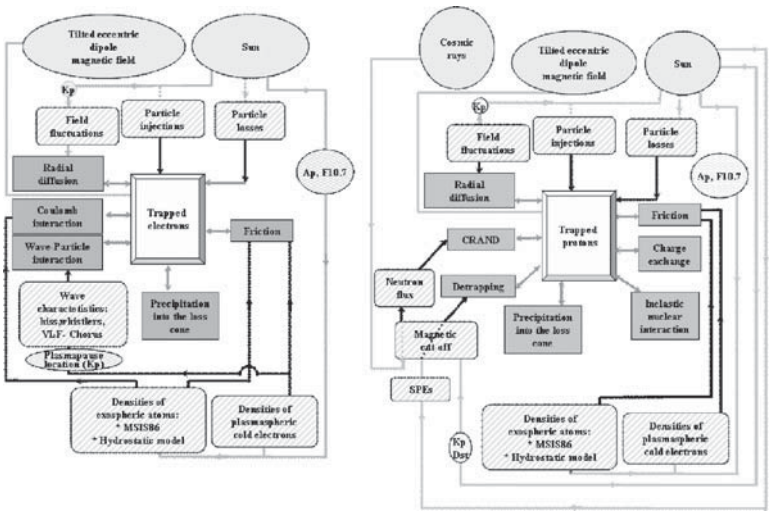


Figure 1. Flow chart of the Salammbô electron (left) and proton (right)

An initial run was performed for the period of the NSF GEM storm of September 1–October 10, 1998 [Bourdarie et al., 2005]. The model was run with the correct Kp values for this period. Here we performed our model run using ONLY the LANL GEO data as input. The results are shown in Fig. 2. The model is initialized with a default state at the beginning of the run representing an average quiet magnetosphere, taken from CRRES measurements. HEO data is used as a test: The top three panels show data in the L^* versus time format, where each color coded vertical bar in the plot represents the flux along the satellites cut through L^* at this time. The top panel shows the actual HEO-3 satellite data for the > 0.63 MeV channel. The next panel shows the model output along the orbit on HEO-3; both these panels share the same color bar. The third panel shows the ratio of model divided by data, and the color bar represents ratios up to 10 in yellow/red graduations and ratios down to 0.1 in blue/dark blue graduations. The bottom panel shows the Dst storm index for reference. Ideally, if model and data agree 100%, this ratio should be 1 (black). Here we see large deviations from 1 in two areas: the outer belts near GEO and the inner region near the slot. The first discrepancy in the outer belts can be explained in terms of the missing model physics as described earlier: whistler chorus interactions are not yet modeled, which are believed to be the agent of electron acceleration in this region. The HEO test-spacecraft has a highly elliptical orbit and cuts through the outer radiation belts at high latitude. The model data here is highly dependent on the assumed pitch angle distribution at the equator, which here comes from a statistical model. We know that energetic

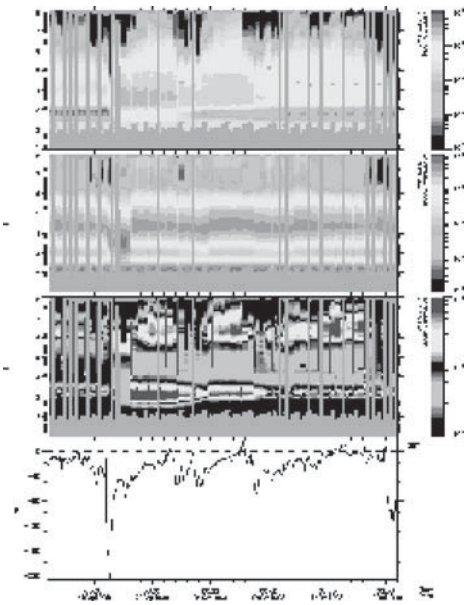


Figure 2. Data, model and comparison outputs − Model + LANL GEO, HEO as test. See text for details

electron pitch angle distributions can be highly variable during storms [Blake et al., 2001]. Obviously the statistical pitch angle representation is not adequate for this time period. The second discrepancy at low L is simply due to wrong initial state and short model run: diffusion is extremely slow in this region and we simply observe a persistence of the initial state.

The limitations of a pure theoretical approach have different origins. The first one is of course due to the physical processes that have been included (or omitted) in the model and the degree of details and the importance of assumptions in their mathematical description (e.g. impact of the quasi-linear theory applied to describe wave particle interaction). At least all major physical processes playing an important role on trapped particles should be included in models. The second one is due to the parameterization of these physical processes from storm to storm. The most common is to deduce empirical formulae from statistical studies, where physical processes are driven by one or more proxies, such as Kp, Dst, and solar wind parameters. This tends to introduce large uncertainties, especially for active periods, when statistics are always poor. Moreover, because physical models for radiation belts are non-linear, it is difficult to estimate the impact of any error due to the parameterization phase. The last limitation comes from the mapping between the invariant model space (which uses magnetic coordinates) to geographic ones. The problem comes from external magnetic field models, which fails to correctly describe the Earth's real field during disturbed periods, especially during the main phase of a storm.

EMPIRICAL APPROACH

Data from any single point measurement in space has traditionally been used to derive information about the local environment at that satellite. There have been some earlier attempts of obtaining a dynamic global state of the inner magnetospheric electron populations based on single spacecraft observations, based on the CRRES data, our community's last mission dedicated to the radiation belts. CRRES alone, on a 5.5-hour resolution, was able to provide a basically complete description of the inner region, across a wide energy range – due to it's ideally suited geosynchronous transfer orbit [Friedel and Korth, 1995]. However, CRRES flew in 1990/1991, and one has to look to other resources for such information today.

Friedel et al. [2000] used a multi-spacecraft synthesis using simple interpolation technique with data from up to 11 spacecraft to assemble a "map" of the inner radiation belt energetic electron population. This simple approach led to radiation belt maps that could represent the dynamics of the inner region on around a 3 hour time scale, but the simplistic interpolation and inter-calibration scheme employed led to many unrealistic local time and radial variations which were clearly not physical, but rather a reflection of insufficient instrument characterization and inter-calibration and insufficient spatial coverage. While measurements from scientific satellites can provide the full three-dimensional particle distribution function and

local magnetic field data (allowing data to be determined at constant adiabatic invariants, which are the coordinates that allow data to be inter-compared throughout the inner magneto-sphere), these data do not provide long time scales in either the past or future, and, when present, have a limited spatial coverage by themselves. Data from the programmatic missions provides excellent time coverage, longevity and spatial coverage, but with particle instrumentation that provides mainly omni-directional data and no magnetic field information.

Next, because different particle species exists in space their measurment is not straightforward. Most radiation monitors detect only energy deposition in materials which can come form any particle with sufficient energy (proton, electron, photon). The consequence is that all measurements have to be analyzed in detail according to our current knowledge of radiation belt global structure and dynamics. This analysis has to focus on instrument contamination, saturation, background, glitches …and cross-calibration [Friedel et al., 2005]. It is obvious that limitations of instruments have to be known for any use of in-situ data. An example of electron data contamination during solar particle event is given in Fig. 3.

The limitations of the empirical approach have different origins. The first one is of course the spatial and energy coverage which is quite poor. Next the instruments are never perfect and each of them have their own limitation, data must be analyzed in term of contamination, saturation, global coherence, etc. Finally, the last limitation is the same as in the theoretical approach discussed above, and comes from the

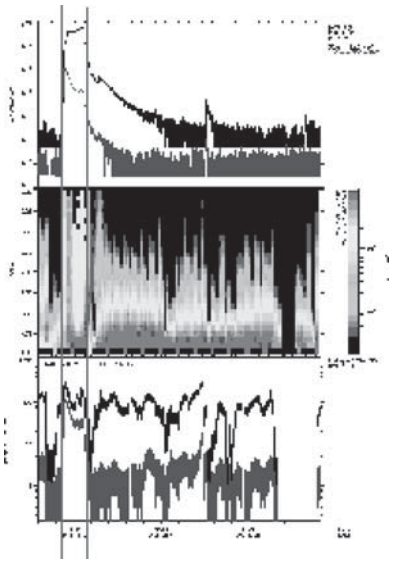


Figure 3. Example of data contamination during a solar flare (top panel is GOES proton data, middle panel is one electron channel on GPS and botton panel is electron data on-board LANL GEO). Contamination is clearly seen between the two vertical bars

mapping between the invariant space (or magnetic coordinates) to geographic ones, which is needed in order to compare or put together data from several sources.

DATA ASSIMILATION

Because the preceding two approaches are limited either by our current physical knowledge of the global magnetospheric system or by in-situ measurements quality or coverage, new techniques have to be developed. A promising way forward is through the use of data assimilation tools. The challenge is thus to utilize the available data in a framework that still allows us to retrieve high fidelity global maps of the radiation belts. Data assimilation techniques range from simple (such as direct insertion) to extremely complex (such as Kalman filters) with associated increase in required computing power.

The first approach, direct insertion is a synthesis between multiple point space measurements and a physics based radiation belt model that makes full use of all the data (electron or proton) available at a given time and uses the model to provide a physical interpolation between the data. The end result is a dynamic and global model of the energetic electron and proton radiation environment at all points in space, which can provide reliable environmental data for locations of satellites that do not carry any energetic particle instrumentation. A run using direct assimilation (LANL GEO and GPS data) with the Salammbô code which assimilates data into the

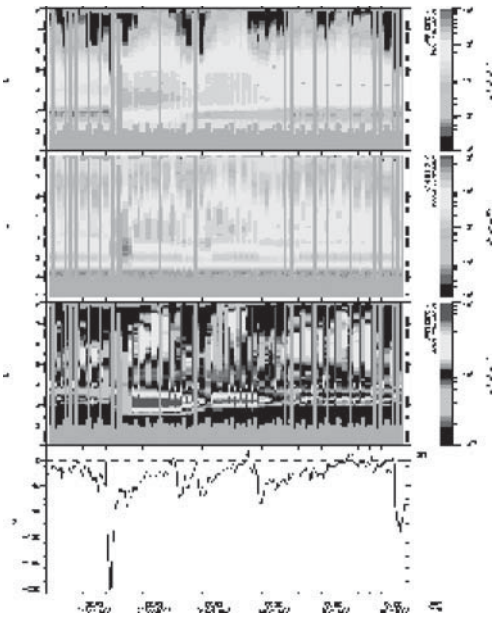


Figure 4. Data, model and comparison outputs. Model + LANL GEO + GPS, HEO as test. See text for details

region down to L = 4 is shown in Fig. 4. [Bourdarie et al., 2005]. As a quick visual comparison between Fig. 2 and Fig. 4 easily shows the model performance is much improved by the inclusion of just one additional satellite in the assimilation process. This is particularly true for the region which is now covered by data input – GPS data is available from L = 4 outward, and in that region the model/data comparison shown mainly black and light yellow indicating performance of model to within a factor of 2–3 of the data. Inclusion of GPS around L = 4 to 5 compensates for the missing physics in the region, while near geo it helps to properly define the pitch angle distribution which is needed to correctly estimate the fluxes at the high latitudes of HEO. The inner region remains badly represented here. It should be noted that here the fidelity of the output heavily depends on the fidelity of the data used. Friedel et al., 2005, estimate that their cross-calibration method provides data fidelity to within a factor of two between the spacecraft used.

A second approach consists in implementing a filtered data assimilation technique, the Kalman Filter (Kalman–Bucy, 1961) which can obtain a better estimate of the true state of the radiation belts than the one obtained from direct assimilation. The Kalman filter estimates the most likely electron or proton phase space density (psd) of the current state of the radiation belts at each defined grid points given the model state and all current and past observations. The Kalman filter is a sequential method that performs a maximum likelihood estimation between weighted current observations and weighted predicted state of the radiation belt at time t. A sequence is composed of two steps: the forecast and update steps.

During forecast step, the Kalman filter takes into account the fact that the initial state and the model are not perfect. The advantage here, compared to direct assimilation, is represented by the propagation forward in time of a matrix containing variance and covariance of errors at each grid point. Thus, this matrix is a $N \times N$ matrix, with N the number of grid points used in Salammbô 3D code (more than 15000).

When a new set of observations becomes available, an update is performed. This step merges the model prediction and current observations, by giving each appropriate weights, to provide an "assimilated" state. These weights are obtained through minimising the trace of the error-covariance matrix based on a probabilistic least square method. Thus, the psd uncertainty is globally sharpened around an optimal estimation of the state of the radiation belts. The benefit of the update consists in the use of the error-covariance matrix to extract the maximum amount of information from the observations and to spread it out over the Salammbô grid. This technique is conceptually appealing but practically unusable for large systems like the radiation belts with more than 15000 states: computation time would be too heavy. Consequently, Monte Carlo techniques can typically be applied in order to sample and thus approximate the error-covariance matrix. One such method based on the Kalman Filter has been developed by G. Evensen since 1994 [Evensen, 1994] for Oceanography and is known as the Ensemble Kalman Filter. We have implemented this technique for the radiation belts using the Salammbô code. The basic idea is to construct an ensemble of m initial states such that the mean of the ensemble is the initial state of the Kalman Filter and whose dispersion is an

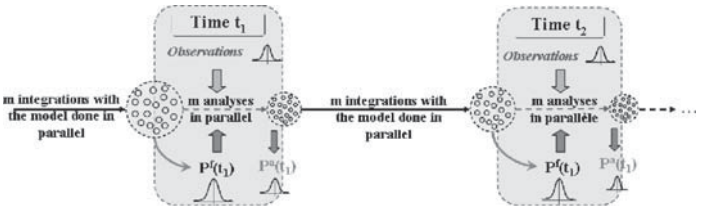


Figure 5. Sketch of the ensemble Kalman Filter

approximation of the initial error-covariance matrix. Consequently, the computational cost is reduced to the propagation and the update of an ensemble matrix A containing m states of N psd grid points, where m is of the order of 100 only!

The time evolution of the covariance approximation is done by evolving all the individual ensemble members using the Salammbô 3D model. Instead of diffusing the uncertainties contained in the error-covariance matrix for the Kalman Filter, the ensemble states are spread into the state space. Only during the update step, the sample covariance of the ensemble A is calculated in order to correct the model forecast using the new set of observations. A sketch of the ensemble Kalman Filter is provided in Fig. 5.

CONCLUSION

Reproducing the time variations of electron and proton radiation belts is not an easy task. Theoretical and empirical approaches have their own intrinsic limitations which might lead to many unrealistic local time and radial variations. Clearly important progress can be made by combining both approaches. The development of data assimilation techniques is very popular nowadays and takes advantages of both models and in-situ data. To this end physical models have to be more and more detailed and radiation belt measurements will have to be done at locations where errors in models are the highest.

REFERENCES

Baker, D.N., et al.: Recurrent geomagnetic storms and relativistic electron enhancements in the outer magnetosphere: ISTP coordinated measurements, J. Geophys. Res. **102**(14), 141–14,148 (1997)

Baker, D.N., Li, X., Blake, J., Kanekal, S.,: Strong electron acceleration in the earths magnetosphere, Adv. Space Res. **21**, 609–613 (1998a)

Baker, D.N., et al.: A strong CME-related magnetic cloud interaction with the earths magnetosphere: ISTP observations of rapid relativistic electron acceleration on may 15, 1997,Geophys. Res. Lett. **25**, 2975–2978 (1998b)

Baker, D.N., et al.: Coronal mass ejections, magnetic clouds, and relativistic magnetospheric electron events : ISTP, J. Geophys. Res. **103** (17), 279–17,291 (1998c)

Blake, J.B., Baker, D.N., Turner, N., Ogilvie, K.W., Lepping, R.P.: Correlation of changes in the outer-zone relativistic-electron population with upstream solar wind and magnetic feld measurements, Geophys. Res. Let. 24, 927–929 (1997)

Blake, J.B., Selesnick, R.S., Baker, D.N., Kanekal, S.: Studies of relativistic electron injections events in 1997 and 1998, J. Geophys. Res. **106** (19) 157–19,168 (2001)

Beutier, T., Boscher, D.: A three-dimensional analysis of the electron radiation belt by the salammbô code, J. Geophys. Res. **100** (14),853–14,861 (1995)

Beutier, T., Boscher, D., France, M.: A three-dimensional analysis of the proton radiation belt by the salammbô code, J. Geophys. Res. **100** (17),181–17,188 (1995)

Boscher, D., Bourdarie, S., Friedel, R., Korth, A.: Long term dynamic model for low energy protons, Geophys. Res. Lett. 25, 4129–4132 (1998)

Bourdarie, S., Boscher, D., Beutier, T., Sauvaud, J.A., Blanc, M.: Magnetic storm modeling in the earths electron belt by the salammbo code, J. Geophys. Res. **101**, (27),171–27,176 (1996)

Bourdarie, S., Boscher, D., Beutier, T., Sauvaud, J.-A., Blanc, M.: Electron and proton radiation belt dynamic simulations during storm periods: A new asymmetric convection-diffusion model, J. Geophys. Res., 102 (17),541–17,552 (1997)

Bourdarie S., Friedel, R.H.W., Fennell, J., Kanekal, S., Cayton, T.E.: Radiation belt representation of the energetic electron environment: Model and data synthesis using the Salammbô radiation belt transport code and Los Alamos geosynchronous and GPS energetic particle data, Space Weather, 3, S04S01, doi:10.1029/2004SW000065 (2005)

Brautigam, D. H., Albert, J.M.: Radial diffusion analysis of outer radiation belt electrons during the October 9, 1990, magnetic storm, J. Geophys. Res. **105** (291), (2000)

Evensen, G., Sequential data assimilation with nonlinear quasi-geostropic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, NO C5, pp 10143, May 1994, (1994)

Falthammar, C.G.: Effects of time-dependent electric fields on geomagnetically trapped radiation, J. Geophys. Res., Space Physics, 70, 2503 (1965)

Friedel, R.H.W., Korth, A.: Long-term observations of keV ion and electron variability in the outer radiation belt from CRRES, Geophys. Res. Lett. 22, 1853–1856 (1995)

Friedel, R.H.W., et al.: A multi-spacecraft synthesis of relativistic electrons in the inner magnetosphere using LANL, GOES, GPS, SAMPEX, HEO and POLAR, Rad. Meas. Jour. 18 ,589–597 (1999)

Friedel, R.H.W., et al.: A multi-spacecraft synthesis of relativistic electrons in the inner magnetosphere using LANL, GOES, GPS, SAMPEX, HEO and POLAR, Adv. in Space Res. 26, 93–98 (2000)

Friedel, R.H.W., Bourdarie, S., Cayton, T.E.: Intercalibration of magnetospheric energetic electron data, Space Weather, 3, S09B04, doi:10.1029/2005SW000153 (2005)

Green, J.C., Kivelson, M.G.: Relativistic electrons in the outer radiation belt: Differentiating between acceleration mechanisms, J. Geophys. Res. 109, A03213, doi:10.1029/2003JA010153 (2004)

Hilmer, R.V., Ginet, G.P., Cayton, T.E.: Enhancement of equatorial energetic electron fluxes near $l = 4.2$ as a result of high speed solar wind steams, J. Geophys. Res. **105** (23),311–23,322 (2000)

Horne, R.B., Thorne, R.M.: Potential waves for relativistic electron scattering and stochastic acceleration during magnetic storms, Geophys. Res. Lett. 25, 3011 (1998)

Horne, R.B., Glauert, S.A., Thorne, R.M.,: Resonant diffusion of radiation belt electrons by whistler-mode chorus, Geophys. Res. Lett. **30**(9), 1493, doi:10.1029/2003GL016963 (2003)

Horne, R.B., Thorne, R.M., Glauert, S.A., Albert, J.M., Meredith, N.P., Anderson, R.R.,: Timescale for radiation belt electron acceleration by whistler mode chorus waves, J. Geophys. Research, 110, A03225 doi:10.1029/2004JA010811 (2005)

Ingraham, J.C., Belian, R.D., Cayton, T.E., Christensen, R., Friedel, R.H.W., Meier, M.M., Reeves, G.D., Yuszewski, M.: Substrom injection of relativistic electrons to geosynchronous orbit during magnetic storms: A comparison of the march,24,1999 and march 10, 1998 storms, Poster, GEM workshop, Snowmass, 19–23 June 2000, (2000)

Kalman, R. E., Bucy R. S.: "New Results in Linear Filtering and Prediction Theory", Transactions of the ASME – Journal of Basic Engineering Vol. 83: 95–107 (1961)

Kanekal, S.G., Baker, D.N., Blake, J.B., Klecker, B., Mason, G.M., Mewaldt, R.A.,: Magnetospheric relativistic electron response to magnetic cloud events of 1997, Adv. Space Res. 25, 1387–1392 (2000)

Li, X.L., et al.: Energetic electron injections into the inner magnetosphere during the jan. 10-11, 1997 magnetic storm, Geophys. Res. Lett. 25, 2561–2564 (1998)

Li, X., et al.: Sudden enhancements of relativistic electrons deep in the magnetosphere during may, 1997 magnetic storm, J. Geophys. Res. 104, 4467–4476 (1999)

Li, X., et al.: Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements, Geophys. Res. Lett. 28, 1887 (2001)

Liu, W.W., Rostoker, G., Baker, D.: Internal acceleration of relativistic electrons by large-amplitude ulf pulsations, J. Geophys. Res. **104**, (17),391–17,407 (1999)

Meredith, N.P., Horne, R.B., Iles, R.H.A., Thorne, R.M., Heynderickx, D., Anderson, R.R.: Outer zone relativistic electron acceleration associated with substorm-enhanced whistler-mode chorus, J. Geophys. Res. **107**(A7), 1144, doi:10.1029/2001JA900146 (2002a)

Meredith, N.P., Horne, R.B., Summers, D., Thorne, R.M., Iles, R.H.A., Heynderickx, D., Anderson, R.R.: Evidence for acceleration of outer zone electrons to relativistic energies by whistler-mode chorus, Ann. Geophys. **20**, 967 (2002b)

Meredith, N.P., Cain, M., Horne, R.B., Thorne, R.M., Summers, D., Anderson, R.R.: Evidence for chorus-driven electron acceleration to relativistic energies from a survey of geomagnetically-disturbed periods, J. Geophys. Res. **108**(A6), 1248, doi:10.1029/2002JA009764 (2003)

Miyoshi, Y., Morioka, A., Obara, T., Misawa, H., Nagai, T., Kasahara, Y.: Rebuilding process of the outer radiation belt during the 3 November 1993 magnetic storm: NOAA and Exos-D observations, J. Geophys. Res. **108**(A1), 1004, doi:10.1029/2001JA007542 (2003)

Paulikas, G.A., Blake, J.B.: Modulation of trapped energetic electrons at 6.6 r by direction of interplanetary magnetic-_eld, Geophys. Res. Let. 3 , 277–280 (1976)

Reeves, G.D., Friedel, R., Henderson, M., Belian, D., Meier, M., Baker, D., Onsager, T., Singer, H.: The relativistic electron response at geosynchronous orbit during the January 1997 magnetic storm, J. Geophys. Res. **103** (17)559–17,570 (1998a)

Reeves, G. D., et al.: The global response of relativistic radiation belt electrons to the January 1997 magnetic cloud , Geophys. Res. Lett. 25, 3265–3268 (1998b)

Reeves, G.D., McAdams, K.L., Friedel, R.H.W., O'Brien, T.P.: Acceleration and loss of relativistic electrons during geomagnetic storms, Geophys. Res. Lett., **30**(10), 1529, doi:10.1029/2002GL016513, (2003)

Shprits, Y., Thorne, R., Reeves G., Friedel, R.: Radial diffusion modeling with empirical lifetimes: Comparison with CRRES observations, Annales Geophys. (2005)

Summers, D., Ma, C., Meredith, N.P., Horne, R.B., Thorne, R.M., Heynderickx, D., Anderson, R.R.: Model of the energization of outer-zone electrons by whistler-mode chorus during the October 9, 1990 geomagnetic storm, Geophys. Res. Lett. **29**(24), 2174, doi:10.1029/2002GL016039, (2002)

CHAPTER 4.2

RADIATION EFFECTS ON SPACECRAFT AND COUNTERMEASURES, SELECTED CASES

WOLFGANG KEIL
EADS ASTRIUM GmbH, Friedrichshafen
Wolfgang.keil@astrium.eads.net

INTRODUCTION

At industry limited data is collected and evaluated on spacecraft (S/C) behaviour in orbit due to particle radiation effects. Information on in orbit performance is mainly at space operation centres (e.g. ESOC) or at science institutes, universities (PI). Lessons learned on space system failure is a task relating to product assurance organisation of space companies (alerts, warning notes issuing departments).

DOCUMENTED DATA ON OBSERVED EFFECTS

Various data collections on past events and effects on S/C are evaluated and documented. The most important are:

- NASA-RP-1390, "Spacecraft System Failures and Anomalies Attributed to the Natural Space Environment", excellent overview but fairly old status from 1996
- or on relevant failures with actual data at S/C operation in "Satellite News Digest", http://www.sat-index.com.

SELECTED CASES

Predicted Effects

The S/C design with respect to the requirement to withstand the environment during lifetime is based on the prediction of the environmental source term (i.e. kind and intensity of particles along the mission trajectory), and the evaluation of the effects on materials, electronics, units and system functions. The evaluation of source term is described in standards (e.g. div. ISO, and in particular in ECSS-E-10-04A). The

J. Lilensten (ed.), Space Weather, 231–240.

evaluation of effects is for the time being described in literature on particle physics, and in the "The Radiation Design Handbook", ESA PSS-01-609, which will be revised by ECSS-10-12A, "Methods for the calculation of radiation received and its effects, and a policy for design margins". An example of predicted particle flux effect on LET, with different orbits (GEO, LEO, polar, and SPE conditions) as parameter is shown in Fig. 1. Nevertheless the prediction on S/C effects is a substantial evaluation for each S/C project, where detailed CAD models are the basis for the definition of the effective shielding at target points of question. Furthermore analyses comparable to FMECA are introduced in order to evaluate the effects

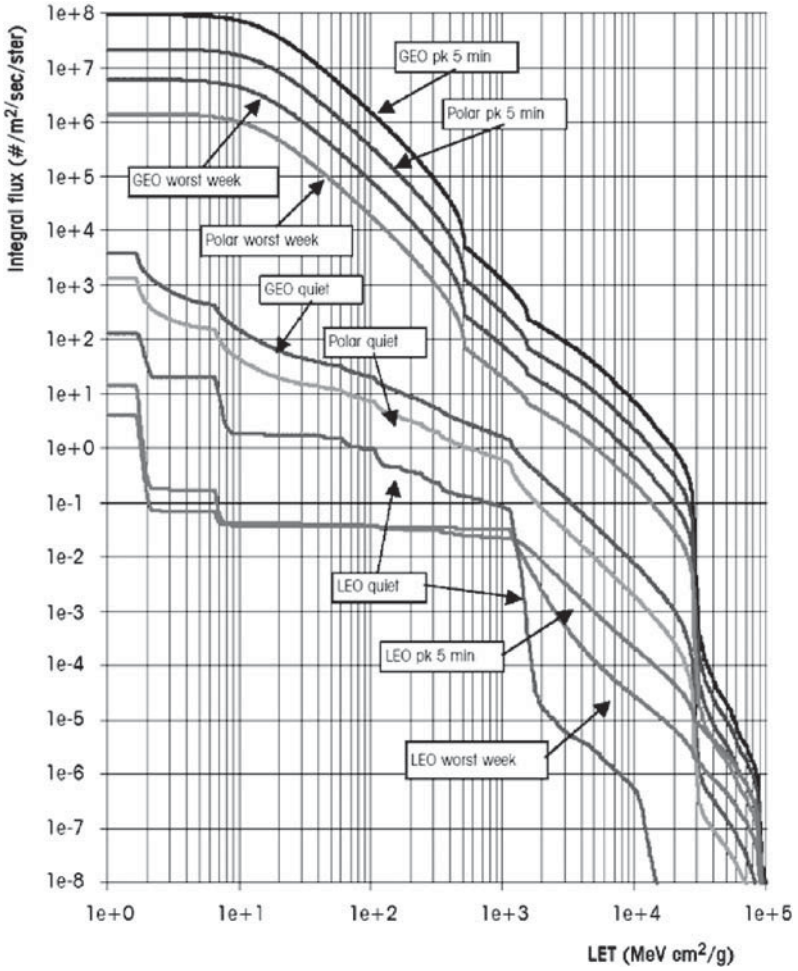


Figure 1. Cosmic ray LET spectra for typical missions (Ref. ECSS-E-10-04A)

on S/C system/subsystem functionality. These analyses are essential for defining optimised countermeasures (e.g. additional shielding, EDAC, filtering).

Manifested Effects

Solar Particle Event (SPE) at 14/15 July 2000 (so-called Bastille event) is a significant impact on ESA scientific satellites where effects on system and electronic components in particular solar cells are observed and investigated.

The proton flux of this SPE , Fig. 2, has not significantly exceeded the 1989 design flare at the SPE from 14/15 July 2000. For particles > 10 MeV its 24 000 pfu is less than 40 000 pfu in maximum (1989 flare).

(pfu) – particle flux unit= $1p^{+}\ cm^{-2}\ sr^{-1}s^{-1}$

The SPE at 14/15 July 2000 h as been the cause of a severe measurable solar cell degradation shown in Fig. 3. As consequence of this event the power dropped about 1.3%. Later SPEs till March 2006 have not induced any comparable strong degradation.

The SPE which is responsible for the first severe solar cell degradation of CLUSTER S/C is characterised with all relevant parameters (proton and electron flux, H_p and K_p indices) in Fig. 4. According to NOAA space weather scale a pfu of ≥ 10000 for this event is a severe solar storm (class S2), occurring about 3 times per solar cycle and has triggered a moderate geomagnetic storm (G2), with $K_p = 6$. The described consequences are due to:

- the solar storm: spacecraft operations: may experience memory device problems and noise on imaging systems; star-tracker problems, and solar panel efficiency can be degraded,

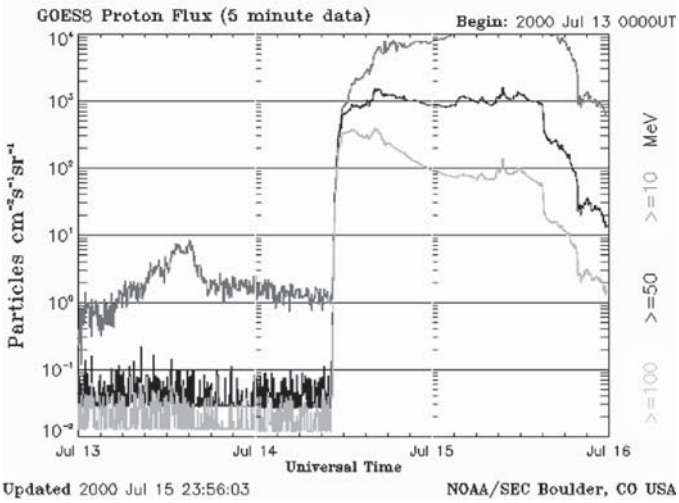


Figure 2. Proton flux at SPE 14/15 July 2000, GOES8 measurements

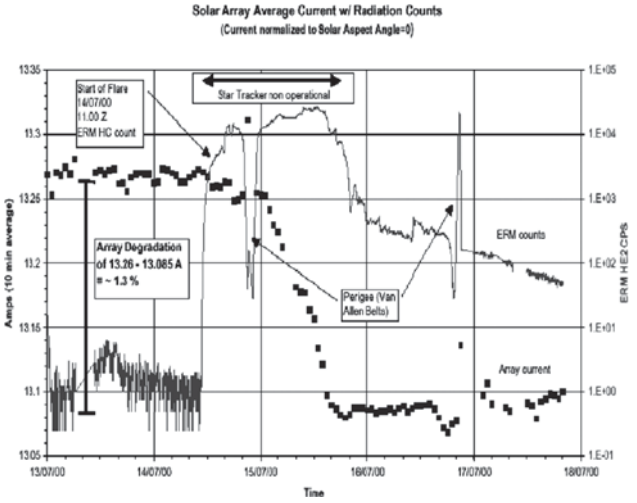


Figure 3. Solar cell current degradation at XMM solar cell and correlation with counts of Radiation Monitor (ERM), as a function of SPE (Bastille event)

- the geomagnetic storm: for spacecraft operations corrective actions to orientation may be required by ground control,

which is herewith confirmed for both spacecraft XMM and CLUSTER, see Fig. 5.

CLUSTER solar cell degradation is $\sim 14.8\%$, from begin of life (BOL) to July 2005, that means in average $< 5\%/\mathrm{y}$, see Fig. 5. Differences on the delivered power

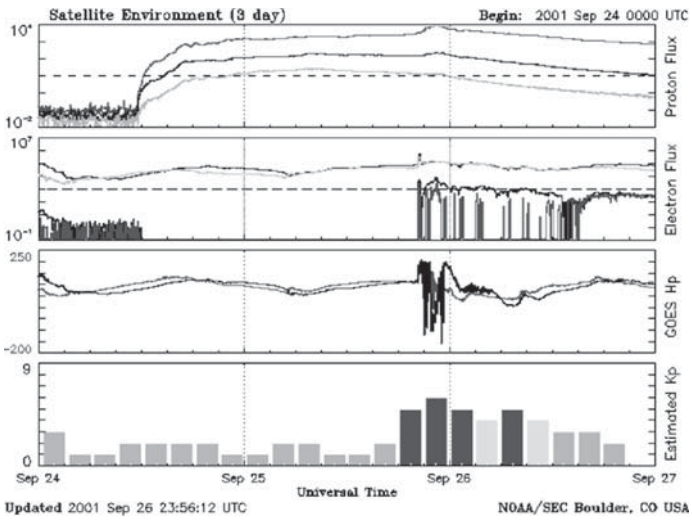


Figure 4. SPE characteristic data at September 2001, GOES measurements

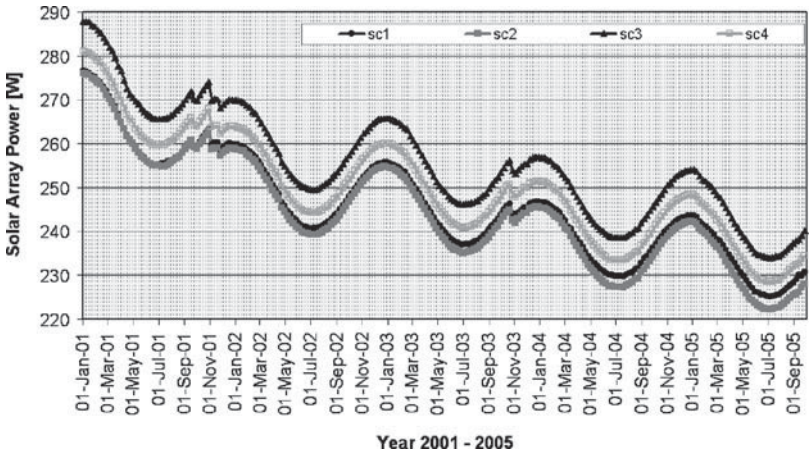


Figure 5. Solar cell power (CLUSTER 4 S/C) as a function of orbit time with significant sudden power degradation at SPEs (September 2001, and October 2003)

are due to different solar cell types (e.g. coverglass thickness). CLUSTER first satellite started from Baikonur Cosmodrome just at the decay of the Bastille SPE, 16 July 2000.

Temporary Effects

An example on temporary bit flips and EDAC behaviour (e.g. mass memory, CLUSTER) is shown in Fig. 6, with the attempt to correlate with SPE.

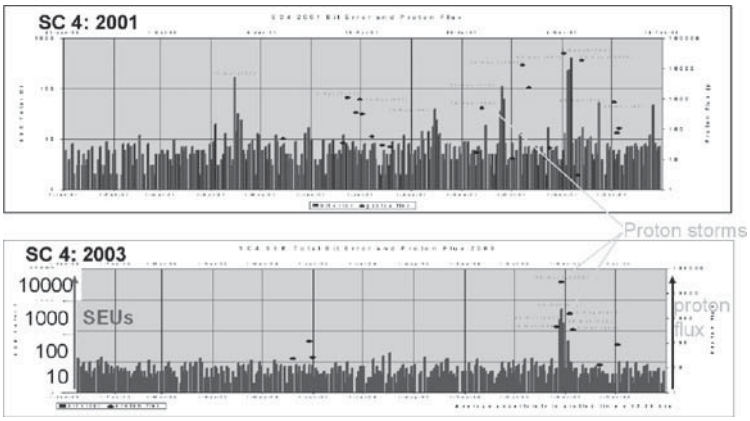


Figure 6. Bit error rates over time (2001, 2003) and dependence on SPE at CLUSTER SC4, (courtesy ESOC/J. Volpp, L. Jagger)

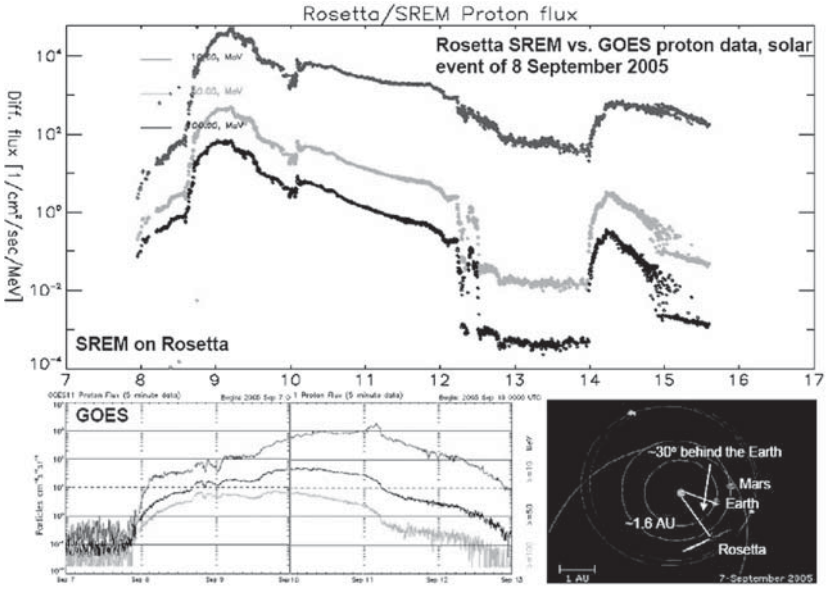


Figure 7. SPE September 2005 on ROSETTA, measured Proton flux at SREM, and position, (courtesy ESTEC, P. Nieminen)

A SPE on 8/9 September 05, hit ROSETTA at the beginning of the weekly non-coverage period. When the signal was acquired for the weekly contact on 15 September the spacecraft was found with the active Star Tracker crashed in INIT mode, and the second Star Tracker (not used for attitude control) in Standby mode.

AOCS had determined the attitude over a period of 6 days using gyroscopes only, and accumulated therefore a drift of about 0.7 degrees, of which 0.3 degrees offset in the High Gain Antenna pointing direction, small enough to allow the RF signal to be received on ground. The recovery activities took most of the ground station pass on 15 September. At the end both Star Trackers were back in tracking mode and the nominal attitude reacquired. The standard radiation monitor (SREM) onboard measured the proton flux, at the same time measured data on GEOS are shown in Fig. 7, with the according position of ROSETTA in relation to earth. Comparing both diagrams show clearly the shock characteristic at ROSETTA, with high gradients in a non earth geomagnetic protected position at about 1.6 AU.

COUNTERMEASURES

To define suitable countermeasures against space radiation effects which are induced by space weather events a detailed knowledge is necessary on:

- The origin of events (environment, source term, type of charged particle)
  - Actual situation (e.g. NOAA GOES measurements, radiation monitoring at orbiting S/C, e.g. XMM-Newton, CLUSTER, ROSETTA).

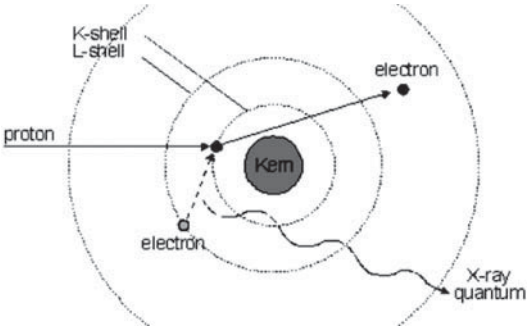


Figure 8. Interaction of Proton with atoms

  - Prediction (e.g. trapped particle models AP-8, AE-8, TREND)
- The cause of effects (particle transport and interaction mechanism with matter leading to activation, and/or secondary particle emission), these are described in basic literature on physics. The most important reaction is the proton interaction with an atom which is in contradiction to ion interaction an indirect ionisation process, see Fig. 8.
- The direct consequence of the physical effects, as shown in Fig. 9:
  - Total Ionizing Dose (effect of ionising secondary particles), described in computer models (e.g. SHIELDOSE, GEANT4)
  - Single Event Effects (SEE), which are induced on powered electronics (SEU, SEL, SEGR, SEB), dealt by computer models (e.g. CREME96, NRL). Examples for the interaction of an ion (GCR) with an electronic part producing a single event upset (SEU) is shown in Fig. 10, and Fig. 11. The principle of a single event latch up (SEL) which is destructive is presented in Fig. 12. At SEL high currents flow in the region of few amperes.
  - Displacement Effects (non ionising effects in lattice induced by protons), a suitable model is available in SPENVIS. Displacement Effects (e.g. solar cell, opt.). These effects are characterised by

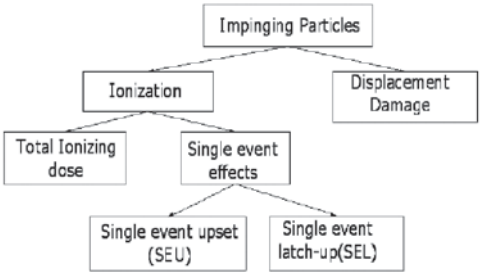


Figure 9. Radiation damage on semiconductor

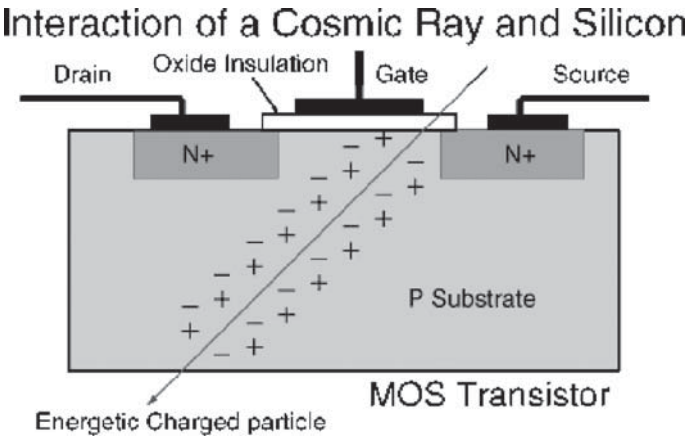


Figure 10. Interaction with galactic cosmic rays (ions) and Si of a MOS Transistor

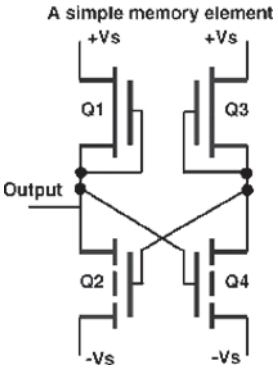


Figure 11. Principle of a Bit flip in Memory Cell (SEU)

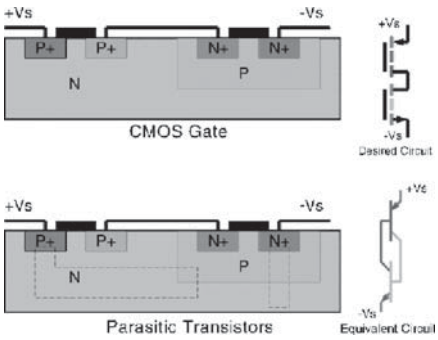


Figure 12. Principle of a single event latch up

  - Non-ionising energy loss (NIEL) manifesting in lattice defects
  - Lattice defects by ejection of atoms from their equilibrium position due to incident particles with suitable kinetic energy
  - Knocked out atom position may be taken by the displacing ion
  - Affected electrical parameters: leakage current, conductivity, mobility of carriers
- The sensitivity of electronic parts can be roughly characterised as shown in Table 1 due to these effects on semiconductors.

Effects on S/C functions are due to the SPE induced effects on electronic parts and materials and which reliability and availability is defined in specific requirements for the S/C mission. Therefore the main differentiation is in payload availability and system availability influenced by space radiation effects.

- Payload Functionality
  - Experiments (affected sensitivity, noise level by activated materials),
  - Instruments (e.g. formation of colour centres at laser crystal, glasses, filters),
  - Sensors (e.g. CCD, APS, HgCdTe),
  - Windows (e.g. darkening of glass, erosion by sputtering)
- Platform System Functionality
  - Power subsystem (e.g. solar cell degradation)
  - Avionics (e.g. star tracker, sun sensor),
  - Propulsion, (impurities by radiolysis or activation at long lasting missions)
  - OBDH (e.g. reduced netto data rates at mass memory due to EDAC operation),
  - TM&TC, (e.g. signal disturbances due to ionospheric conditions)
  - Thermal subsystem (e.g. material degradation optical parameter: α/ε)

The countermeasures which are implemented by Design are:

- Selection of suitable radiation hardened electronic parts and materials, supported by irradiation Tests (total dose Co-60, SEE and material tests with ions/protons)
- Implementation of shielding (intelligent selection of absorbing material with low atomic number Z elements),
- Implementation of EDAC, TMR, filters etc.
- Redundancy of boards, units, sensors (not useful for weak devices)

Table 1. Radiation sensitivity of selected electronic parts

- CMOS (SOS/ SOI)
- CMOS
- APS
- Standard bipolar (bad low dose rate performance, some degrade unbiased)
- Low power Schottky bipolar
- NMOS DRAMs
- CCD (ideal particle counter, SOHO)

Countermeasures by operational measures are still in development.

- Operational concept (e.g. XMM at entering of radiation belts no operation of payload) presently not applicable for stochastic SW events, which needs a reliable prediction capability
- Enhancement of S/C simulator modelling to take into account Space Weather scenario for safeguarding autonomy concepts.

CHAPTER 4.3

AIRCRAFT CREW RADIATION EXPOSURE IN AVIATION ALTITUDES DURING QUIET AND SOLAR STORM PERIODS

PETER BECK
ARC Seibersdorf Research (ARCS), Seibersdorf, Austria

Abstract: The European Commission Directorate General Transport and Energy published in 2004 a summary report of research on aircrew dosimetry carried out by the EURADOS working group WG5 (European Radiation Dosimetry Group, http://www.eurados.org/). The aim of the EURADOS working group WG5 was to bring together, in particular from European research groups, the available, preferably published, experimental data and results of calculations, together with detailed descriptions of the methods of measurement and calculation. The purpose is to provide a dataset for all European Union Member States for the assessment of individual doses and/or to assess the validity of different approaches, and to provide an input to technical recommendations by the experts and the European Commission. Furthermore EURADOS (European Radiation Dosimetry Group, http://www.eurados.org/) started to coordinate research activities in model improvements for dose assessment of solar particle events. Preliminary results related to the European research project CONRAD (Coordinated Network for Radiation Dosimetry) on complex mixed radiation fields at workplaces are presented. The major aim of this work is the validation of models for dose assessment of solar particle events, using data from neutron ground level monitors, in-flight measurement results obtained during a solar particle event and proton satellite data. The radiation protection quantity of interest is effective dose, E (ISO), but the comparison of measurement results obtained by different methods or groups, and comparison of measurement results and the results of calculations, is done in terms of the operational quantity ambient dose equivalent, $H^*(10)$. This paper gives an overview of aircrew radiation exposure measurements during quiet and solar storm conditions and focuses on dose results using the EURADOS In-Flight Radiation Data Base and published data on solar particle events

INTRODUCTION

The Austrian Nobel price winner Victor F. Hess discovered in 1912 during balloon ascent increasing ionizing radiation and described these effects of cosmic origin (Hess 1912). It has been discovered that the Earth is continuously bombarded with high-energy particle radiation from outer space and the Sun. The intensity of the

J. Lilensten (ed.), Space Weather, 241–267.

cosmic radiation is partly decreased by the magnetic field associated with the Sun's solar wind and by the Earth's magnetic field. Galactic energetic charged particles are mostly protons (~ 85%) and helium ions (~ 12%), the rest includes nuclei of all known elements and some electrons. Their energy extends up to about 10^{20} eV. These particles interact with the atmosphere producing secondary radiation, which together with the primary incident particles give rise to ionizing radiation exposure throughout the atmosphere decreasing in intensity with depth from the altitude of supersonic aircraft down to sea level (Fig. 1). Further solar energetic charged particles can contribute to the aircraft crew exposure through occasional so-called solar particle events (SPE's). These are produced by sudden, sporadic releases of energy in the solar atmosphere (solar flares), and by coronal mass ejections (CMEs). During such events a large number of mainly high-energy protons are produced and an increased fluence of particles at aviation altitudes may be observed.

The increase of civil aviation flight altitudes, supersonic flights and space travel, together with updated knowledge about cosmic radiation, motivated already in 1966 the International Commission on Radiological Protection (ICRP) to consider the biological effects of the cosmic radiation to aircraft crew. The ICRP recommendation 1990 stated further that exposure to cosmic radiation during jet flight in aircraft should be included as part of occupational exposure of aircraft crew (ICRP 1991). This initiated a number of new dose measurements onboard aircraft (Beck et al. 1999; O'Sullivan et al. 2002). Computer programs suitable for predicting route doses were further developed and new ones were designed. A large fraction of the available dose measurements results published during the last ten years

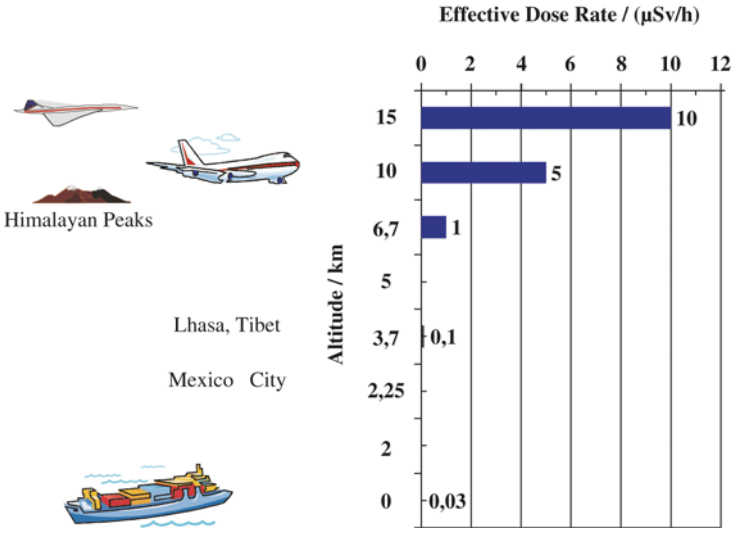


Figure 1. Approximate values of ionizing radiation exposure at different altitudes due to interaction of galactic cosmic radiation with the Earth atmosphere

is compiled in a EURADOS report, published by the European Commission and summarizes descriptions of the instruments and techniques used (Lindborg et al. 2004a). For further information see also the proceedings of the workshops in Luxembourg, 1991 (McAulay 1993) and in Dublin, 1998 (Kelley et al. 1999).

Cosmic radiation has been known for almost a century and for many years airlines cruising at very high altitudes have had their aircraft equipped with instruments to detect possible sudden increases in the dose rate due to solar activities (Hurne 1999). The annual average dose to aircraft crew may become similar to or even larger than that of other occupationally exposed groups (Bartlett 1999, 2004; Spurny et al. 2002; van Dijk 2003). Figure 2 shows average annual doses to workers in the various occupations. The legal consequences of the ICRP Publication 60 were considered by the European Council in its Basic Safety Standards (Directive 96/29/Euratom) (European Community 1996). It says that each member state shall make arrangements for undertakings operating aircraft to take account of exposure to cosmic radiation of air crew who are liable to be subject to exposure to more than 1 mSv per year. The undertakings shall take appropriate measures, in particular: to assess the exposure of the crew concerned, to take into account the assessed exposure when organizing working schedules with a view to reducing the doses of highly exposed aircraft crew, to inform the workers concerned of the health risks their work involves, to apply Article 10 (of the Directive 96/29/Euratom) to female air crew. Article 10 deals with special protection during pregnancy and

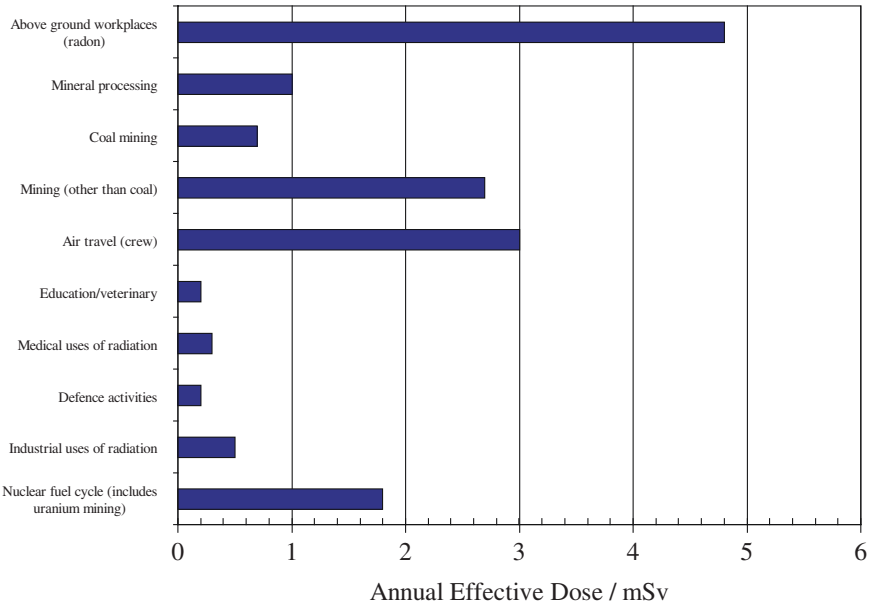


Figure 2. Comparison of radiatio exposure of different occupational workers published by the United Nations Scientific Committee on the Effects of Atomic Radiation (United Nations 2000)

breastfeeding. Its first paragraph reads: "As soon as a pregnant woman informs the undertaking, in accordance with national legislation and/or national practice, of her condition, the protection of the child to be born shall be comparable with that provided for members of the public. The conditions for the pregnant women in the context of her employment shall therefore be such that the equivalent dose to the child to be born will be as low as reasonably achievable and that it will be unlikely that this dose will exceed 1 mSv during at least the remainder of the pregnancy."

Civil aviation is an international business and it is essential that it is regulated in a similar way in different countries. The civil aviation authorities co-operate through an organisation called the Joint Aviation Authorities (JAA, http://www.jaa.nl/). It is an associated body of the European Civil Aviation Conference (ECAC, http://www.ecac-ceac.org/) representing the civil aviation regulatory authorities of a number of European States, which have agreed to co-operate in developing and implementing common safety regulatory standards and procedures. It issues Joint Aviation Requirements (JARs), which usually are implemented as national regulations. The European radiation protection Basic Safety Standards Directive (European Community 1996) is considered in JAR-OPS 1.390 (JAR 2001).

There are two radiation quantities used in aircraft crew dosimetry: One is the protection quantity effective dose E, which is most often used in regulations. This quantity is related, through probability coefficients, to the stochastic health effects in humans that ionizing radiation might cause. E may be used to quantify the degree of protection a regulatory body considers reasonable. E is, however, not a measurable quantity. Therefore, for measurements a quantity called ambient dose equivalent, $H^*(10)$, is defined. The two quantities can be related to each other. However, while $H^*(10)$ is independent of the irradiation geometry, E is not. To be able to convert a value of $H^*(10)$ into a value of E, and in some instances to be able to correctly interpret the indication of an instrument, the direction distribution of the radiation field has to be known. As a simplifying assumption, in most reports about that topic, the radiation field onboard aircraft is taken to be isotropic. The movement of crew may tend to make the field approximately isotropic and the high energies of some radiation components will result in the radiation field geometry being a less important parameter. Both E and $H^*(10)$ have the same unit, sievert (Sv). Sometimes, when it is less important, or when it is obvious which quantity is meant, the general word dose is used instead of the precise names. The cosmic radiation exposure of the body is essentially uniform and the maternal abdomen provides no effective shielding to the foetus. As a result, the magnitude of equivalent dose to the foetus can be put equal to that of the effective dose received by the mother (Lindborg et al. 2004a).

Calculations can be made directly of effective dose per unit time as a function of geographic location, altitude and solar cycle phase, for the assumed field geometry (taken as isotropic). When folded with flight and staff roster information, estimates of effective dose for individuals are obtained. The role of calculations in this procedure is unusual in routine radiation protection. Because effective dose is not

directly measurable, in order to validate assessed values of effective dose based on calculations, calculations are also made of ambient dose equivalent rate or ambient dose equivalent and these compared with values determined by measurements traceable to national standards.

INVESTIGATIONS DURING QUIET SOLAR PERIODS

The in-flight radiation exposure depends on the latitude, longitude, and altitude. This dependence is caused by the influence of the Earth's geomagnetic field on cosmic rays and their subsequent transport through the atmosphere. The additional influence of the Sun's varying activity during the 11-year solar cycle also affects the radiation exposure. The parameters used describing these effects are the heliocentric potential (HCP), and the solar deceleration potential (SDP), both in units of megavolts (MV). The differences between the two approaches are described in (Lindborg et al. 2004a). The comparison of time differential in-flight data has ideally to be done exactly at the same flight position (altitude, longitude and latitude) and under the same solar condition (HCP or SDP). Usually the UTC time combines the link from all in-flight measurement data. In the framework of EURADOS working group WG5 ARC Seibersdorf research established a data base[1] for comparing time differential in-flight dose data. Dose measurements gathered by eleven international research institutes have been collected, analysed and provided to the public (Lindborg et al. 2004a; Bect et al. 2006). The database is organized according the parameters altitude, geographical position (latitude and longitude), and the Sun's activity:

Altitude: Primary cosmic radiation interacts with elements of the top layer of atmosphere producing secondary particles of different types. Primary and secondary cosmic rays create cascades of particles which penetrate the atmosphere. The build-up process of secondary particles competes with simultaneous processes which lead to a reduction of particles fluence rate (Heinrich 1999). As a result fluence rate changes with the depth of atmosphere. Starting with the top of atmosphere and going downwards the fluence rate first increases up to around 20 km of altitude reaching an upper limit, known as Pfotzer maximum (Heinrich 1999), and afterwards decreases down to ground level. In the database the altitude h is the Standard Barometric Altitude (SBA) which is the altitude determined by a barometric altimeter by reference to a pressure level and calculated according to the standard atmosphere definition. The SI unit is the metre but in aviation the unit flight level (FL) is also used, which is one hundredth of the standard barometric altitude expressed in feet.

Geographical position can be described by latitude and longitude and these can be reduced to one single parameter called vertical cut-off rigidity, r_c. Here we use for r_c the unit GV. A particle can enter the atmosphere if its magnetic rigidity, r_p, proportional to the ratio of its momentum and charge, is greater than the vertical cut-off rigidity of the Earth's magnetic field at the point of entry (Heinrich 1999).

[1] EURADOS In-Flight Aircrew Radiation Data Base

A particle with rigidity below the value of r_c is deflected and can not penetrate the atmosphere. In the database a matrix of vertical cut-off rigidity values determined by Shea et al. (1987) is used. These values have been determined for vertical direction of incidence at the altitude of 20 km, for the magnetic field in the year 1990.

Solar activity: To determine the influence of the solar modulation on the intensity of the cosmic rays a diffusion – convection model has been developed by the National Aeronautics and Space Administration (NASA) – Johnson Space Centre (JSC) (Badhwar 1997; Neill in press; Badhwar and Neill 1996). In this model the strength of the solar modulation is described by a parameter called solar deceleration potential (Φ) which calculation is based on CLIMAX (University of Delaware 2006) neutron monitor records and its unit is MV. The value of Φ at a time T depends on CLIMAX neutron monitor count rate averaged over ± 14 days around time $T' = \mathrm{T} - 95$ days, and the polarity of Sun's magnetic field.

In order to provide consistency between different dose measurement methods, all dose rate values in the database are expressed in terms of ambient dose equivalent rate $\dot{H}^*(10)$. Several types of active and passive instruments are used for dose measurements, such as ionization chambers, neutron monitors, tissue equivalent proportional counters (TEPC), Si-spectra dosemeter (LIULIN) and track detectors (PADC[2]). Detailed description of the instruments, measurement methods and the calibration procedures are given in EURADOS final report (Lindborg et al. 2004a).

The Database contains more than 600 in-flight investigations with 15 000 dose rate measurements covering a wide range of altitude, latitude, longitude and solar activity. A summary of maximum parameter ranges is given in Table 1.

The most frequent flight altitude during all in-flight investigations is 10.7 km (FL 350) (see Fig. 3). Regularly used are also 10.1 km (FL 330) and 11.3 km (FL 370). Measurements at altitudes higher than 12.2 km (FL 400) or lower than 8.5 km (FL 280) are very rare. Most of the measurements were performed on the northern hemisphere of the Earth. There are fewer data in the equatorial region and

Table 1. Ranges of parameters in the Database

Description	Data
Number of in-flight measurements	642
Time period	May 1992–Feb 2005
Range of solar deceleration potential	471 MV – 1 320 MV
Range of geographical longitude	180° West to 180° East
Range of geographical latitude	87° North to 62° South
Range of vertical cut off rigidity	0–17.4 GV
Range of barometric altitude	up to 16 500 m
Number of data sets	15 384

[2] PADC detector: track detector based on *poly allyl diglycol carbonate*; sometimes known by the trade name CR-39.

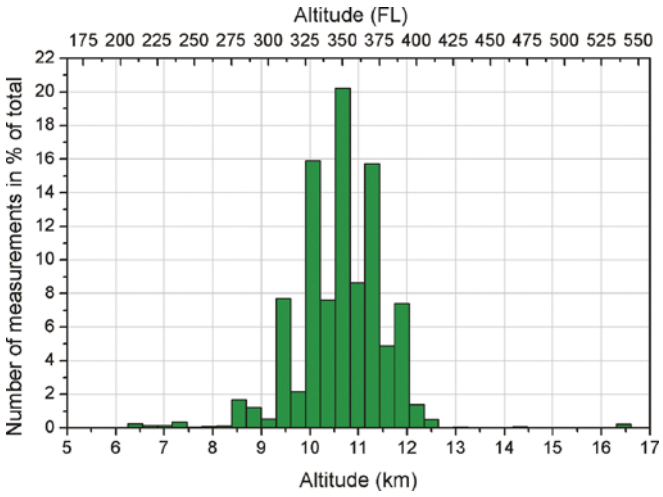


Figure 3. Frequency distribution of in-flight measurements as a function of altitude

the southern hemisphere. Approximately 50% of all measurements were performed at vertical cut-off rigidity less than 2 GV. Figure 4 shows the frequency distribution of in-flight measurements for cut-off rigidity less than 18GV.

The database contains measurements done between the 1992 and the 2005 which corresponds to one cycle of solar activity (471–1320 MV) and are summarized in Tables 2 and 3. Figure 5 shows all investigated flight routes of in-flight measurements on a map of the cut off rigidity r_c in GV for 20km standard

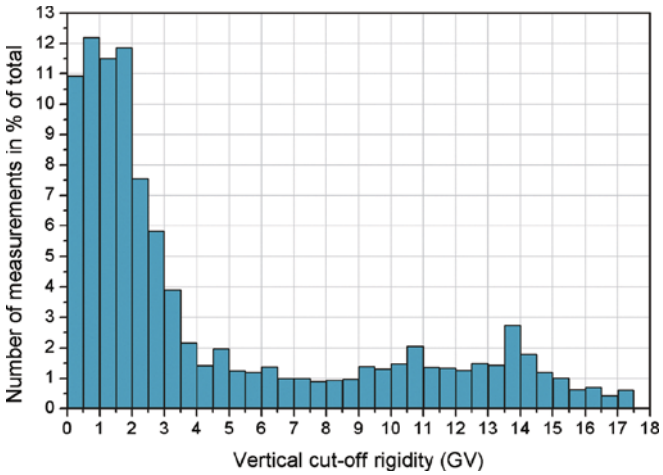


Figure 4. Frequency distribution of in-flight measurements as a function of vertical cut-off rigidity

Table 2. List of in-flight investigations

Institute	Primary investigator	Dose assessment method	Number of data*	Time Period
APAT	Tommasino	TEPC (tissue equivalent proportional counter) and other active instruments	55	1997
ARCS	Beck	TEPC and other active instruments	4896	1997–2005
CIEMAT	Saez-Vergara	TEPC and other active instruments	5680	2001–2002
GSF	Schraube, Regulla	active instruments	308	1992–1993
IRSN	Bottollier	TEPC	49	2002
NPI	Spurny	Si-spectra dosemeter (MDU-Liulin)	811	2001
NPL	Taylor	TEPC	146	2000–2003
NRPB	Bartlett	track detectors and TLDs	19	1997–2003
PTB	Schrewe	active instruments	1240	1997–1999
PTB	Wiegel	Bonner Spheres, Ionization Chamber	43	1998
RMC	Lewis	TEPC	342	1999
SSI	Lindborg	TEPC	66	1998–2003

*EURADOS Radiation In-Flight Data Base.

barometric altitude (SBA) based on the Earth's magnetic conditions in 1990 (Shea et al. 1968; Sheq and Smart 2001). A lower r_c shows a lower magnetic shielding, and higher values show a higher shielding. Table 2 shows a list of instruments used, with their abbreviations and the measurement integration time. For ionization chambers and neutron monitors, an integration time of 5 to 10 minutes were agreed; for TEPC instruments 25 minutes up to 60 minutes; for passive dosemeter an integration time of 32 hours over 16 flights was used to get adequate accuracy.

As examples, we show in Fig. 6 and Fig. 7a comparison between in-flight measurements of the in-flight data base with a model prediction developed by ARCS Latocha et al. (in press), and a calculation program called EPCARD (Schraube et al. 2002). Figure 6 presents ambient dose equivalent rate as a function of altitude for a solar deceleration potential between 470 MV and 490 MV and cut-off rigidity lower then 2 GV. Both for the ARCS model and the EPCARD calculations values of $\Phi = 475$ MV and $r_c = 1$ GV are assumed. Figure 7 shows ambient dose equivalent rate as a function of vertical cut-off rigidity. It presents measurements performed between 9.9 km (FL325) and 10.2 km (FL335) of altitude and solar deceleration potential between 470 and 610 MV. Dose rate calculations were done for $h =$ 10.06 km (FL330) and $\Phi = 475$ MV. Experimental and calculated data agree within the measurement uncertainty of about 25%.

Figure 8 (taken from (Lindborg et al. 2004a) shows calculated estimates of the minimum number of flight hours, which are needed to obtain annual effective dose

Table 3. List of dose rate assessment methods during in-flight investigations

Abbreviation	Dose assessment method measurement/ calculation	Measurement intervals
NM+IC (ARCS)	Combined neutron monitor (NM) LB6411 and ionization chamber (IC) RSS (Beck et al. 1999a, b)	5 min
NMX+IC (PTB)	Combined neutron monitor NE-NM2 with lead converter (NMX) and ionization chamber (Schrewe 1999; Schrewe et al. 2000)	5–20 min
ACREM (ARCS)	Combined GM detector and transport code calculations (Beck et al. 1999a, b)	5 min
NMX+Halle(GSF)	Combined neutron monitor NE-NM2 with lead converter (NMX) and low level scintillation detector DLM7908 (Regulla and Davis 1993; Regulla and Schraube 1996)	6 min
TEPC-log (ARCS)	TEPC detector, 12 cm sphere, logarithmic amplifier [1, 5, 6]	30–60 min
TEPC (ARCS)	TEPC (HAWK) (O'Sullivan et al. 2002; Lindborg et al. 2004a)	30–60 min
TEPC (RMC)	TEPC (FAR WEST detector) (Green et al. 2000; Lewis et al. 2001)	25 min
TEPC (SSI)	TEPC instruments based on the variance method (Kyllönen et al. 2001)	30–60 min
TEPC (CIEMAT)	TEPC (HAWK) (Lindborg et al. 2004a; Saez Vergara et al. 2002; Romero et al. 2004)	25 min
NMX+IC (CIEMAT)	Combined neutron monitor with tungsten converter (NMX) SWENDI-2 and ionization chamber (IC) RSS (Lindborg et al. 2004a)	5 min
LIULIN (NPI)	Si-Spectra-dosemeter developed originally for space (MDU-Liulin) (Spurny and Dachev 2003)	30 min
Track Detector (NRPB)	Box with 36 PADC and 30 TL dosemeters (Barlett et al. 2000, 2001, 2003)	16×120 min
TEPC (NPL-PPARC)	TEPC (HAWK) (Taylor et al. 2002)	30 min
EPCARDv3.2	European Program Package for the Calculation of Aviation Route Doses (Schraube et al. 2002)	single point calculation
TEPC (IRSN)	TEPC (HAWK) (Bottollier-Depois et al. 2004)	30 min
TEPC (APAT)	TEPC (HANDI) (Tommasino 1999)	60 min
NMX+IC (APAT)	Combined neutron monitor LINUS with tungsten converter (NMX) and ionization chamber RSS (IC) (Wiegel et al. 2002)	5 min
BSS+IC (PTB)	Bonner Spheres (BSS) and ionization chamber (IC) (Beck et al. 1999a; Wiegel et al. 2002).	30–60 min

values of 1 mSv and 6 mSv. The calculations were performed for January 1998, i.e. around solar minimum activity and therefore about the highest possible doses from cosmic radiation at flight altitudes. The three lowermost lines are for the equatorial region, the three uppermost for the polar region. For each region, the flight altitudes chosen are 9 144 m (30 000 ft) lower line, 12 192 m (40 000 ft) and 15 240 m (50 000 ft) upper line.

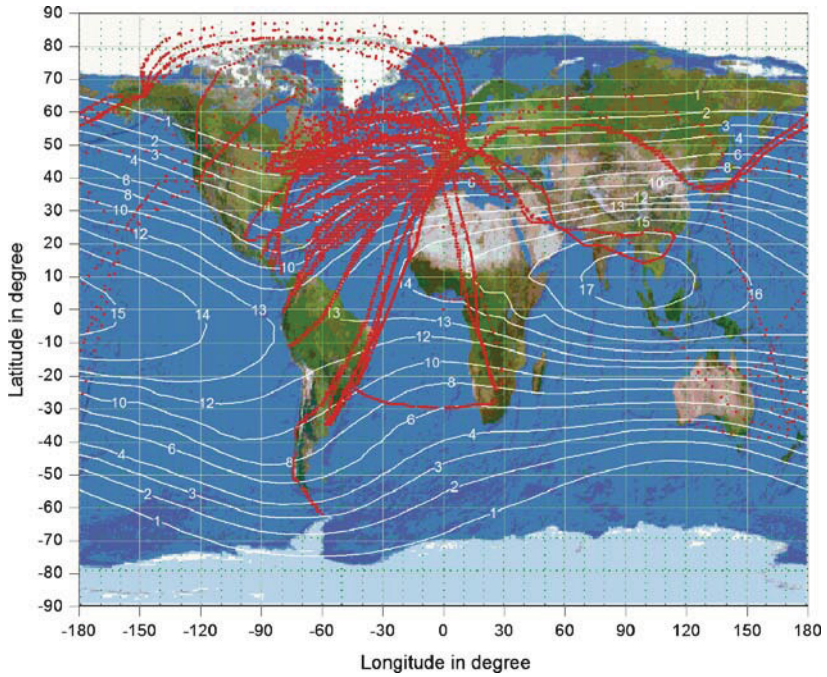


Figure 5. Investigated flight routes between 1992 and 2005 shown at vertical cut-off rigidity r_c in GV at 20 km altitude

INVESTIGATION DURING SOLAR STORM PERIODS

Radiation exposure at flight altitudes due to solar events have been discussed in the EURADOS working group report (Lindborg et al. 2004a). The radiation assessment of these events caused by specific space weather conditions are still under research. Again, by an initiative of EURADOS a coordinative research project were started on modelling and validation of the radiation exposure due to solar particle events. The investigations are part of the European research project CONRAD (Coordinated Network for Radiation Dosimetry) (Beck et al. 2006). Work package 6 is concerned with the co-ordination of research in EU member states on the evaluation of complex mixed radiation fields at workplaces and is organized in two task-groups. One task-group existing of 13 members from 12 different European research institutes is working on aircraft crew radiation workplaces. The objective of this task-group is to coordinate research activities in model improvements for radiation dose assessment due to solar particle events. The results will aid European research, increase the efficiency of resource utilization, and facilitate the technology transfer to practical application and support the development of standards. Main objectives of the task-group on solar energetic particle events will be the validation of models for dose assessment of solar particle events, using data from neutron ground level monitors,

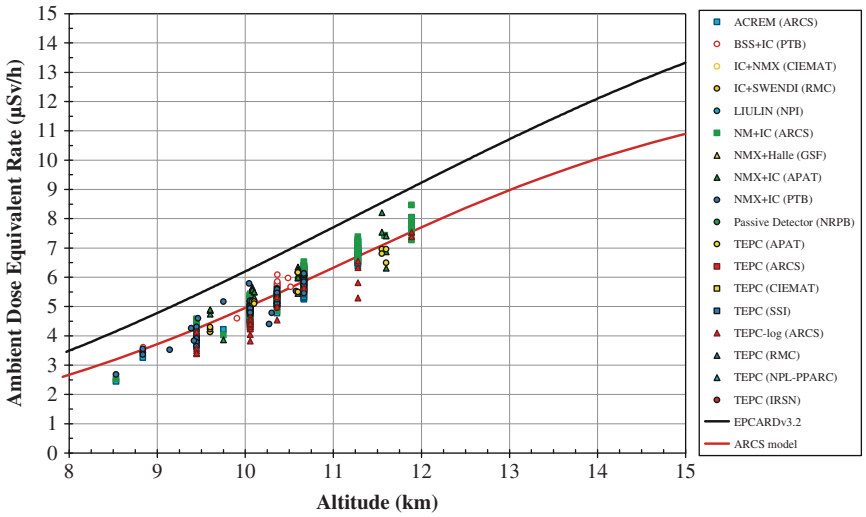


Figure 6. Ambient dose equivalent rate as a function of altitude for measurements (colour points), EPCARDv3.2 calculation (black line) and the ARCS model (red line). Measurements are selected from the database for $\Phi = [470\text{–}490]$ MV and $r_c = [0\text{–}2]$ GV. EPCARD and model calculations assumed $\Phi = 475$ MV and $r_c = 1$ GV

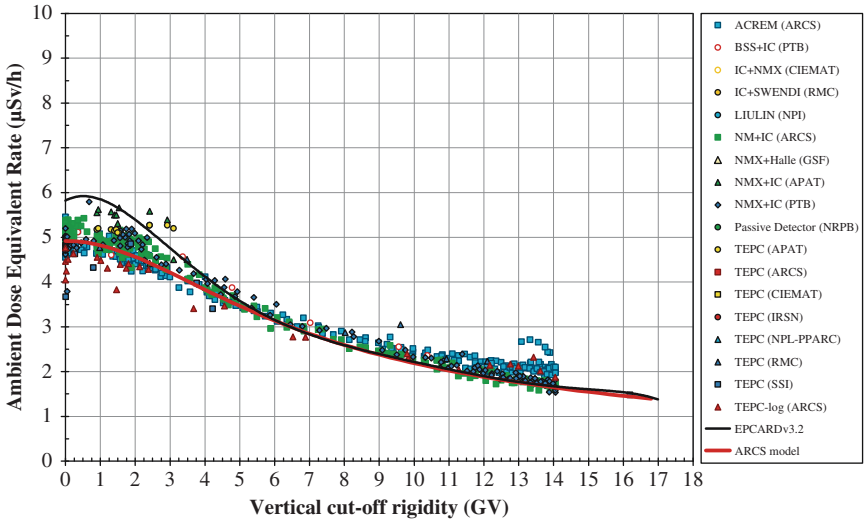


Figure 7. Ambient dose equivalent rate as a function of vertical cut off rigidity for measurements (colour points), EPCARDv3.2 calculations (black line) and the ARCS model (red line). Measurements are selected from the database for $\Phi = [470\text{–}610]$ MV and $h = [9.9\text{–}10.2]$ km (FL325–FL335). Calculations assumed $\Phi = 550$ MV and $h = 10.06$ km (FL330)

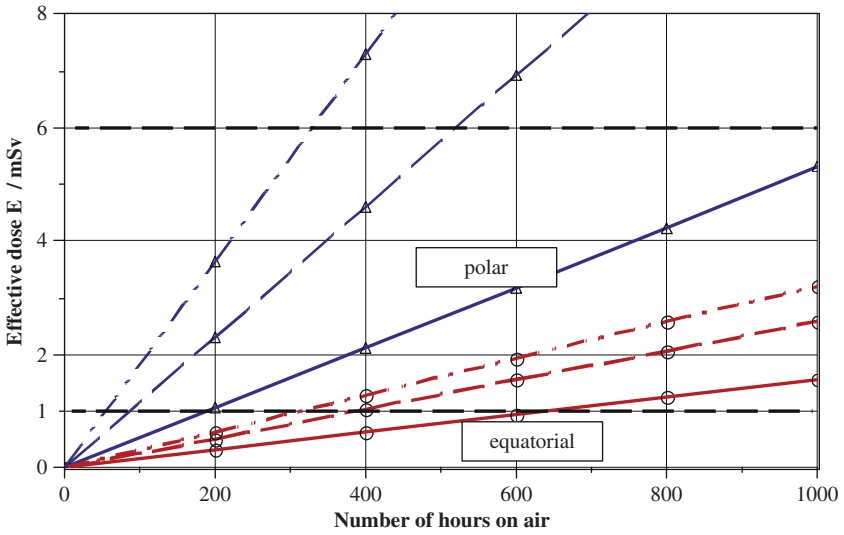


Figure 8. Diagram to enable estimate of the minimum number of flight hours to obtain annual effective dose values of 1 mSv and 6 mSv. The three lowermost lines – indicated by sphere symbols – are for the equatorial region; the three uppermost – indicated by triangular symbols – are for the polar region. For each region, the flight altitudes chosen are 9 144 m (30 000 ft) [lower line], 12 192 m (40 000 ft) and 15 240 m (50 000 ft) [upper line] (Lindborg et al. 2004a)

in-flight measurement results obtained during a solar particle event and proton data measured by instruments onboard satellite.

Solar energetic events (e.g. solar flares, coronal mass ejections) are huge particle outbursts occurring at the surface of the sun, accompanied with emission of radiation from radio frequencies up to gamma rays and GeV protons, with variable time profile (minutes to several hours and days), different extension of the location of origin and different involvement of solar mass (Allkofer 1975). There is a correlation existing between the sun spot numbers and the frequency of the appearance of solar particle events. Figure 9 shows a correlation between solar flares and sunspot numbers. Sun spot numbers vary in a time period of 11 years and indicate the so called solar cycle by increasing and decreasing of sun activity.

While several 1 000 solar energetic events were observed in the past, since 1942 only some 70 of these events have lead to effects observed on the ground, so called ground level events (GLE). Table 4 shows a list of all registered GLEs between February 1942 and January 2005. On average about one event per year was observed. Figure 10 shows a correlation between GLEs and sun spots demonstrating that the occurrence of GLEs is very likely around the maximum of sun spot frequency, slightly higher before and after the maximum, but clearly less probable during minimum. The energy spectra of solar protons vary greatly from event to event and are very different to the spectra of galactic cosmic rays. In general the spectra are much softer but can involve greatly increased intensities at low energies.

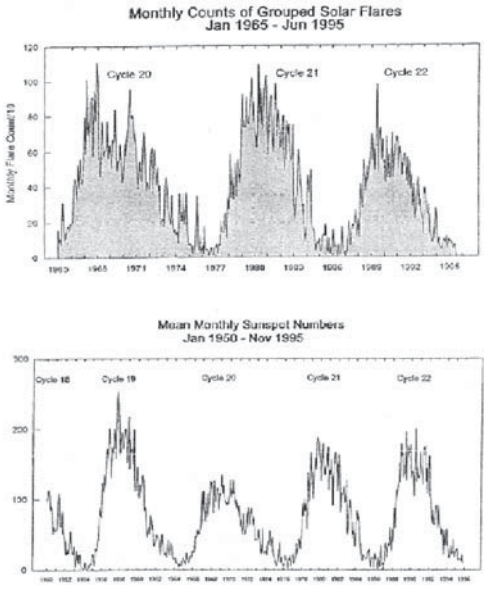


Figure 9. Correlation between solar flares and sun spot numbers

Table 4. Ground level event enhancements list (http://neutronm.bartol.udel.edu/)

EVENT#	Day	Month	Year
1	28	February	1942
2	07	March	1942
3	25	July	1946
4	19	November	1949
5	23	February	1956
6	31	August	1956
7	17	July	1959
8	04	May	1960
9	03	September	1960
10	12	November	1960
11	15	November	1960
12	20	November	1960
13	18	July	1961
14	20	July	1961
15	07	July	1966
16	28	January	1967
17	28	January	1967
18	29	September	1968
19	18	November	1968
20	25	February	1969

(Continued)

Table 4. (Continued)

EVENT#	Day	Month	Year
21	30	March	1969
22	24	January	1971
23	01	September	1971
24	04	August	1972
25	07	August	1972
26	29	April	1973
27	30	April	1976
28	19	September	1977
29	24	September	1977
30	22	November	1977
31	07	May	1978
32	23	September	1978
33	21	August	1979
34	10	April	1981
35	10	May	1981
36	12	October	1981
37	26	November	1982
38	07	December	1982
39	16	February	1984
40	25	July	1989
41	16	August	1989
42	29	September	1989
43	19	October	1989
44	22	October	1989
45	24	October	1989
46	15	November	1989
47	21	May	1990
48	24	May	1990
49	26	May	1990
50	28	May	1990
51	11	June	1991
52	15	June	1991
53	25	June	1992
54	02	November	1992
55	06	November	1997
56	02	May	1998
57	06	May	1998
58	24	August	1998
59	14	July	2000
60	15	April	2001
61	18	April	2001
62	04	November	2001
63	26	December	2001
64	24	August	2002
65	28	October	2003
66	29	October	2003
67	02	November	2003
68	17	January	2005
69	20	January	2005

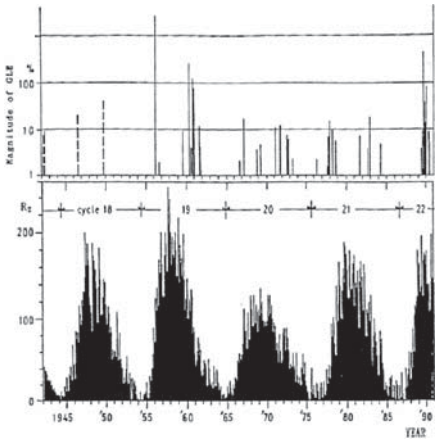


Figure 10. Correlation between GLEs and sun spot numbers

The events of interest here are those with hard spectra extending to several GeV. Figure 11, taken from (Dyer et al. 2003), shows the calculated proton energy spectra of some major solar events compared to cosmic radiation during solar maximum. The increase of the lower energy part of the spectra leads to a significant increase

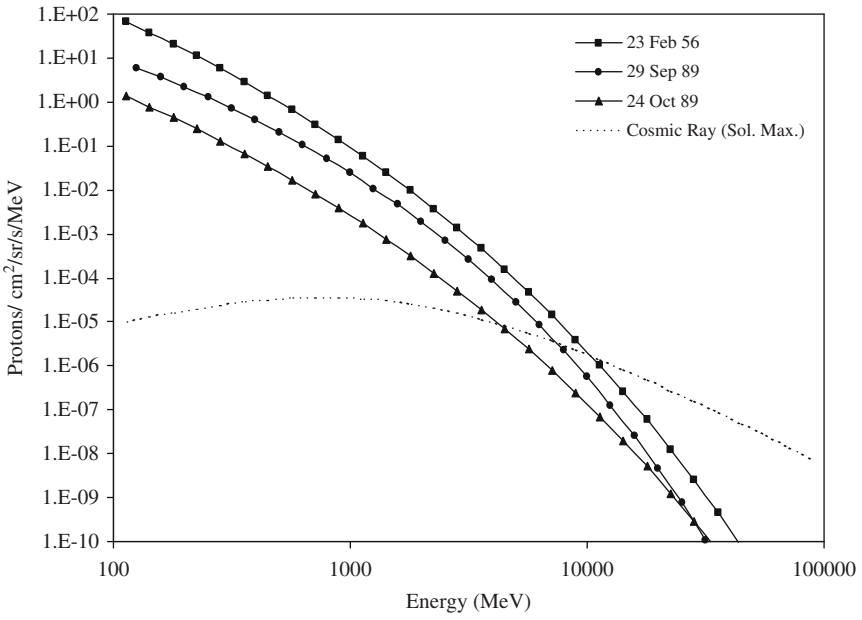


Figure 11. Energy spectra of solar protons during solar events compared to cosmic ray protons during solar maximum

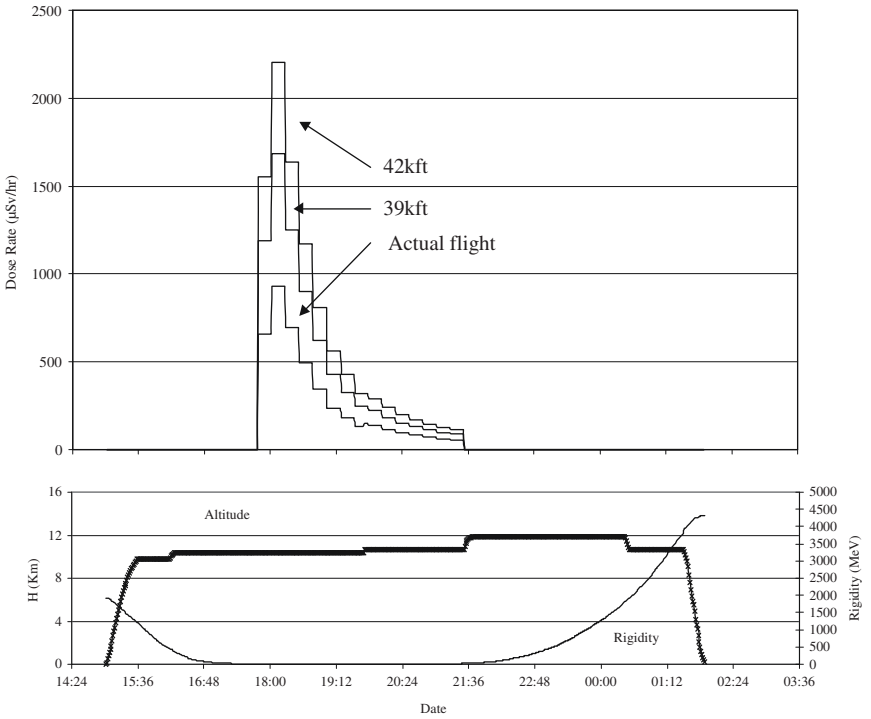


Figure 12. Estimation for the dose rate during a fight between London and Los Angeles on 23 February 1956 for 12.8 km (42 000 ft), 11.9 km (39 000 ft), and 10.5 km (35 000 ft)

of the radiation exposure on board aircraft where the magnetic shielding is less such as close to the Polar region. Usually in the equatorial region the radiation exposure shows almost no increase due to a ground level event. Summarizing these effects a radiation exposure up to 2mSv/h in 12 km altitude (39 000 ft) can be found in the literature (Dyer et al. 2003). Figure 12 (taken from (Dyer et al. 2003)) shows estimates for the dose rate profile during a flight between London and Los Angeles on 23 February 1956.

It should be noted that solar particle spectra are difficult to measure and calculate as they cover a very wide energy range. It is necessary to combine information from both space borne detectors, covering the range up to several hundred MeV, with ground level neutron monitors covering GeV energies. Additional complications arise from the anisotropic nature of certain of these energetic events and from concurrent geomagnetic storms that can increase the penetration of particles at high latitudes.

Measurements of the radiation exposure on board aircraft from the galactic cosmic radiation component have been done during the last decade quite extensively (Lindborg et al. 2004a). Some of the research institutes installed radiation monitors for long-term investigations at aircraft (Spurný and Dachev 2003; Beck et al. 2005).

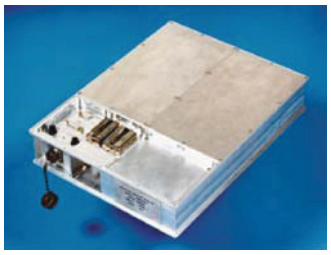
For those institutes it was possible to gather also data during sporadic solar particle events. For monitoring of radiation exposure due to solar events active radiation instruments were used such as tissue equivalent proportional counter (TEPC), Geiger-Müller counter or Si-semiconductor detectors or spectrometers. Figure 13 shows several types of active radiation monitors used by different institutes.

Radiation measurement results of the radiation exposure were performed during several solar particle events (Bartlett et al. 2002). Figures 14 and 15 show in-flight measurement results during GLE60 on 15 April 2001 (Bartlett et al. 2002; Spurný and Datchev 2001). Both measurements show a significant increase of the radiation exposure during the flight for a time period of up to some 3 hours. The total increase in terms of the radiation quantity $H^*(10)$ was about 50% due to that ground level event.

Figure 16 shows the whole scenario of radiation exposure, the proton fluence rate measured by satellite, and the ground level neutron monitor count rate during a solar storm which happened between October and November 2003 (Halloween storms) (Beck et al. 2005). The radiation monitor used which was fix-installed on-board a Lufthansa Airbus A340, is a TEPC showing the low- and high-LET contribution separately. The ratio is significantly different between GLE65 on 28 October 2003 and the following Forbush decrease. Figure 17 shows the relative deviation of the radiation exposure in the radiation quantity $H^*(10)$ during the solar storm period compared with quiet periods before and after the storm. While the variation of radiation exposure for the same flight routes is about $\pm 10\%$ in 12 km flight altitude, the deviation during the Halloween storm was about $\pm 40\%$ due to GLE65 and the following Forbush decrease. On the other hand the Forbush decrease that started on the 29 October 2003 (see Fig. 16) has led to few days lasting rather important decrease of the exposure onboard aircraft. The measurements performed with Liulin equipment onboard a Czech Airlines aircraft during the flight Sofia-Prague just in the deep decrease showed that the exposure was about 28% lower when compared to normal solar activity conditions (Spurný et al. 2005).

The radiation exposure due to solar particle events were estimated by some members of the EURADOS task-group based on models, neutron ground level monitor data, proton data measured by satellites, and using in-flight data measured onboard aircraft during GLEs (e.g. SiGLE, QARM). Several institutes of the task-group are working on complex calculation methods based on SPE proton input spectra, atmospherics and magnetic models, and using high energy transport Monte-Carlo codes (e.g. FLUKA, GEANT4, and MCNPX). Figure 18 show the radiation exposure estimations for flights between Paris and New York with Concorde aircraft and between Paris and San Francisco for sub-sonic flights during 31 GLEs (1994 to 2001) (Lantos and Fuller 2003). Total doses up to some 5mSv per flight were calculated. Most of the increasing radiation dose due to GLEs is in the order of some 100 μSv per flight. Figure 19 shows a world dose map calculated for the radiation conditions during the recently occurred GLE69 on 20 January 2005 based on the SiGLE model (Lantos 2006). Figure 20 shows a time profile of calculated dose rate data during GLE 60 on 15 April 2001 during a flight from FRA to DFW.

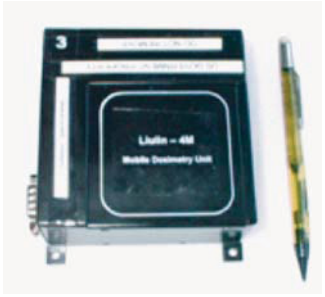
Tissue Equivalent Proportional Counter (TEPC, HAWK) installed on board Airbus A340, measurements on board space shuttle and Boeing 747 (Lindborg et al., 2004a; Beck et al., 1999b; Lindborg et al., 2004b)



Cosmic Radiation Effects & Activation Monitor (CREAM) as installed onboard space shuttle, Boeing 747, executive jets, and Mir (Dyer et al. 1999; Dyer et al. 2005)

Air Crew Radiation Exposure Monitor (ACREM, GM-monitor) installed on Board Airbus A340, connected to aircraft flight bus (Beck et al. 1999a)

Dose Telescope (DOSTEL, Si dose and spectrometer) measurements onboard space shuttle, space station and onboard aircraft (Beaujean et al., 2005)



Liulin Si dose and spectrometer measurements onboard space shuttle, space station and onboard several aircraft (Spurný and Dachev, 2002)

Cosmic Radiation Effects & Activation Monitor (CREAM): "Concorde" version flown on BA Concorde, SAS Boeing 767 and NASA WB-57F (Dyer et al., 1999)

Figure 13. Several active dosimeter instruments used for long term investigations onboard space- and aircraft

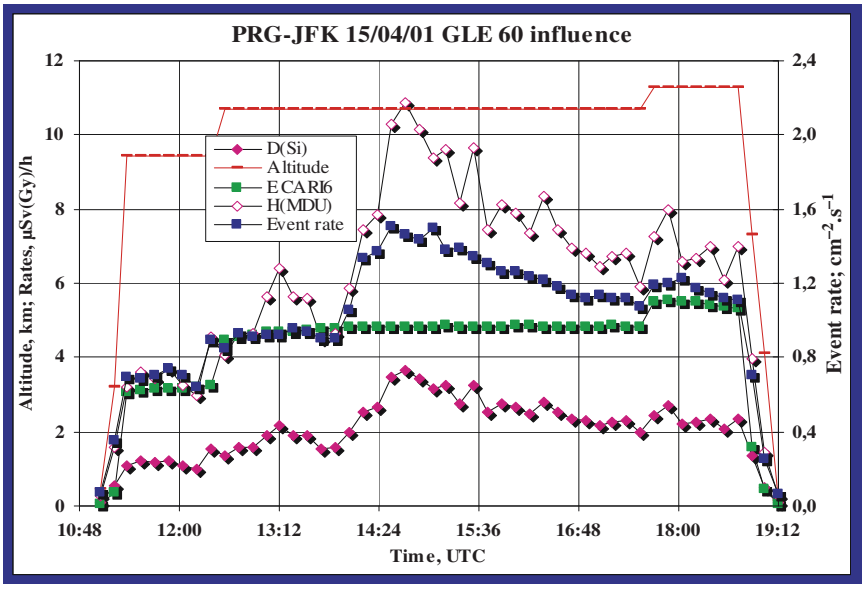


Figure 14. LIULIN measurements of GLE 60 during PRG-JFK flight (Spurný and Datchev, 2001)

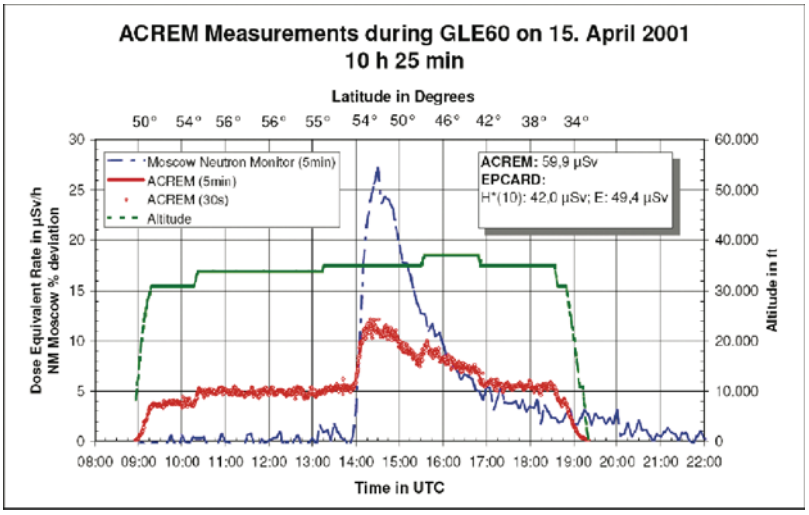


Figure 15. ACREM in-flight measurement of GLE 60 during FRA-DFW flight and comparison with neutron monitor data from the ground station at Moscow (Bartlett et al., 2002)

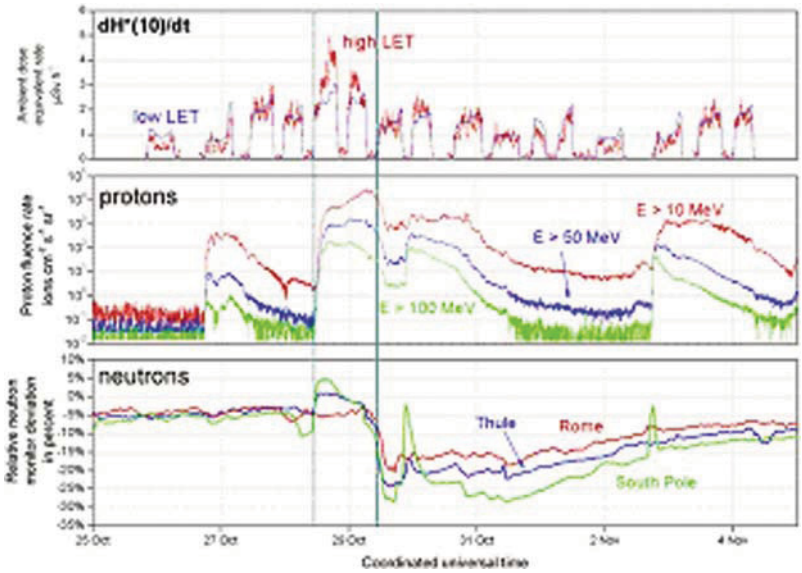


Figure 16. TEPC measurements during the Halloween storms between October and November 2001. GLE65, GLE66 and GLE67 occurred during these time period (Beck et al., 2005)

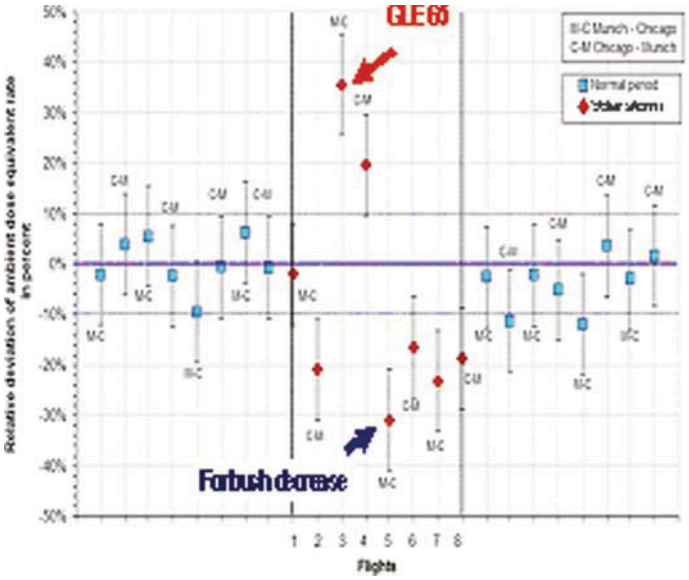


Figure 17. Observation of about 40% increase and 30% decreases of the radiation exposure in 12 km due to GLE 65 and the following Forbush decrease (Beck et al.,2005)

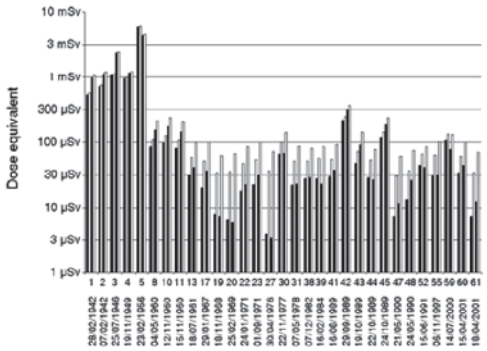


Figure 18. Worst case calculation for 31 GLEs for flights between Paris and New York with Concorde (1st and 2nd bar) and Paris to San Francisco with subsonic aircraft (3rd and 4th bar). GLE contribution (black bar), total dose (white bar). The GLE numbers along horizontal axis are those of Table 4. The GLEs taken into account are those significant in terms of radiation dose

The radiation dose data were calculated by the QARM model (Lei et al. 2004). A comparison with the measured data shown in Fig. 15 indicates a difference of order of a factor of 2 in dose rate. Since disturbances of the magnetic field were not taken into account in that preliminary comparison of measured and calculated data, the agreement seems reasonable. Further investigations will be needed to improve the models and provide reliably dose rate estimations.

SUMMARY AND CONCLUSION

Cosmic radiation is included by ICRP in radiation protection recommendations. The European Council Directive 96/29/EURATOM (1996) gives requirements for the protection of aircraft crew. The Joint Aviation Regulations (of the Joint Aviation

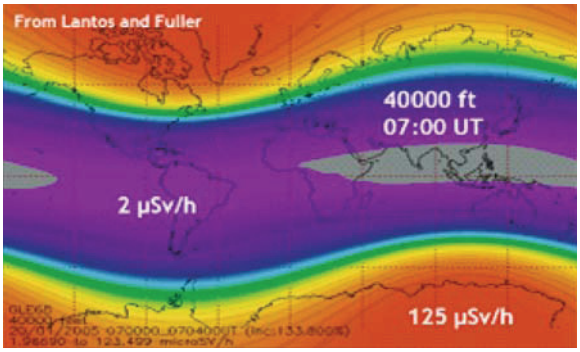


Figure 19. World dose map calculated for the conditions during GLE 69 on 20 January 2005 based on SiGLE model (Lantos, 2006)

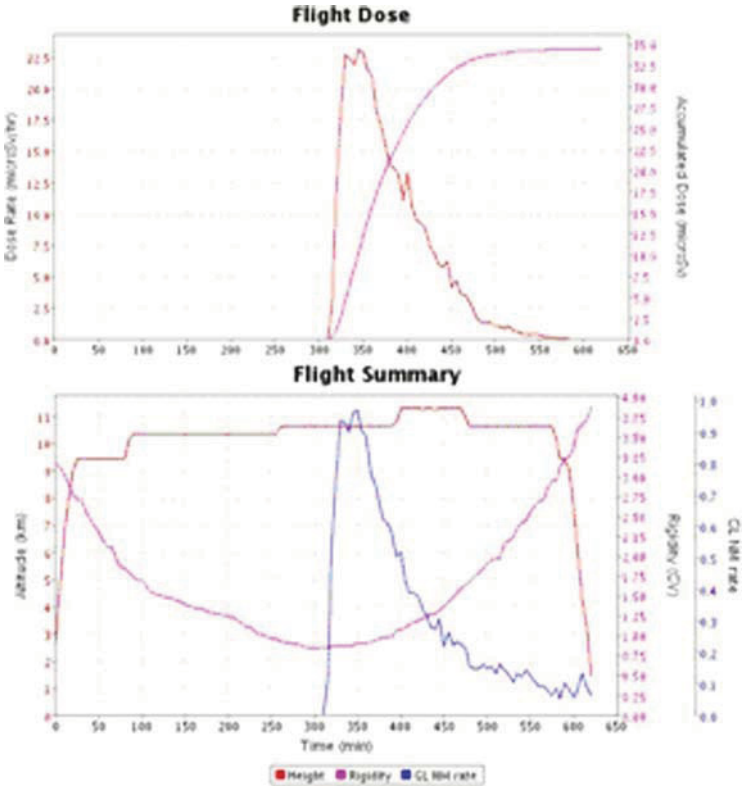


Figure 20. Calculated time profile dose rate data during GLE 60 on 15 April 2001 on a flight from FRA to DFW. Radiation data were calculated by the QARM model (Beck in press)

Authority which covers the activities of the civil airlines of 35 European states) introduced similar requirements in 2001. Operators should make arrangements to give information and provide education regarding the risks of occupational exposure to radiation to their air crew, air crew being defined as flight crew, cabin crew and any person employed by the aircraft operator to perform a function on board the aircraft while it is in flight. Female air crew should be made aware of the need to control doses during pregnancy and to notify their employer if they become pregnant so that any necessary dose control measures can be introduced.

In general, no controls are necessary for a crew member whose annual dose can be shown to be less than 1 mSv. Operators whose aircraft crew may receive an effective dose greater than 1 mSv, generally those operators whose aircraft operate above 8km (26 000ft), should carry out an assessment, by computer program prediction, of the maximum annual dose to which their air crew are liable. The details of these assessments of exposure must be recorded. Operators should adjust an air crew member's roster to reduce exposure. National regulations or guidance may require

that individual monitoring be carried out if doses exceed 6 mSv per annum, with full record keeping, and, in some cases, medical surveillance.

The favoured approach to individual monitoring is to assess doses from calculations of dose rate as a function of altitude, geomagnetic latitude and longitude or cut-off, and phase of solar cycle, combined with flight profiles and staff rosters. The calculated doses should be supported or validated by measurements.

The provisions of Article 10 of the EC Directive apply to pregnant air crew and, once pregnancy is declared, the protection of the foetus should be comparable with that provided for members of the public. This means that, once the pregnancy is declared, the employer must plan future occupational exposures such that the equivalent dose to the foetus should be kept as low as reasonably achievable, and such that it is unlikely to be greater than 1 mSv during the remainder of the pregnancy. The cosmic radiation exposure of the body is essentially uniform and the maternal abdomen provides no effective shielding to the foetus. As a result, the magnitude of equivalent dose to the foetus can be put equal to that of the effective dose received by the mother. In accordance with the requirement of keeping doses as low as reasonably achievable, some operators have determined that pregnant aircraft crew ceases flying duties on declaration of pregnancy.

Aircraft capable of operating at altitudes greater than 15 km (49 000ft) should carry an active radiation monitor, which monitors current levels of radiation, to detect any significant short-term variation in radiation levels during flight. In principle, the need to detect high dose rates could be achieved by some means other than an on-board monitor e.g. satellite or ground based solar monitoring systems. However, at present, such techniques provide no more than retrospective estimates of dose. Further coordinated European research activities are carried out by EURADOS experts.

The results of the detailed EC-supported research on the radiation exposure due to galactic cosmic radiation have been published as European Commission project reports and elsewhere.

A EURADOS expert group co-ordinates since 2005 research on effects of solar energetic particles at aircraft altitudes, and is carrying out validation of the models used to evaluate the radiation exposure on board aircraft due to GLEs.

The radiation exposure to aircraft crew depends on flight route, flight time and solar cycle phase. UNSCEAR 2000 gives an occupational effective dose value of 3 mSv p.a. during quiet solar periods. The occurrence of solar particle events, which lead to a further radiation exposure to aircraft crew is very rare; one average about one event per year. Theses events are most likely before and after the solar maximum. The radiation exposure has been estimated to between several 100 μS and some milli sievert for one civil aviation flights close to the poles during the occurrence of a solar event. The cosmic radiation exposure to aircraft crew leads in total to an increase of the cancer risk of about 1% over the whole working live.

Although considerable progress has been made within the research programmes carried out during the last full solar cycle, there remains a need for research and

development support in this new area of radiation protection for a number of reasons:

- The annual doses are significant compared to mean annual doses of radiation workers in the nuclear and medical sectors, and, relatively, more aircraft crew are both young and female, characteristics linked to higher lifetime risk. Furthermore, more than 50% of the doses to aircraft crew are due to high-LET radiation, for which the risk factor is less well determined.
- The radiation field is unique in terms of both the range of particle types and energies, and additional response characterization of instruments is needed. National regulations in Member States require validation by measurement of dose estimates made using calculation methods. There is a need to define procedures and common approaches to analysis, calibration and traceability of such measurements.
- There has been no complete assessment of the accuracy of measurement or calculation methods; this is desirable.
- There is a need for co-operative procedures between Member States to share information on solar particle events which give increased dose rates at aviation altitudes, to define common procedures for the notification of such events, and to summarize assessments of resultant doses to aircraft crew.

It may also be thought useful to maintain an expert group on aircraft crew dosimetry instrumentation in order to ensure the quality of dosimetry and record keeping generally, but also for possible epidemiological investigations.

ACKNOWLEDGEMENTS

The author is indebted to many colleagues for their published work, their help in collaborative research projects, and discussions on this topic. They are too numerous to mention individually, but I should like to specifically acknowledge Klaus Duftschmid and Keran O'Brien who introduced me to cosmic radiation effects in the atmosphere when I entered that topic in 1994.

REFERENCES

Allkofer, O.K.: Introduction to Cosmic Radiation (München: Karl Thiemig). ISBN 3-521-06098-5. (1975)

Badhwar, G.D.: The Radiation Environment in Low-Earth Orbit, Radiation Research, **148**(5), 3–10 (1997)

Badhwar,G.D., Neill,O.: Galactic cosmic radiation model and its applications, Adv. Space Research, **17**, 7–17 (1996)

Bartlett, D.: Radiation Protection Concepts and Quantities for the Occupational Exposure to Cosmic Radiation, Radiat. Prot. Dosimetry, **86**(4), 263–268 (1999)

Bartlett, D.T.: Radiation protection aspects of the cosmic radiation exposure of aircraft crew, Radiat. Prot. Dosimetry, **109**(4), 349–355 (2004)

Bartlett, D.T., Hager, L.G., Irvine, D., Bagshaw, M.: Measurements on Concorde of the Cosmic Radiation Field at Aviation Altitudes, Radiat. Prot. Dosimetry, **91**(4), 365–376 (2000)

Bartlett, D.T., Hager, L.G., Tanner, R.J., Steele, J.D.: Measurements of the High Energy Neutron Component of Cosmic Radiation Fields in Aircraft Using Etched Track Dosemeters, Radiat. Meas. **33**(3), 243–253 (2001)

Bartlett, D., Beck, P., Heinrich, W., Pelliccioni, M., Roos, H., Schraube, H., Silari, M., Spurny, F., d'Enrico, F.: Investigation of Radiation Doses at Aircraft Altitudes during a Complete Solar Cycle. April 14, 2002 (2002)

Bartlett, D.T., Hager, L.G., Tanner, R.J.: The Determination Using Passive Dosemeters for Aircraft Crew Dose: Results for 1997 ER-2 Flights. Wilson, J. W., Jones, I. W., Maiden, D. L., and Goldhagen, P. Atmospheric Ionizing Radiation (AIR): Analysis, Results and Lessons Learned from the June 1997 ER-2 Campaign, 321–332 (2003) Langley, VA, USA, NASA. Ref Type: Report

Beaujean, R., Burmeister, S., Petersen, F., Reitz, G.: Radiation exposure measurement onboard civil aircraft, Radiat. Prot. Dosimetry, **116**(1–4), 312–315 (2005)

Beck, P. et al.: Validation of Modelling the Radiation Exposure due to Solar Particle Events in Aircraft Altitudes. IRPA Proceedings . 2006. Second European IRPA Congress on Radiation Protection. 15–19 May 2006 Paris, France. Ref Type: In Press

Beck, P., Ambrosi, P., Schrewe, U., O'Brien, K.: ACREM, Aircrew Radiation Exposure Monitoring. OEFZS-G-0008. 1999a. ARC Seibersdorf research. OEFZS Report. Ref Type: Report

Beck, P., Bartlett, D., O'Brien, K., Schrewe, U.J.: In-flight Validation and Routine Measurements, Radiat. Prot. Dosimetry, **86**(4), 303–308 (1999b)

Beck, P., Latocha, M., Rollet, S., Stehno, G.: TEPC reference measurements at aircraft altitudes during a solar storm, Adv. Space Res. **36**(9), 1627–1633 (2005)

Beck, P., Bartlett, D., Lindborg, L., McAulay, I., Schnuer, K., Schraube, H., Spurny, F.: AIRCRAFT CREW RADIATION WORKPLACES: COMPARISON OF MEASURED AND CALCULATED AMBIENT DOSE EQUIVALENT RATE DATA USING THE EURADOS IN-FLIGHT RADIATION DATA BASE, Radiat. Prot. Dosimetry ncl029 (2006).

Bottollier-Depois, J.F. et al.: Exposure of aircraft crew to cosmic radiation: on-board intercomparison of various dosemeters, Radiat. Prot. Dosimetry, **110**(1–4), 411–415 (2004)

Dyer, C. et al.: Implications for space radiation environment models from CREAM & CREDO measurements over half a solar cycle, Radiat. Meas. **30**, 569–578 (1999)

Dyer, C.S., Lei, F., Clucas, S., Smart, D.F., Shea, M.A.: Calculations and observations of solar particle enhancements to the radiation environment at aircraft altitudes, Adv. Space Res. **32**, 81–93 (2003)

Dyer, C., Lei, F., Hands, A., Clucas, S., Jones, B.: Measurements of the atmospheric radiation environment from CREAM and comparisons with models for quiet time and solar particle events, IEEE Trans. Nucl. Sci. **52**(6), 2326–2331 (2005)

European Community: Council Directive 96/29/EURATOM of 13 May 1996 Laying Down the Basic Safety Standards for Protection of the Health of Workers and the General Public Against the Dangers Arising from Ionising Radiation. Official Journal of the European Communities, **39**(L159) (1996)

Green, A.R., McCall, M.J., Lewis, B.J., Bennett, L.G.I.: Cosmic Radiation Exposure of Aircrew Transport. RMC-CCE-NSE-00-02, 1–2 (2000) Ref Type: Report

Heinrich, W., Roesler, S., Schraube,H.: Physics of Cosmic Radiation Fields, Radiat. Prot. Dosimetry, **86**(4), 253–258 (1999)

Hess,V.F.: Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten, Physikalische Zeitschrift 13. Jahrgang 21 (22), 1084–1091 (1912)

Hume, C.: *Cosmic Radiation – An Aircraft Manufacturer's View*, Radiat. Prot. Dosimetry, **86**(4), 335–336 (1999)

ICRP *Recommendations of the International Commission on Radiological Protection* (Oxford: Pergamon Press. 1991)

Joint Aviation Authorities (JAR). Joint Aviation Requirements JAR-OPS 1 Commercial Air Transportation (Aeroplanes) Subpart D – Operational Procedure JAR-OPS-1.390 Cosmic radiation. JAR-OPS-1.390 (2001) Ref Type: Report

Kelly, M., Menzel, H.G., Schnuer, K., Ryan,T.: Editorial – Cosmic Radiation and Aircrew Exposure, Radiat. Prot. Dosimetry, **86**(4), 245 (1999)

Kyllönen, J.-E., Lindborg, L., Samuelson, G.: Cosmic Radiation Measurements On-board Aircraft with the Variance Method, Radiat. Prot. Dosimetry, **93**(3), 197–205 (2001)

Lantos, P.: Radiation doses potentially received on-board aeroplane during recent solar particle events. Radiat.Prot.Dosim. (2006). Ref Type: In Press

Lantos, P., Fuller, N.: History of the solar flare radiation doses on-board aeroplanes using semi-empirical model and Concorde measurements, Radiat. Prot. Dosim. **104**(3), 199–210 (2003)

Latocha, M., Beck, P., Rollet, S.: Cosmic Radiation Exposure at Aircraft Crew Workplaces. Conference Proceedings. 2006. Second European IRPA Congress on Radiation Protection. 15–19 May 2006 Paris, France. 15-5-0006. Ref Type: In Press

Lei, F., Clucas, S., Dyer, C., Truscott, P.: An atmospheric radiation model based on response matrices generated by detailed Monte Carlo simulations of cosmic ray interactions, IEEE Trans. Nucl. Sci. **51**(6), 3442–3451 (2004)

Lewis, B.J., McCall, M., Green, A.R., Bennett, L.G.I., Pierre, M., Schrewe, U.J., O'Brien, K., Ferrari,A.: Aircrew Exposure from Cosmic Radiation on Commercial Airline Routes, Radiat. Prot. Dosimetry, **93**(4), 293–314 (2001)

Lindborg, L., Bartlett, D., Beck, P., McAulay, I., Schnuer, K., Schraube, G., Spurny, F.: [140], 1–271. 2004a. Cosmic Radiation Exposure of Aircraft Crew: Compilation of Measured and Calculated Data. Luxembourg, Belgium, European Commission, Office for Official Publication of the European Communities. Radiation Protection. Ref Type: Report

Lindborg, L., Bartlett, D., Beck, P., McAulay, I., Schnuer, K., Schraube, H., Spurny, F.: Cosmic Radiation Exposure of Aircraft Crew: Compilation of Measured and Calculated Data, Radiat. Prot. Dosimetry, **110**(1–4), 417–422 (2004b)

McAulay, I.R.: Round Table Discussion – Radiation Exposure of Civil Aircrew, Radiat. Prot. Dosimetry, **48**(1), 135–140 (1993)

Neill, O.: Badhwar–O'Neill galactic cosmic ray model update based on advanced composition explorer (ACE) energy spectra from 1997 to present. Advances in Space Research (2005) Ref Type: In Press

O'Sullivan, D., Bartlett, D., Beck, P., Bottollier-Depois, J.-F., Schrewe, U., Lindborg, L., Tommasino, L., Zhou, D.: Recent Studies on the Exposure of Aircrew to Cosmic and Solar Radiation, Radiat. Prot. Dosimetry, **100**(1–4), 495–498 (2002)

Regulla, D., David, J.: Measurements of Cosmic Radiation on Board Lufthansa Aircraft on the Major Intercontinental Flight Routes, Radiat. Prot. Dosimetry, **48**(1), 65–72 (1993)

Regulla, D., Schraube, H.: Radiation exposure of aircrews in civil aviation. Köln, October 23, 1996, 375–380 (1996)

Romero, A.M., Saez-Vergara, J.C., Rodriguez, R., Dominguez-Mompell, R.: Study of the ratio of non-neutron to neutron dose components of cosmic radiation at typical commercial flight altitudes, Radiat. Prot. Dosimetry, **110**(1–4), 357–362 (2004)

Saez Vergara, J.C., Romero, A.M., Rodriguez, R., Muñiz, J.L., Dominguez-Mompell, R., Merelo, F., Ortiz, P., Ruano, J.: Monitoring of the Cosmic Radiation at IBERIA Commercial Flights: One Year Experience of In-flight Measurements. Proceedings of the Radiation Protection and Shielding Division Topical Meeting, (2002)

Schraube, H., Leuthold, G.P., Heinrich, W., Roesler, S., Mares, V., Schraube, G.: European Program Package for the Calculation of Aviation Route Doses. -32 (2002) Neuherberg, Germany, GSF. Ref Type: Report

Schrewe, U.J.: Air Crew Radiation Exposure Monitoring – Results from the in-flight Measurement Program of the PTB: Summary of the Radiation Monitoring Data. PTB-6.31-99-1, -C5. 1999. Braunschweig, Physikalisch Technische Bundesanstalt (PTB). Ref Type: Report

Schrewe, U.J., Newhauser, W.D., Brede, H.J., DeLuca, P.M.Jr.: Experimental kerma coefficients and dose distributions of C, N, O, Mg, Al, Si, Fe, Zr, A-150 plastic, Al2O3, AlN, SiO2 and ZrO2 for neutron energies up to 66 MeV, Physics in Biology and Medicine, **45**(3), 651–683 (2000)

Shea, M.A., Smart, D.F., McCall, M.: A Five Degree by Fifteen Degree World Grid of Trajectory-Determined Vertical Cutoff Rigidities. Canadian Journal of Physics, [46], 1098–1101 (1968)

Shea, M.A., Smart, D.F., Gentile, L.C.: Estimating Cosmic Ray Vertical Cutoff Rigidities as a Function of the McIlwain L-parameter for Different Epochs of the Geomagnetic Field, Phys. Earth Planet. Inter. **48**, 200–205 (1987)

Shea, M.A., Smart, D.F.: Vertical Cutoff Rigidities for Cosmic Ray Stations Since 1955. 2001, 4063–4066 (2001)

Spurný, F., Datchev, Ts.: Measurements in an Aircraft during an Intense Solar Flare, Ground Level Event 60, on the 15th April 2001., Radiat. Prot. Dosim. **95**, 273–275 (2001)

Spurný, F., Dachev, T.: Aircrew onboard Dosimetry with a Semiconductor Spectrometer, Radiat. Prot. Dosim, **100**, 525–528 (2002)

Spurný, F., Dachev, T.: Long-term monitoring of the onboard aircraft exposure level with Si-diode based spectrometer, Adv. Space Res. **32**, 53–58 (2003)

Spurný, F., Kudela, K., Dachev, C.: Strong Forbush decrease registered in onboard aircraft dose, Adv. Space Res. **36(9)**, 1634–1637 (2005)

Spurný,F., Malušek,A., Kováø,I.: Individual dosimetry of Czech company aircrew 1998–2000. Inter.Conf. on Occupational Radiation Protection: Protecting of Workers against Exposure to Ionising Radiation, August 26, 2002, 397–400 (2002)

Taylor, G.C., Bentley, R.D., Conroy, T.J., Hunter, R., Jones, J.B.L., Pond, A., Thomas, D.J.: The Evaluation and Use of a Portable TEPC System for Measuring In-flight Exposure to Cosmic Radiation, Radiat. Prot. Dosimetry, **99**(1–4), 435–438 (2002)

Tommasino, L.: In-Flight Measurements of Radiation Fields and Doses, Radiat. Prot. Dosimetry, **86**(4), 297–301 (1999)

United Nations: Sources and Effects of Ionizing Radiation. United Nations Scientific Committee on the Effects of Atomic Radiation, UNSCEAR 2000 Report to the General Assembly, with scientific annexes. No. E.00.IX.3 (Volume I: Sources, No. E.00.IX.4 (VolumeII: Effects), 1–1220 (2000) New York, United Nations sales publications. Ref Type: Report

University of Delaware. The Bartol Research Institute, Neutron Monitor Programme. http://ulysses.sr.unh.edu/NeutronMonitor/Misc/neutron2.html (2006) Ref Type: Electronic Citation

van Dijk, W.E.: Dose assessment of aircraft crew in the Netherlands, Radiat. Prot. Dosimetry **106**(1), 25–31 (2003)

Vergara, J.C.S., Gutierrez, A.M.R., Jimenez, R.R., Roman, R.D.-M.: In-flight measured and predicted ambient dose equivalent and latitude differences on effective dose estimates, Radiat. Prot. Dosimetry, **110**(1–4), 363–370 (2004)

Wiegel, B., Alevra, A.V., Matzke, M., Schrewe, U.J., Wittstock, J.: Spectrometry with the PTB Neutron Multisphere Spectrometer (NEMUS) at Flight Altitudes and at Ground Level, Nuclear Instruments and Methods in Physics Research, **A476**, 52–57 (2002)

CHAPTER 5.0

THE MAGNETIC ENVIRONMENT – GIC AND OTHER GROUND EFFECTS

JURGEN WATERMANN

Geomagnetism and Space Physics Programme, Danish Meteorological Institute, Copenhagen, Denmark

The term 'Space Weather' is nowadays frequently used by the informed society, and often with more or less different meanings behind it, among which we find the following. It is used to characterize in physical terms the state and dynamics of the Earth's space environment (geospace) with respect to – or in response to – solar activity. In consequence, it is sometimes considered to be part of solar-terrestrial physics. But it is, strictly speaking, not limited to the coupling between the Sun and the planet Earth, although we choose to use it here in this sense. It has been associated with understanding and describing the direct and indirect effects of solar activity on humans and technological systems wherever in the solar system they may happen to occur. But just like terrestrial weather exists with or without the presence of human beings, with or without technological systems in operation, exists space weather independent of the presence of human life and technology.

Part of the sometimes ambiguous use of the term 'Space Weather' stems from the fact that the concept of Space Weather rests on two main columns, research and applications, which are supporting each other to their mutual benefit. Research largely overlaps with the traditional category of solar-terrestrial physics, and applications means turning research results into technical products, operational procedures, information and advisory services and the like – products which eventually serve society. The logical chain starts with scientific basic research on solar-terrestrial relations. The results are used by service developers and providers who turn them into products which are offered to or requested by service users. The requirements with which research and applications are confronted are often very different. Research aims at understanding the physics behind the phenomena and building physically accurate models. It builds on observations not necessarily taken in real-time. Application is closely related to and almost always depending on

J. Lilensten (ed.), Space Weather, 269–275.

the availability of real-time or near-real-time data in order to offer timely results which enable concerned individuals or authorities to assess the potential effect of the prevailing space weather conditions, choose suitable products or strategies and launch appropriate operational procedures if deemed necessary.

The majority of those concerned with space weather associates 'strategies' and 'procedures' with 'mitigation' or 'protection', but this is too limited a view. Space weather application means being aware of the consequences of space weather events whether benign or malignant. It includes, for instance, the positive experience of watching the launch of coronal mass ejections and enjoying impressive auroral displays.

In this chapter we wish to consider the effect of solar activity on technological systems which exist on planet Earth. We focus our view on the geomagnetic environment and – according to the title of this chapter – specifically on geomagnetically induced currents (GIC) and other ground effects. More precisely, we are concerned with one particular type of space weather effects, namely geomagnetic effects which are the direct result of geomagnetic field perturbations originating in ionospheric and magnetospheric electric current systems which are observed on the ground and which have a potential impact on the quality of our life, the performance of our technology and the state of our economy. A direct impact of the magnetic effects of adverse space weather conditions on human health has so far not been confirmed, unlike, for instance, the effect of enhanced radiation associated with a solar storm. Therefore we are in this chapter concerned with geomagnetic effects on ground-based technological systems only. Here we can follow different categorization schemes. The two main schemes are based on either different physical parameters or different techno-economical parameters.

The two physically oriented categories of geomagnetic effects on technological systems concern

- systems and operations which are sensitive to the magnetic field amplitude, ΔB. They include magnetic anomaly surveys (e.g., aeromagnetic surveys) and directional wellbore drilling.
- systems and operations which are sensitive to the magnetic field time derivative, $\partial B/\partial t$. They include electric power transmission grids, oil and gas pipelines and long-distance communication cables.

There is a significant difference between these two parameters. ΔB often exhibits large-scale characteristics reflecting the large-scale pattern of ionospheric and magnetospheric currents. Among them are the sub-auroral signatures of substorm enhanced electrojets (a negative excursion of the magnetic northward component, known as 'bay'), the evolution of a magnetic storm (a deep decrease of the magnetic northward component indicates the 'storm main phase') and the southward resp. northward deviation in the morning resp. afternoon auroral electrojet zone which makes ΔB qualitatively predictable (e.g., Kelley, 1989). In contrast, the orientation of $\partial B/\partial t$ behaves much more randomly and is therefore much more difficult to predict (Pulkkinen et al., 2005). The latitudinal dependences of the peak amplitudes of ΔB and $\partial B/\partial t$ are also different. ΔB peaks in the nominal auroral oval

during quiet and moderately disturbed times, and its amplitude peak tends to move further equatorward during severe magnetic storms. This is not the case with $\partial B/\partial t$, its peak amplitude is found at nominal auroral latitudes also during severe storms even though ΔB peaks at subauroral latitudes (Watermann, private communication).

The two techno-economically oriented categories of geomagnetic effects on technological systems concern

- systems which may suffer equipment damage as a result of enhanced geomagnetic activity. They include electric power transmission grids and gas and oil pipelines where the damage in the former case can be immediate and in the latter cumulative and long-term.
- systems which are not directly damaged by large geomagnetic perturbations but whose operational performance degrades during geomagnetically active times. They include magnetic anomaly surveys, directional wellbore drilling and communication via long-distance cables.

The importance of observations of geomagnetic field variations for the physical understanding of solar-terrestrial coupling was recognized long time ago. It triggered attempts to categorize geomagnetic field variations which led to the development of various geomagnetic indices. Selection and definition of geomagnetic indices reflect the state of the physical understanding of the external geomagnetic field and the state of the technical development of measurement devices and data processing methods. Consequently geomagnetic indices are under continuous evolution. Originally devised as tools to support research, the value of geomagnetic indices for space weather applications was later discovered and is nowadays controversely discussed. The article by Menvielle and Marchaudon gives an overview over the present system of formally accepted indices, discusses the value of indices for space weather monitoring by highlighting their role in a comprehensive case study of the Sun–Earth chain of events during a major space storm, and points out the shortcomings and limitations of some of the presently used indices. The paper ends with realistic suggestions for developing indices which will be better adapted to specific space weather related needs.

Let us turn to the first physics-based category concerning the role of ΔB. Technological systems and operations which are negatively affected by the amplitude of magnetic field variations are usually systems which depend on the availability of reliable reference levels for the magnetic field. A magnetic anomaly survey which seeks to map the spatial distribution of the magnetostatic crustal field needs to avoid being distorted by temporal variations of the geomagnetic field resulting from external sources. If a quasi-permanent magnetic reference station is located in close vicinity of the survey area its data can be used to distinguish between spatial and temporal variations, and the survey measurements can be corrected for temporal variations thus retaining solely spatial variations. If the fixed reference magnetometer is too far away from the actual survey site (which is usually the case in aeromagnetic surveys at high latitudes where external magnetic field perturbations can vary substantially over short distances) a reliable correction is not possible.

Survey teams have so far tended to cancel a survey flight if magnetic activity is high or else discard survey data taken during magnetically active times and repeat the measurements in a quiet period. However, first steps have been initiated to assess the potential for developing methods based on a chain or an array of magnetometer stations for computing the temporal magnetic variation at a virtual reference site, in fact the survey site (Watermann et al. 2005).

Another example from the same physical category addresses the problem of accurate directional wellbore drilling. In early days boreholes were drilled nominally vertical. But methods became more sophisticated, and nowadays directional drilling is quite common. Today wells can be drilled in practically any direction, limited only by certain geological structures which require special attention (e.g., faults and soft sediments), and by the drilling equipment which is subject to certain geometrical constraints (for instance, sharp borehole bends are not permitted). Since directional drilling is mostly controlled by borehole magnetometers its accuracy is very sensitive to magnetic contamination by the drill gear and to variations of the external geomagnetic field. In the past wellbore survey managers used to monitor the temporal variation of the geomagnetic field at a remote reference station and decided a posteriori whether the borehole data could be accepted and retained or whether they had to be discarded and the measurements possibly be repeated (in case of too many discarded data points).

In recent years the method has changed. Survey managers try to incorporate real-time geomagnetic field variation data into the control parameters of ongoing wellbore operations (Reay et al. 2005). At least at subauroral and mid-latitudes it is possible to construct a virtual magnetic observatory at the wellbore location from measurements taken at real observatories in the same region. The article by Bowe and McCulloch illustrates from a user's point of view the current status of magnetically controlled directional drilling and the immense value of real-time geomagnetic observations for the current level of the technology.

These two examples from the first physics category fall into the second techno-economical category, namely technological systems for which high geomagnetic activity is a disturbing nuisance but does not result in equipment damage.

The second physics category concerns the role of $\partial B/\partial t$. The effects are a manifestation of basic laws of electromagnetic theory, namely Maxwell's equations combined with Ohm's law. The central theme of this category are the electric field and current induced by a time-varying magnetic field (which led to the notion of geomagnetically induced currents – GIC). The physical basis of the induction effect can be understood in the following way. Let us consider a plane harmonic electromagnetic wave with angular frequency ω propagating from the ionosphere down to the Earth's surface. We assume for simplicity of demonstration that the relative permeability of the ground is unity ($\mu = \mu_0$) and the electric conductivity of the substratum is uniform and finite ($0 < \sigma = \text{constant} < \infty$). Space charges cannot build up in the ground ($\rho \approx 0$) so that the permittivity, $\varepsilon = \varepsilon_r \cdot \varepsilon_0$, plays in our case no role for $\nabla \cdot (\varepsilon E)$. Maxwell's equations and Ohm's law for the electric field E, the magnetic induction B, and the current density j, read

$$(1) \qquad \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad \nabla \times \vec{B} = \mu_0 \left(\vec{j} + \frac{\partial (\varepsilon \vec{E})}{\partial t} \right)$$

$$\nabla \cdot (\varepsilon \vec{E}) = \rho \qquad \nabla \cdot \vec{B} = 0$$

$$\vec{j} = \sigma \vec{E}$$

We restrict our considerations to waves in the ULF/ELF bands which has the consequence that displacement currents are negligible compared to Ohm's currents when assuming a quantitatively realistic ground conductivity distribution. The overall magnetic field (which is the superposition of both, the primary and the induced components) and the induced electric field can then be described by symmetric diffusion equations.

$$(2) \qquad \nabla^2 \vec{E} = \mu_0 \sigma \frac{\partial \vec{E}}{\partial t} \qquad \nabla^2 \vec{B} = \mu_0 \sigma \frac{\partial \vec{B}}{\partial t}$$

It becomes immediately clear that a non-zero time derivative and a non-zero conductivity are necessary conditions for GIC to occur. The fact that the induction process is governed by a diffusion equation demonstrates that the instantaneous value of the induced electric field is determined by the history of the process, i.e by the instantaneous and past values of the full magnetic field.

$$(3) \qquad E_x(t) = \frac{1}{\sqrt{\pi \mu_0 \sigma}} \int_0^\infty \frac{\partial B_y(t-\tau)}{\partial t} \frac{\mathrm{d}\tau}{\sqrt{\tau}}$$

Solving the diffusion equation shows that the induced electric field (and current) in an arbitrarily chosen x-direction are proportional to a changing magnetic field in the y-direction (which is rotated 90° clockwise from the x-direction). Note that the induced electric field decreases and the induced current increases with increasing electric conductivity.

$$(4) \qquad E_x^2 = \frac{\omega}{\mu_0 \sigma} B_y^2 \qquad j_x^2 = \frac{\omega \sigma}{\mu_0} B_y^2$$

Space weather effects on technological systems resulting from large time derivatives of the magnetic field variation have been observed since long. The common prerequisite for such effects to become noticed is the existence of long-distance lines. Observations of anomalous behaviour of long telegraph lines are probably the earliest examples. Since communication lines can be very long, particularly those crossing oceans, GIC have always been experienced in long-haul wire communication. They have contributed to either impeding or facilitating telegraph transmission (depending on the direction of the induced electric current) although undesired GIC are not known to directly damage telegraphic equipment. Since about 20 years electric trans-oceanic cables have gradually been replaced by fiber optic cables. But that has not eliminated the effect of GIC since the signals propagating

in fiber optics need to be amplified at periodic spacing, and the repeater elements are powered via electric wires running parallel to the optical fibres. A presentation by Lanzerotti given in this session discussed relevant experiences on long-haul communication lines in more detail. The presentation does not form part of this volume, but some of the material was published earlier by Lanzerotti (2001).

With the implementation of very long electric power transmission lines at high latitudes GIC effects became a concern for power plant and network operators. A number of cases have actually attained wide attention where the electric power supply ceased in the wake of geomagnetic disturbances affecting power networks, see Boteler et al. (1998) for a list of events prior to the most recent decade. Space weather effects on power transmission grids thus belong to the first techno-economical category which deals with technological systems which may suffer severe equipment damage. For that reason much effort has been and is being spent on (i) understanding quantitatively the physical processes that eventually lead to large GIC, (ii) developing algorithms to forecast GIC, and (iii) finding technical solutions which help to avoid damaging consequences of adverse space weather conditions. It should be kept in mind that the advanced state of the compound inter-European power supply network bears the consequence that GIC effects suffered in a high-latitude country can eventually affect the stability of the power supply in countries at lower latitudes.

The article by Pulkkinen discusses the spatio-temporal structure of $\partial B/\partial t$ (which is closely related to GIC and thus can serve as a proxy for GIC), presents a stastistical analysis of magnetic field fluctuations for a selection of events, discusses the implications of the characteristic results of the analysis and comments on the possibility to develop strategies for GIC prediction. In this context one must keep in mind that GIC in power transmission systems depend in a complex way on the history of ΔB, the ground conductivity distribution and the structure of the power line network so that $\partial B/\partial t$ can serve as a proxy for GIC only to a certain extent (e.g., Pulkkinen 2003).

The discussion of the problem of GIC effects on power transmission lines continues with an article by Elovaara (who represents a power network operator) for whom GIC are nearly an everyday topic. His article gives detailed information on the technical issues associated with GIC in power transmission lines and describes mitigation strategies developed in Finland and equipment design solutions adapted to the Finnish power grid.

GIC can have a long-term effects on the integrity of pipelines. They change the nominal soil-pipe electric potential and counteract the effect of cathodic protection. Cathodic protecion is used to keep $(OH)^-$ ions away from the pipe steel. It is most efficient if the pipe potential is maintained at about 850–1150 mV negative with respect to the surrounding soil. If the potential changes through the effect of an induced electric field an undesired GIC can flow if the pipeline is not perfectly insulated against the ground. The pipeline material can be oxidized and the pipeline degrades faster. An earlier than planned refurbishment of the pipeline system will become necessary. The build-up of GIC in power lines and in pipelines follows

the same physical principle. The main difference between the two systems is the grounding. While power lines are grounded only at transformer stations the pipelines are grounded all along their way, although via high resistance coating material. Boteler (2000) quotes a conductance of 0.057 S/km for a sample pipeline in North America.

The collection of presentations given at the ESWW-2 and partially reproduced here provides convincing evidence that GIC are space weather effects which are well-known to researchers and equipment operators. But that does not mean that they are under full control. And it is equally clear that GIC are not the only space weather effects in which the magnetic field variation plays a dominant role. Aeromagnetic anomaly surveys and directional wellbore drilling are other examples where magnetic ground effects of space weather events play an important role. Geomagnetic effects of space weather events will therefore remain to be a topic of attraction for researchers and operators, also in the years to come.

REFERENCES

Boteler, D.H., Pirjola, R.P., Nevanlinna, H.: The effects of geomagnetic disturbances on electrical systems at the earth's surface. Adv Space Res. 22, 17–27 (1998)

Boteler, D.: Geomagnetic effects on the pipe-to-soil potentials of a continental pipeline. Adv Space Res. 26, 15–20 (2000)

Kelley, M.: The Earth's ionosphere – plasma physics and electrodynamics. Academic Press, Inc., San Diego (1989)

Lanzerotti, L.J.: Space weather effects on communications. In: Daglis, I.A. (ed) Space storms and space weather hazards. Kluwer Academic, Dordrecht/Boston/London, pp. 313–334 (2001)

Pulkkinen, A.: Geomagnetic induction during highly disturbed space weather conditions: Studies of ground effects, Ph.D. thesis, Finnish Meterological Institute (2003)

Pulkkinen, A., Lindahl, S., Viljanen, A., Pirjola, P.: Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system, Space Weather 3:doi:10.1029/2004SW000123 (2005)

Reay, S.J., Allen, W., Baillie, O., Bowe, J., Clarke, E., Lesur, V., Macmillan, S.: Space weather effects on drilling accuracy in the North Sea, Ann Geophys. 23:3081–3088 (2005)

Watermann, J., Rasmussen, O., Stauning, P., Gleisner, H.: Temporal versus spatial geomagnetic variations along the west coast of Greenland. Adv. Space Res. 37, 1163–1168 (2006)

CHAPTER 5.1

GEOMAGNETIC INDICES IN SOLAR-TERRESTRIAL PHYSICS AND SPACE WEATHER

M. MENVIELLE[1] AND A. MARCHAUDON[2]

[1] *CETP/CNRS/IPSL, Centre d'étude des Environnements Terrestre et Planétaires, Saint Maur, France (michel.menvielle@cetp.ipsl.fr) et Département des Sciences de la Terre, Université Paris Sud XI, Orsay, France*

[2] *LPCE/CNRS, Laboratoire de Physique et Chimie de l'Environnement, Orléans, France*

Abstract: Geomagnetic indices play a significant role in describing the magnetic configuration of the Earth's magnetosphere. In the past 15 years, they have become a key parameter in Space Weather research, being commonly used to detect and describe Space Weather events. Research is currently being carried out into using them for forecasting. The objective of this paper is to contribute to a better understanding of the usefulness, potential and limitations of geomagnetic indices in Space Weather research and applications

INTRODUCTION

The Sun emits a permanent but variable supersonic flow of hot plasma: the solar wind. The magnetic field of the Sun is dragged by the solar wind into the interplanetary medium, thus giving rise to the Interplanetary Magnetic Field (IMF). Close to the Earth, the solar wind is slowed down through the bow shock and streams around the Earth's magnetic obstacle. This results in compressing the Earth's magnetic field in the dayside, and in stretching it in a long tail in the nightside, giving rise to the magnetosphere cavity.

Merging between anti-parallel IMF and dayside magnetospheric field is the dominant mechanism by which energy, momentum and plasma are transferred from the solar wind into the Earth's magnetosphere. Other mechanisms such as diffusion, viscous interaction or impulsive plasma penetration contribute also to this transfer. During merging process, newly reconnected field lines are opened and are eventually dragged anti-sunward by the solar wind flow. The solar wind plasma entering the magnetosphere flows along magnetic field lines and is simultaneously

J. Lilensten (ed.), Space Weather, 277–288.

transported anti-sunward by the magnetic field lines convection. Part of this plasma is stored inside the magnetotail in the nightside. On both sides of the magnetotail equatorial plane, the stretched opened field lines are anti-parallel and can reconnect mainly during violent episodes called substorms. The plasma stored in the tail is then released in the nightside ionosphere and the new closed field lines created by reconnection are dragged sunward by the magnetic tension. The dynamics of the Earth's magnetosphere is variable and depends upon IMF orientation and solar wind properties. The magnetospheric convection maps along magnetic field lines in the high-latitude ionosphere. The ionospheric convection gives then a condensed view of the general dynamics of the magnetosphere and shows several cells whose number and shape vary with IMF orientation (see Cowley, 1982, for a complete review).

The plasma dynamics in the magnetosphere is associated with current flows, which produce magnetic signatures at the Earth's surface, the so-called irregular transient variations of the magnetic field (see, e.g. Menvielle and Berthelier (1991) and references therein for a more detailed description). There are several sources of currents in the magnetosphere (see Fig. 1):

- the magnetopause current is essentially due to the interaction between the solar wind and magnetosphere plasmas. This current flows eastward on the dayside magnetopause, around the polar cusps (neutral points of the terrestrial magnetic field), and westward on the nightside magnetopause where it is called the tail

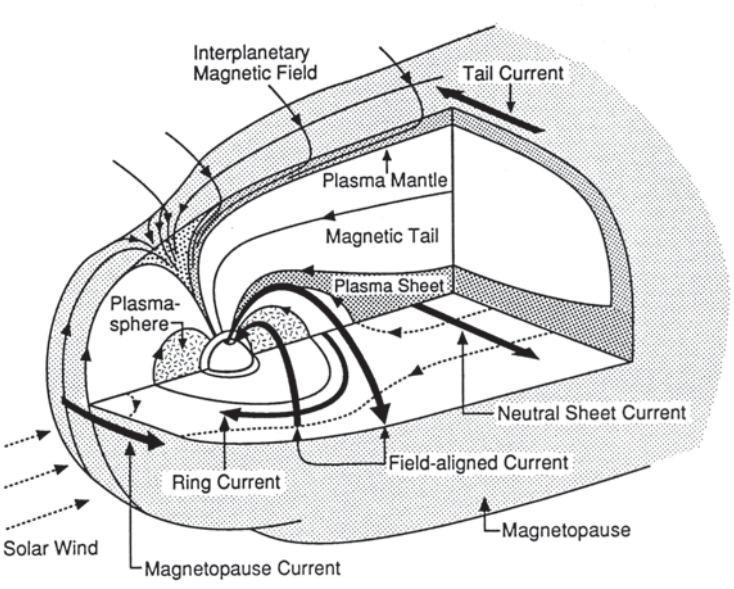


Figure 1. Sketch of the magnetosphere. The arrows indicate the current flows associated to the plasma dynamics in the magnetosphere (see Section 1 for further details; from Kivelson and Russel, 1995)

current. This circuit is closed by the cross-tail current sheet flowing eastward in the magnetic equatorial plane of the magnetotail;

- the ring current is the magnetic signature of the inner magnetosphere radiation belts. It flows westward around the Earth and is mainly due to pressure gradients. Plasma diffusion along the magnetic field lines and plasma injections result in local enhancement of the ring current intensity;
- the field-aligned currents (FACs) flow along highly conductive field lines and link the magnetosphere and ionosphere current systems. The large-scale distribution of FACs consists of two concentric zones encircling the magnetic pole. Currents flow in opposite direction on each side of the noon-midnight plane and also between the two concurrent currents. The low-latitude currents called Region-2 are mainly due to pressure gradients in the inner magnetosphere and to the divergence of the ring current, while the high-latitude currents called Region-1 flow at the interface between open and closed field lines and result from direct or indirect interaction between the solar wind and the magnetosphere boundary layers. The noon and midnight systems of currents are more complicated: their patterns depend upon IMF conditions for the former, and upon substorm activity for the latter;
- the currents flowing in the high-latitude ionosphere are associated with ionosphere convection. Due to medium anisotropy, two types of currents co-exist. The Pedersen currents flow parallel to the convection electric field, i.e. duskward through the polar cap and dawnward on the auroral zones: they close the Region-1 and Region-2 of field-aligned currents. The Hall currents flow perpendicular to the convection electric field, i.e. anti-sunward along the auroral zones: they are called auroral electrojets, and are strongly enhanced during substorms.

The transient magnetic variations observed at geomagnetic observatories installed at the Earth' surface can therefore be considered as the output of a complex highly non-linear magnetosphere-ionosphere filter with the interplanetary conditions at the Earth's location as inputs. Monitoring the magnetic transient variations at the Earth' surface then provides information on the magnetosphere and ionosphere response to its forcing by the solar wind and IMF.

The present geomagnetic indices are briefly described in Section 2, paying particular attention to their physical meaning and limitations. In Section 3, the contribution of geomagnetic indices to Solar-Terrestrial physics is illustrated using a case study. Finally, suggestions of new indices more appropriate to describe the activity state of the magnetosphere during storm events, and associated Space Weather effects are presented in section 4.

PRESENT GEOMAGNETIC INDICES

Continuous recordings of the Earth magnetic field variations at permanent geomagnetic observatories started during the second half of the nineteenth century. It appeared very rapidly that they are the signature of a complex system. Summarizing quantities that extract pertinent, reliable and concentrated information from the

observations were introduced. The first geomagnetic index has thus been proposed in 1905 in order to distinguish between the days of a single month, so that a proper choice of the five quietest days per month might be made. It becomes rapidly obvious that the daily basis is not well suited to derive geomagnetic indices. In what follows, we briefly review the geomagnetic indices of common use in Solar-Terrestrial physics and Space Weather studies. For further details, the reader is referred to Menvielle and Berthelier, (1991), Rangarajan (1989), or to the Introduction of IAGA Bulletins 32 series.

The main planetary geomagnetic indices are the Kp index introduced by Bartels in 1949, and the am and aa indices introduced by Mayaud in 1968 and in 1973 respectively (Mayaud, 1968, 1973). These indices are called K-derived planetary geomagnetic indices, because they are derived from K indices measured at network of stations. Introduced by Bartels, (1939), the K index is a pure code, which marks the class in which falls the range of the horizontal magnetic components variations during a 3-hour UT interval. Because it is more convenient to use quantities expressed in physical units, the K-index can be converted back to equivalent amplitude, the aK-index, using mid-class values. Equivalent amplitudes aK are proxies of the energy density embedded in the irregular magnetic variations and the relationship between aK and the magnetic energy density is a statistical one (Menvielle, 1979). The three-hour time resolution of the K-derived indices limits their interest for detailed studies of the magnetosphere response to the solar wind forcing. The Kp index is the average of "standardized K-indices" derived from observed K at each subauroral latitude station of the Kp network (Fig. 2a), by means of conversion tables that aim at eliminating LT and seasonal features. On the contrary, the am index is a weighted average of the aK equivalent amplitudes measured at each subauroral latitude station of the am network (Fig. 2b). The aa index is the weighted average of the aK equivalent amplitudes measured at two antipodal observatories one in Western Europe, the other in Eastern Australia: it provides a simple means of monitoring planetary geomagnetic activity continuously back to 1868. The am, aa, and Kp indices are statistically related to the overall magnetosphere energy status. Consequently, they should provide close information. However, the am index gives better estimation of the magnetosphere state, because of the historical context. Designed at the end of the forties (i.e. during the cold war, and before the international Geophysical Year, IGY), the Kp network (Fig. 2a) is heavily weighted towards Europe and Northern America while on the contrary, the am network (Fig. 2b) takes advantage of the better situation after the International Geophysical Year (IGY) and the end of the cold war, and shows a strong improvement in the observatory distribution. Then to conclude, as clearly stated by Menvielle and Berthelier (1991): "Kp is generally used from force of habit. (...) In statistical studies am is obviously the best choice. The use of aa is justified when one needs very long series of indices, or on the other hand when a quick available indication of geomagnetic activity is desirable."

Other indices aim at characterizing only part of the magnetic activity of the magnetosphere. The Dst index measures the variations in the geomagnetic North component H

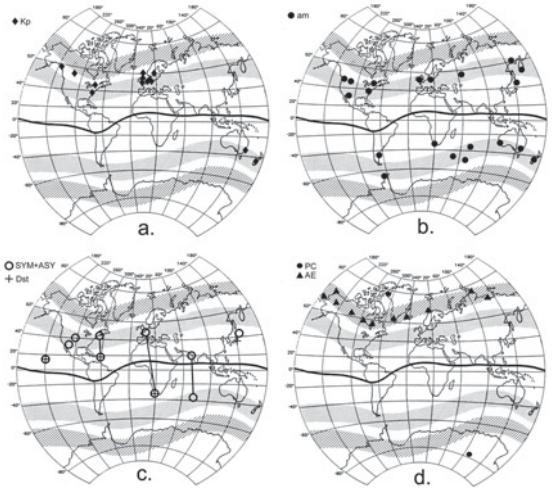


Figure 2. Geographical world maps on which are indicated the positions of stations belonging to the different networks used in deriving geomagnetic indices, as for 2005. A solid line indicates the position of the dip equator. The average extension of the auroral zone is sketched by the hatched area, that of the subauroral region by the shaded area (after Berthelier, 1993).
Panel a: the Kp network. Panel b: the am network. Panel c: the Dst, SYM and ASY networks Dst: Crosses; SYM and ASY: Circles). The SYM and ASY network stations connected by a solid line are replaced by each other in the index computation, depending on the availability and the condition of the data of the month. Panel d: the AE network. The grey triangle corresponds to the station of Cape Wellen (Russia), which is closed since 1996

at four low latitude observatories (Fig. 2c) (Sugiura, 1965; Sugiura and Kamei, 1991); it is derived on a one-hour basis. It aims at monitoring the axi-symmetric part of the magnetosphere currents. It is mainly sensitive to the ring current and to the magnetopause Chapman-Ferraro currents. The SYM and ASY indices have been proposed by Iyemori (1990); their network is also shown in Fig. 2c. Considering (i) the distribution of geomagnetic observatories, and (ii) the present knowledge on the behaviour of the non-symmetric part of the ring current magnetic field, the SYM-H and ASY-H indices can be considered as a state-of-the-art solution for monitoring the ring current magnetic field. The SYM-H index is essentially the same as Sugiura's hourly Dst index, but with the advantage of being derived on a one-minute basis, and from a set of six stations, or groups of stations. It is worth noting that both Dst and SYM-H zero values have no physical meaning. The ASY-H index measures both the direct and the unloading response of the magnetosphere. In particular the signature of substorm onsets takes the form of a sharp positive peak in ASY-H.

The auroral activity indices (AE, AU, AL, A0, classically referred to as AE indices or simply AE) have been introduced by Davis and Sugiura (1966). They are at present based on the transient variations in the geomagnetic North component observed at a network of 11 observatories distributed in longitude over the auroral oval (see Fig. 2d). The AE indices are produced on a one-minute basis.

The AU and AL indices are intended to represent a measure of the maximum current density of the eastward and westward auroral electrojets, respectively. The AE index ($AE = (AU–AL)/2$; $AL < 0$) aims at representing global auroral electrojet activity. It monitors the magnetic activity produced by enhanced ionosphere currents in the auroral zone, mostly related to the magnetosphere-ionosphere coupling through the field aligned currents. The A0 index ($A0 = (AU + AL)/2$) aims at representing the symmetry between eastward and westward electrojets. The AE stations are located at standard auroral oval latitudes. They may fail to properly capture the magnetic signature of the auroral phenomena during periods of intense geomagnetic activity, as a result of the associated equatorward motion of the auroral oval.

Finally, the Polar Cap (PC) magnetic index has been proposed by Troshichev et al.(1988), and revised by Troshichev et al. (2006). In fact, there are two PC indices: the PCn (Northern hemisphere) and the PCs (Southern hemisphere). In each hemisphere, the PC index is based on the magnetic disturbances observed at a single near-pole station (see Fig. 2d) and is produced on a one-minute basis. It aims at monitoring the magnitude of the transpolar convection electric field which drives the transpolar part of the polar ionosphere current system, and it is linearly correlated in a statistically optimal way with the solar wind merging electric field. As a result, increasing PC values can be interpreted as increasing dayside merging. When using PC indices, in particular during exceptional events, one should however keep in mind that each individual value relies on a statistical analysis of data from a single station.

CONTRIBUTION TO SOLAR TERRESTRIAL PHYSICS: A CASE STUDY

We present in this section a case study: the 2003, May 29th–30th intense magnetic storm, essentially from the geomagnetic indices point of view. This violent geomagnetic event, described into details by Hanuise et al. (2006) has a complex structure, since it corresponds to several successive interplanetary shocks and pressure pulses hitting the magnetosphere. Variations of solar wind and IMF parameters and geomagnetic indices during the storm are presented in Fig. 3.

On May, 29th, before the storm impact at 12:30 UT, the By and Bz components of the IMF are fairly weak and stable, and the solar wind pressure is very low (~ 2 nPa) (Fig. 3, panels a to c). The storm onset is characterized by strong increase of the amplitude of all the IMF components and of the solar wind pressure. Between 12:30 and 19:00 UT, the IMF-By is variable, oscillating between -10 and $+10$ nT and the IMF-Bz becomes strongly negative around -10 nT. Such a southward IMF-Bz at the beginning of the storm causes a strong coupling with the magnetosphere. Between 19:00 and 02:00 UT (on May, 30th), the IMF-By, still variable, has extreme values of ± 20 nT; the IMF-Bz becomes variable with larger extreme values (-30 and $+20$ nT). After 02:00 UT (on May, 30th), the IMF-By is still variable and the IMF-Bz turns strongly positive up to $+30$ nT. The enhancement and rotation of the IMF-Bz is characteristic of a magnetic cloud. The solar wind pressure shows

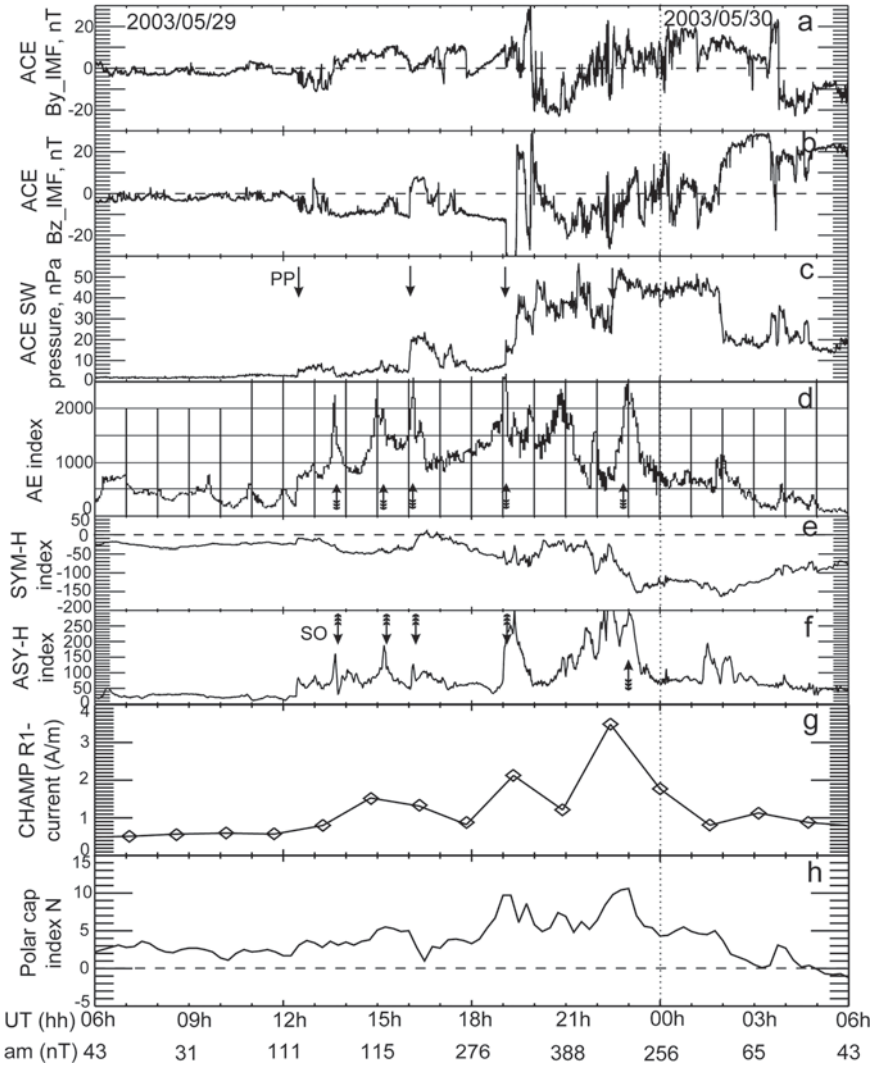


Figure 3. Characterization of the main phase of the magnetic storm (from Hanuise et al., 2006). Panels a to c: IMF and solar wind pressure derived from ACE observations (PP: pressure pulse). Time is delayed by 36 minutes in order to match magnetosphere data. Panel d: AE index. Panels e and f: SYM-H and ASY-H indices (SO: substorm onset). Panel g: R-1 current intensity (latitudinally integrated current density) in the 15:00–16:00 MLT sector derived from CHAMP magnetometer data. Panel h: PCn index

four successive and strong increases, at 12:30 UT, 16:00 UT, 19:00 UT, 22:30 UT (up to 50 nPa) followed at 02:00 UT (on May, 30th) by a strong decrease down to 20 nPa, associated with the IMF-Bz turning. The first and third pressure pulses are associated with interplanetary shocks (characterized by sharp velocity variations).

We first consider the am index (values indicated at the bottom of Fig. 3), which gives a global magnetic view of the magnetosphere state. Before the storm onset, it remains relatively stable around $\sim$ 30–40 nT. From the storm impact, at 12:30 UT, the index increases from $\sim$ 100 nT to $\sim$ 400 nT, reaching its maximum during the 20:00–24:00 UT period on May, 29th. The maximum of the magnetic activity occurs when the IMF-Bz components turns positive. This result is quite surprising, as the coupling between the magnetosphere and the solar wind is weaker during positive IMF-Bz. The high solar wind pressure and the accumulation of magnetic energy in the magnetosphere during the first part of the storm can maybe explain this behaviour. The magnetic activity becomes quiet again at the end of the storm, after 04:00 UT on May, 30th.

The storm onset is also associated with a step-like increase of the SYM-H index (Fig. 3, panel e). Then, between 12:30 and 23:00 UT, this index decreases down to less than -150 nT, also by steps which are associated with the four solar wind pressure pulses causing magnetosphere compressions and ring and magnetopause current intensifications. Two successive minima are observed at 23:15 and 02:00 UT (on May, 30th). Then, SYM-H re-increases smoothly until the end of the storm.

As shown by the AE index, the auroral activity is also rather small before the storm onset (Fig. 3, panel d). After 12:30 UT and the arrival of the first shock, the AE index strongly increases, indicating an enhancement of the auroral electrojet intensity. The three successive peaks observed between 12:00 and 18:00 UT are characteristics of substorm onsets (SO) in the magnetotail and associated enhancements of the electrojets, as consequences of the storm impact on the magnetosphere. A more pronounced increase is observed after 18:30 UT. It is most likely related to the arrival of the most prominent solar wind shock. Around 22:30 UT, upon the arrival of the last pressure pulse, the AE index start to grow again to high values, showing again successive peaks probably due to substorm onsets. AE reaches its maximum at about 23:00 UT; afterwards, it significantly decreases to moderate values. The electrojet intensity is probably decreased because of the reduced magnetic field merging between northward interplanetary and terrestrial magnetic fields.

The ASY-H index (Fig. 3, panel f) shows roughly the same profile as the AE index, especially the peaks due to substorm onsets. This result is in agreement with the fact that the ASY-H index is sensitive to intensifications of the ring current through plasma injections from the magnetotail.

The PCn index (Fig. 3, panel h) has variations similar to those of AE. The rather small values of the index before the magnetic storm onset correspond to low intensity transpolar electric field. After 12:25 UT (first shock), the PCn index increases fairly smoothly, thus indicating a progressive enhancement of the polar cap current. A more pronounced increase of PCn is observed after the arrival at the magnetosphere of the most prominent shock (19:00 UT). After the last pressure pulse (around 22:30 UT), the PCn starts growing again to high values and reach its peak at about 23:00 UT. After 02:00 UT (on May, 30th), the IMF-Bz turns strongly northward and PCn decreases down to low values, even if the end of the storm

does not occur before 04:00 UT. This result illustrates the high sensitivity of the PC index to dayside merging.

To finish, we study the dayside field-aligned currents (FACs) which are directly related to the energy transfer from the interplanetary medium to the magnetosphere. We use the magnetic data of the CHAMP low-altitude satellite during several successive passes of the satellite in the afternoon sector of the auroral oval (15:00–16:00 MLT), to estimate their intensity (Fig. 3, panel g). Strong intensifications of the FACs are observed during the first part of the storm, to more than 10 times the typical quiet time value. The northward turning of the IMF (after 02:00 UT on May, 30th) is followed by a strong decrease of the FACs intensity. These observations show that the dayside FACs intensity is, like PC, a very good monitor of the solar wind-magnetosphere coupling.

This case history illustrates how geomagnetic indices monitor the magnetosphere behaviour. Key periods and events can be identified, which mark the magnetosphere evolution in response to the solar wind/IMF forcing.

WHICH INDICES FOR SPACE WEATHER STUDIES?

The case study presented in the previous section illustrates that the present geomagnetic indices enable one to depict the magnetosphere behaviour in response to the solar wind/IMF forcing. However, there are still some insufficiencies with the existing indices. First, all these indices are planetary ones, consequently they cannot describe local phenomena. Moreover, part of these indices have a low time resolution and cannot describe the temporal evolution of some processes. In this section, we present some examples of new indices allowing a better description of the magnetosphere.

Let us stress that new magnetosphere indices should keep in future the same basic properties as at present. Their definition and derivation scheme should be clearly stated; their physical meaning should be clear; the data series should be homogeneous. Free of charge access is mandatory for all geomagnetic indices based upon data provided by institutes funded by public state agencies.

The three hour time resolution of the K-derived planetary indices is directly inherited from the K index definition. It drastically limits their contribution to the time monitoring of the magnetosphere dynamics. As stated in Section 2, aK equivalent amplitudes are proxies of the magnetic energy density. Because of the morphology of the irregular variations, this relation is in fact mostly valid for 3-hour long intervals at sub-auroral latitudes (Menvielle, 1979). Since the magnetic energy density is directly related to the square of the modulus of the magnetic field, the running mean square (rms) of the two horizontal components can be used as proxy of the magnetic energy density. Menvielle (2003) showed that substituting the rms to the aK in the aa derivation scheme leads to a geomagnetic index that is statistically closely related to the aa index. Furthermore, it makes it possible to derive planetary magnetic activity indices with a time resolution down to 15 minutes (Fig. 4).

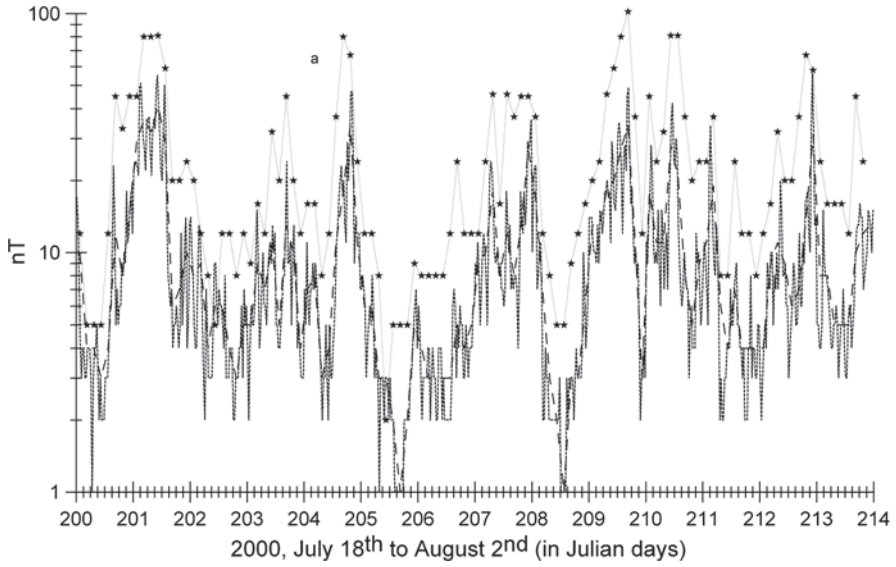


Figure 4. Comparison between aa indices (stars) and aa-like indices based upon the rms of the two horizontal components of the irregular variations during time interval of length $\tau = 180$ minutes (hatched line) and $\tau = 30$ minutes (solid line) (from Menvielle, 2003)

The overview of the magnetosphere dynamics presented in Section 1 evidences that the magnetic activity depends on LT (Local Time), i.e. on longitude. The planetary character of all the indices makes it impossible to get precise information on the LT, or MLT (Magnetic Local Time) dependence of the magnetosphere activity. This suggests deriving geomagnetic indices bearing information on the longitude modulation of the geomagnetic activity, i.e. its variation with longitude during a given UT time interval. Although not used in the case study presented in section 3, longitude sector indices have already been proposed by Menvielle and Paris (2001). These indices, called aλ indices correspond to the equivalent amplitudes computed for the sectors used in the am derivation. There are accordingly 9 longitude sector indices (5 in the Northern hemisphere, and 4 in the Southern one) computed for each 3-hour time interval; aλ indices are expressed in nT. They open the way for statistical studies of the longitude modulation of space weather effects (see, e.g., Lathuillère and Menvielle, 2005). As already mentioned for the planetary geomagnetic indices, their 3-hour resolution limits their contribution to detailed monitoring of space weather events. Derivation of similar longitude sector indices with a better time resolution is thus worth being considered.

The actual indices of the auroral activity are also planetary indices and do not discriminate between dayside and nightside phenomena. However, at auroral latitudes, the daytime activity is directly related to magnetic reconnection on the dayside magnetopause, while the night time activity is driven by magnetic stresses and reconnection in the magnetotail. Introducing two auroral activity geomagnetic

indices monitoring auroral daytime and night time activity respectively, would thus makes it possible to discriminate between the solar wind/IMF forcing and the effects of magnetosphere tail processes. On the other hand, the results presented in Section 3 show that the intensity of the dayside FACs can be used for monitoring the coupling between the solar wind and the magnetosphere. It would be very interesting to define a new index based on such parameter, taking advantage of the ongoing magnetic Ørsted and CHAMP missions and of the forecoming SWARM mission.

Consider now regional and local space weather effects. One of the most widely studied is the impact of Geomagnetically Induced Current (GIC) on industrial activities (e.g., pipe line corrosion, induction in electric power lines). The key physical quantity is the locally induced geoelectric field over the region where the conducting structures (pipe lines, power lines) are installed. This induced field is driven by the time rate of change of magnetic field, the geometry of the pipe network, and properties of the medium (described in terms of e.g., surface impedance). This space weather effect is mostly important in auroral regions. Geomagnetic indices dedicated to monitoring GICs should therefore be settled on a proxy of the time derivative of the vector magnetic field $\mathbf{H}$, $\partial\mathbf{H}/\partial t$ in the area of interest. The great space and time variability of the transient magnetic variations in the auroral zones suggests that local variations of $\partial B/\partial t$ cannot be properly accounted for by a global index. The differences between $\partial B/\partial t$ observed at two nearby observatories, Ottawa, Canada and Fredericksburg, USA (Balch, 2004), illustrate the necessity of 'regional' indices deduced from a rather dense network of magnetic stations at a 'regional' scale (typically 100 km spacing between stations).

The example of GICs illustrates a crucial point: when considering local space weather issues, one should keep in mind that the user needs critically depend on the application. The best solution is using information on the magnetic activity from nearby geomagnetic station(s), for deriving an index which answer the specific user needs. Such index generally takes benefit from results of academic research activities, and its definition turns out to be a compromise between what should be an ideal index and the available data.

REFERENCES

Balch, C.: Limitations of geomagnetic indices and new trends in specifying geomagnetic activity, Space Weather Week, NOAA Space Environment Center, Boulder, Colorado, USA, Oral presentation (2004)

Bartels, J.: The standardized index, Ks, and the planetary index, Kp, IATME Bulletin 12b, 97 (1949)

Berthelier, A.: The geomagnetic indices: derivation, meaning and uses in Solar Terrestrial physics, in STPW-IV proceedings, Hruska J., Shea M.A., Smart D.F. and Heckman G., editors. US Gov. Publication Office, Boulder, 3–20 (1993)

Cowley, S.W.H.: The causes of convetion in the Earth's magnetosphere, A review of developments during the IMF, Rev. Geophys. **20**, 531–565 (1982)

Davis, T.N., Sugiura, M.: Auroral electrojet activity index *AE* and its universal time variation, J. Geophys. Res. **71**, 785–801 (1966)

Hanuise, C., Cerisier, J.-C., Auchère, F., Bocchialini, K., Bruinsma, S., Cornilleau-Wehrlin, N., Jakowski, N., Lathuillère, C., Menvielle, M., Valette, J.-J., Vilmer, N., Watermann, J., Yaya, P.: From the Sun to the Earth: impact of the 27–28 May 2003 solar events on the magnetosphere, ionosphere and thermosphere, Ann. Geophys. **24**, 129–151, (2006)

Iyemori, T.: Storm-time magnetospheric currents inferred from mid-latitude geomagnetic field variations, J. Geomag. Geoelec. **42**, 1249–1265 (1990)

Kivelson, M.G., Russel, C. T.: Introduction to Space Physics, Cambridge University Press, Cambridge, New York, Melbourne (1995)

Lathuillère, C., Menvielle, M.: WINDII thermosphere temperature perturbation for magnetically active situations, J. Geophys. Res. **109**, A11304, doi:10.1029/2004JA010526 (2004)

Mayaud, P.N.: Indices Kn, Ks, Km, 1964–1967, 156 p., Centre National de la Recherche Scientifique, Paris (1968)

Mayaud, P.N.: A hundred year series of geomagnetic data, 1868–1967: indices aa, storm sudden commencements, 256 p., IUGG Publ. Office, Paris (1973)

Menvielle, M.: A possible geophysical meaning of K indices. Ann. Géophys., **35**, 189–196 (1979)

Menvielle, M.: On the possibility to monitor the planetary activity with a time resolution better than 3 hours, in Proceedings of the Xth IAGA Workshop on Geomagnetic Instruments Data Acquisition and Processing, L. Loubser editor, HMO publication, 246–250 (2003)

Menvielle, M., Berthelier, A.: The K-derived planetary indices: description and availability, Rewiews Geophys. **29**, 415–432. Correction, Rev. Geophys. **30**, 91 (1991)

Menvielle, M., Paris, J.: The aλ longitude sector geomagnetic indices, Contrib. Geophys. Geod. **31**, 315–322 (2001)

Rangarajan, G.K.: Indices of geomagnetic activity, in *Geomagnetism*, J.A. Jacobs editor, Academic Press, San Diego, Calif., USA, 323–384 (1989)

Sugiura, M.: Hourly values of equatorial Dst for the IGY, Ann. Int. Geophys. Year, **35**, 9–45 (1965)

Sugiura, M., Kamei, T.: Equatoriel Dst index, 1957–1986, Berthelier, A., and Menvielle, M., editors, IAGA Bulletin 40, 246 p., ISGI Publ. Off. Saint Maur (1991)

Troshichev, O.A., Andrezen, V.G., Vennerstrøm, S., Friis-Christensen, E.: Magnetic activity in the polar cap – A new index, Planet. Space Sci., **36**, 1095–1102 (1988)

Troshichev, O., Janzhura, A., Stauning, P.: Unified PCN and PCS indices: Method of calculation, physical sense, and dependence on the IMF azimuthal and northward components, J. Geophys. Res. **111**, A05208, doi:10.1029/2005JA011402 (2006)

CHAPTER 5.2

THE VALUE OF REAL-TIME GEOMAGNETIC REFERENCE DATA TO THE OIL AND GAS INDUSTRY

JAMES BOWE AND SIMON MCCULLOCH
Halliburton, Tay Facility, Aberdeen AB21 0GL, United Kingdom

INTRODUCTION

A primary requirement for an economically successful oil or gas field development is to drill the boreholes along predetermined trajectories within the reservoir. Known as directional drilling, the geological objectives for these boreholes are often only a few tens of metres wide at distances of several thousands of metres away from the surface drilling location. Improving the quality of real-time borehole positioning measurements is a key element in keeping these wells on target.

Uncertainty with the magnetic environment is a concern for the oil and gas exploration industry because magnetically referenced survey tools remain the predominant source of real-time borehole positioning data. Referred to as measurement-while-drilling (MWD) systems, most share a fundamentally similar sensor configuration. Three accelerometers measuring the gravitational field vector are used to determine borehole inclination. Borehole direction is derived from measurements of the horizontal component of the geomagnetic vector by three fluxgate magnetometers. Both sets of sensors are in alignment with each other. One sensor in each set is aligned along the axis of the tool (and thereby the borehole), with the other two sensors aligned cross-axially and orthogonal to one another.

The use of real-time geomagnetic data now enhances while-drilling magnetic surveys to accuracy levels previously only achievable with post-drilling gyroscopic surveys. This higher level of accuracy has transformed the way boreholes are surveyed especially in areas of higher geomagnetic inclination, bringing both engineering and financial benefits to the industry.

J. Lilensten (ed.), Space Weather, 289–298.

ISSUES AFFECTING MAGNETIC MWD SURVEY DATA

Two sources of error are specific to magnetically referenced survey data. Firstly, errors resulting from the magnetic interference associated with the drilling environment in the form of *drill string magnetism* and *magnetized drilling fluid*. Secondly, errors resulting from the uncertainty in the local geomagnetic field vector. Errors in the magnetic declination (D) affect every calculated survey direction, whilst errors in the magnetic inclination (I) and field intensity (F) have major implications for real-time data quality control and the azimuth correction procedures that are routinely used within the borehole surveying industry. The earlier these errors can be detected and corrected during the drilling process the less detrimental they will be to borehole placement.

Magnetic Interference from the Drilling Environment

Almost the entire drill string is composed of ferrous steel, with only a short section of non-magnetic steel components in which the magnetic survey instrument is positioned. Ideally a magnetic survey tool will be spaced in the non-magnetic section at such a point that the effects of drill string magnetism are negligible. However, there are significant advantages in reducing the amount of non-magnetic steel in a drill string, as well as placing the MWD tool as close to the drill bit as possible. So, it is now rare that a magnetic MWD sensor is placed in a truly non-magnetic environment. At the location of the directional sensor, drill string magnetism acts along the axis of the tool. As a result, measurements from the axial magnetometer are corrupted by a bias error, and this bias causes an error in the calculated direction of the borehole.

Drilling operations normally require that the borehole is filled with a fluid[1]. As a result, MWD sensors are operating in constant proximity to this fluid and magnetically susceptible particles suspended within the fluid create an error by shielding the magnetometers from the full geomagnetic field. These particles have two potential sources: additives containing hematite and ilmenite (used to increase the weight of the fluid) and abrasion of steel from the drill string and casing that lines the upper sections of a borehole. It seems that much of this material is so fine it remains in suspension and is not easily removed during the normal circulation of the fluid system. Consequently, scale factor errors are generated on the cross-axial magnetometers. At geomagnetic latitudes such as the North Sea, it is not uncommon to generate scale factor errors large enough to produce direction errors of 5 degrees (Torkildsen et al. 2004).

---

[1] Drilling fluid serves several functions. It stabilizes the borehole against wall collapse, prevents the ingress of formation fluids, lubricates the drill bit, powers down-hole drilling tools and provides the medium for the transportation of rock cuttings from the borehole. The fluid is pumped down the inside of the drill string, out through the bit and back to surface via the annulus between the outer wall of the drill string and the borehole wall.

The effects of both of these error sources can be mathematically corrected. The technique converges upon the set of values for bias, scale factor and misalignment corrections that minimizes the residual error in the calculated total field (Estes and Walters 1989). Accurate estimates of I and F values are essential for a reliable calculation of these errors.

Uncertainty of the Geomagnetic Field

Accurate transformation of the tri-axial sensor measurements from the tool's XYZ axes system to Earth's North, East, and Vertical axes system requires an accurate value of D. Errors in D caused by a crustal anomaly will affect every magnetic survey direction in that area. At higher geomagnetic latitudes the errors in D caused by short-term field variation can be a problem when measuring the performance of drilling tools during critical directional steering operations.

MWD magnetometer performance is monitored by comparing the measured values of F and I with the theoretical estimates of these components for the same time and place. Charts and computerized global models of the main-field (based on spherical harmonics) still provide the sole source of theoretical geomagnetic vector data on most oil and gas wells drilled today. Table 1 compares the estimate uncertainties for the BGGM[2] main-field model that are typical for the North Sea (Williamson 2000) with the usual tolerances that MWD operators use to define acceptable magnetometer data quality. It shows how difficult it can be to distinguish sensor performance variations from main-field model uncertainty at mid to higher geomagnetic latitudes.

The inherent uncertainty of main-field models as a consequence of local crustal anomalies and external-field variation is well understood. The industry has established acceptable levels of geomagnetic uncertainty that would have a negligible detrimental impact on magnetic survey accuracy (Russell et al. 1995). Table 2 compares estimates of the uncertainty of the BGGM and two geomagnetic field-referencing techniques (which are explained in the next section) with the industry's target values. At geomagnetic latitudes similar to (or greater than) the North Sea,

Table 1. Main-field model uncertainties and MWD measurement tolerances

	Inclination (I) [deg]	Intensity (F) [nT]
North Sea main-field model uncertainty*	0.40	260
MWD magnetometer measurement tolerance	±0.50	±300

*Values supplied by British Geological Survey.
Uncertainties are quoted at 95% confidence levels.

[2] British Geological Survey Global Geomagnetic Model.

Table 2. Typical uncertainty levels with geomagnetic models in the North Sea

	Declination (*D*) [deg]	Inclination (*I*) [deg]	Intensity (*F*) [nT]
Industry Targets[a]	0.10	0.05	50
BGGM[b]	1.00	0.40	260
IFR[c]	0.30	0.30	120
IIFR[c]	0.20	0.10	70

IFR In-field Referencing, *IIFR* Interpolation In-field Referencing.
[a]After Russell et al. 1995.
[b]After Williamson 2000.
[c]Values supplied by British Geological Survey.
Uncertainties are quoted at 95% confidence levels.

only the Interpolation In-field Referencing technique approaches the industry's target levels of acceptable uncertainty. With the best quality data available it can now match the industry's targets on specific projects.

A METHOD FOR REAL-TIME GEOMAGNETIC MEASUREMENTS

The method most commonly used to derive real-time values of the local geomagnetic vector at a drilling location is known as *Interpolation In-field Referencing* (IIFR) (Russell et al. 1995). In essence, the main-field component is obtained from the BGGM whilst the crustal-field is calculated from locally mapped data, typically with measurements made at a sample interval of 5–10 m along survey tracks with 2–6 km of separation. Downward continuation of the main-field is included; and, provided the mapped area is sufficiently large, it is also applied to the crustal-field. Continuous measurements of the time varying external-field are collated from a number of fixed magnetic observatories by the British Geological Survey, Edinburgh, and interpolated for the drilling location. Where significant geomagnetic vector variation exists along a proposed borehole, interpolations are conducted for multiple locations to account for the gradients. In this way, real-time values of the local geomagnetic vector can be generated at 1-minute intervals throughout the drilling process for any location where suitable observatories exist within a region. As MWD data are pumped to surface, the time-stamped measurements are correlated with the IIFR values; and the corrected survey is then calculated. The IIFR values are updated and available via a secure website at 10-minute intervals, sufficient to enable real-time corrections in all but a very few instances.

The use of established observatories to generate virtual real-time observatories unique to individual drilling sites has major financial and practical benefits, particularly for remote offshore developments (Williamson et al. 1998). In the case of Alaska, a single bespoke observatory at Prudhoe Bay covers all the drilling operations in that area as they are sited along similar geomagnetic latitudes with relatively small local time differences.

The correction technique using geomagnetic reference data based on just the main and crustal-field components is commonly referred to as *In-field Referencing* (IFR).

BENEFITS OF REAL-TIME GEOMAGNETIC REFERENCE DATA

In the North Sea the average reduction achieved by the addition of the external-field in uncertainty of the estimate of the geomagnetic vector is about 20% when compared to estimates based on an IFR model (Reay et al. 2005). The differential between IIFR and IFR estimate uncertainties increases with increasing geomagnetic latitude and the phase of the geomagnetic activity cycle. As a result, in the northern North Sea the reduction in the uncertainty of the estimate of F between IIFR and IFR models can be as great as 75%.

Figure 1a compares MWD measured values of I and F with the IIFR modelled values for a single tool run in a North Sea borehole. Only one measured value (for I) lies within the tolerance limits, and this lack of correlation between the values would be of alarm to drill-site engineers trying to monitor instrument performance. However, real-time assessment of the data using multi-station analysis[3] techniques showed significant axial bias and cross-axial scale factor errors from a combination of drill string magnetism and magnetized drilling fluid. Figure 1b shows the same data after applying mathematical corrections for the bias and scale factor errors. All the MWD data now lie within the quality control limits. Because IIFR was used to

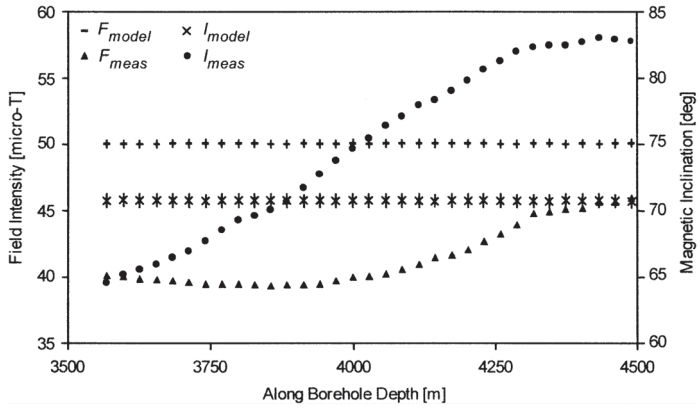


Figure 1a. Comparison of measured MWD values (F_{meas}, I_{meas}) with IIFR modelled estimates (F_{model}, I_{model}). Drill-site quality control tolerances for the survey measurements are shown as vertical error bars on the IIFR base-line values illustrating the degree of poor correlation between both sets of data

[3] A range of algorithms in widespread use by mainstream MWD operators that analyze sets of raw sensor data in order to detect and correct systematic errors.

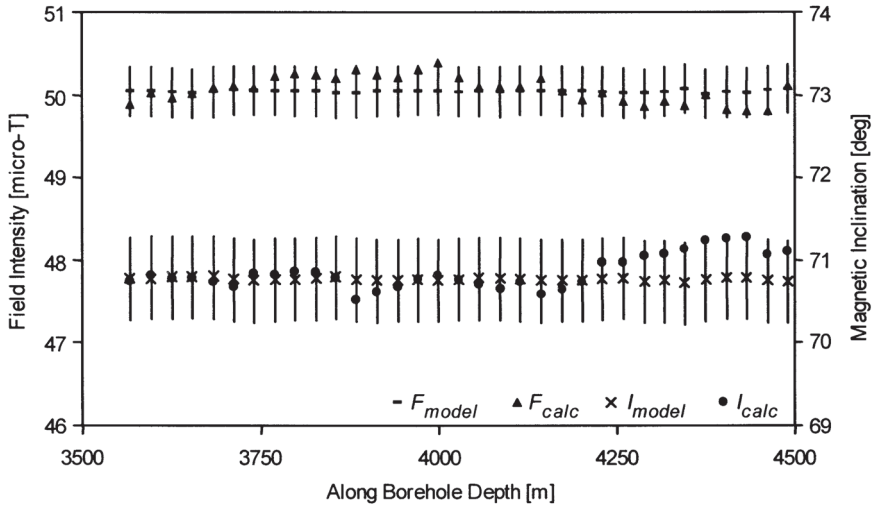


Figure 1b. Comparison of the mathematically corrected MWD values (F_{calc}, I_{calc}) and the IIFR modelled estimates (F_{model}, I_{model}). Drill-site quality control tolerances for the survey measurements are shown as vertical error bars on the IIFR base-line values

calculate the geomagnetic reference vector, a higher level of confidence could be placed on the mathematical corrections than would have been possible using either IFR or main-field models.

The difference between the raw measured direction and the corrected direction values from the previous example is illustrated in Fig. 2. The value of being able to assess these data and correct them in real-time is clear. Without a real-time correction the borehole would have been drilled in a direction several degrees away from the planned trajectory. This difference would have resulted in missing the

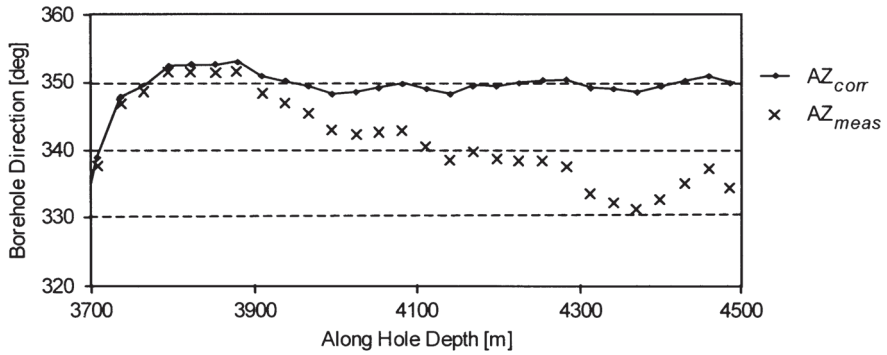


Figure 2. Comparison between measured and corrected MWD directions

target area in the reservoir and therefore, if detected, would have required remedial directional drilling to realign the borehole.

The mathematical corrections for drill string magnetism (and field attenuation effects) are highly sensitive to errors in the local I and F values at certain attitudes (Fig. 3). Real-time geomagnetic reference data (IIFR) reduces both the overall level of calculated azimuth error and the zone in which major azimuth errors will occur. In the North Sea, IIFR reference data reduces the potential error in the correction for axial drill string magnetism to levels comparable with low latitude locations like Brunei.

For geomagnetic latitudes similar to those of the North Sea, the added value of the time varying external-field in the local geomagnetic vector is sometimes questioned. It is argued that over an entire well the typical scale of vector variation seen at these latitudes will be relatively small and that a few anomalous surveys taken during periods of severe disturbance will have little impact on the overall placement of the borehole. However, an element of hindsight exists with this view since it ignores the effect that short-term geomagnetic vector variations can have on real-time drilling operations. For example, when drilling rates of penetration are high, large amounts of magnetic survey data are acquired over a short time interval, so a period of disturbance can compromise a significant proportion of the survey. Furthermore, many magnetic storm episodes last for days rather than hours (Fig. 4). Consider a case where a directional driller is attempting to steer a well in a particular direction to intersect a small geological target. The plan requires a constant 4 degrees of azimuth turn for every 30 m of drilled distance to the target entry point, but the MWD survey indicates only 2.5 degrees of turn were achieved since the survey taken 30 m previously. Is the drilling assembly underperforming, or has the declination changed by 1.5 degrees since the preceding survey was taken an hour earlier? Should the decision be made to drill ahead, or pull out of the hole to change the drilling assembly for a more aggressive configuration? A round trip out of the hole and back in will usually cost several tens of thousands of dollars in rig time. Although such judgments are unlikely on the basis of a single survey,

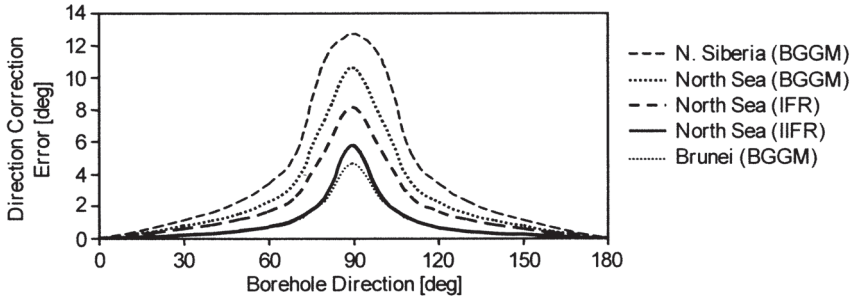


Figure 3. Typical drill string magnetism correction errors for horizontal boreholes as a consequence of local I and F uncertainties (2σ values have been used)

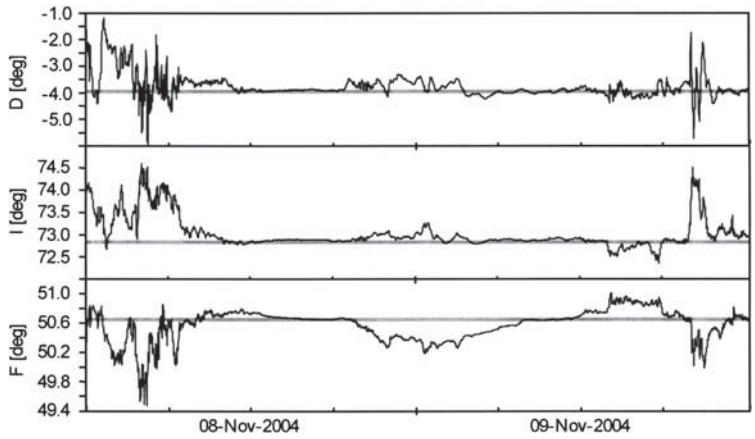


Figure 4. Geomagnetic vector disturbance in the northern North Sea over 48 hours showing significant periods when the vector values were well outside the Industry's acceptable levels of uncertainty (shaded grey). This magnetic storm episode lasted for 96 hours and operators using real-time geomagnetic reference data were able to continue drilling throughout this period with a high degree of confidence in the reliability of real-time magnetic survey measurements (data courtesy of British Geological Survey)

a few more surveys taken during a period of moderate geomagnetic disturbance could easily force an incorrect decision to pull out of the hole.

To ensure that geological objectives are intersected as intended, boreholes are directed toward reduced 'drilling targets'. These are calculated by subtracting the 3-dimensional uncertainty associated with the borehole survey from the dimensions of the original geological target. The error sources inherent with a survey measurement generate an ellipsoid of positional uncertainty around each calculated survey point along the borehole. This creates a cone of uncertainty surrounding the borehole's calculated path in which the actual borehole position may lie. As each type of survey tool has its own error model, based on test stand performance and empirical data, the statistical uncertainty of every survey can be calculated. For target penetration, a typical statistical confidence level would be 95% (2σ). Subtracting the 2σ uncertainties of the major and minor axes of the calculated ellipses of uncertainty for the proposed survey from the dimensions of the geological target produces the inner (drilling) target. In this way, penetration by the borehole anywhere within the drilling target automatically ensures that the geological objective has been achieved at a confidence level of 95%. As previously described, the sensitivity of azimuth correction algorithms to errors in the geomagnetic reference and the uncertainty with declination can be large at geomagnetic latitudes similar to the North Sea, and potentially unmanageable at high latitudes. This level of uncertainty with the external-field means that its omission from the geomagnetic reference has an impact on borehole positional uncertainty much larger than its geomagnetic component size would imply. As a consequence, industry

models of magnetic survey instrument performance generate significantly reduced positional uncertainties in the horizontal plane when the external-field is included in the local geomagnetic reference (Williamson 2000). Reducing the uncertainty in borehole positioning enables smaller reservoirs to be developed and longer reservoir sections to be drilled with a single well (Fig. 5.).

In conclusion, real-time geomagnetic reference data adds significant value in the positional control of directional boreholes in areas of higher geomagnetic latitude. It improves the quality control of magnetic survey data, reduces the uncertainty associated with mathematical correction techniques, and improves directional drilling tool performance monitoring. Reducing uncertainty with MWD survey data has been a contributory factor in extending the range of oil and gas reservoirs that can now be efficiently exploited.

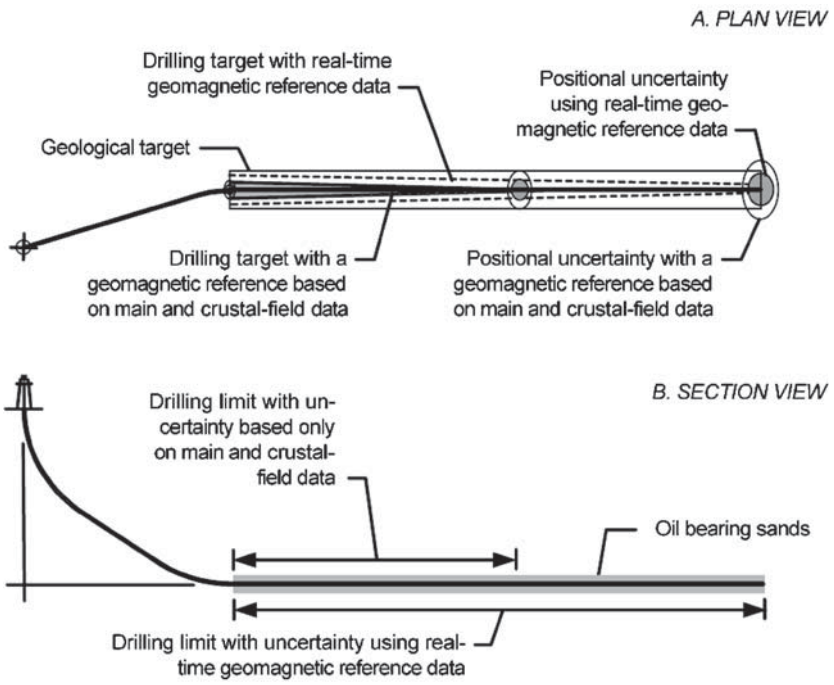


Figure 5. Drilling target size as a function of the geomagnetic reference data type. As the positional uncertainty of a borehole survey increases with increasing drilled depth, the drilling target size decreases proportionally, to a point where the survey uncertainty is equal to the dimension of the geological target (see plan view). This is the limit to which a borehole can be drilled through a geological target at the required level of confidence, with respect to its geometric positioning. Beyond this point there is an increased statistical risk that the actual borehole position will stray outside the boundaries of the geological target. At higher geomagnetic latitudes, the application of a real-time geomagnetic reference significantly lengthens the section of reservoir that can be penetrated by a horizontal borehole when compared to a geomagnetic reference based on just main and crustal field values (see section view)

REFERENCES

Estes, R., Walters, P.: Improvement of azimuth accuracy by use of iterative total field calibration technique and compensation for system environmental effects. Society of Petroleum Engineers, paper SPE 19546 (1989)

Reay, S.J., Allen, W., Baillie, O., Bowe, J., Clarke, E., Lesur, V., Macmillan, S.: Space weather effects on drilling accuracy in the North Sea. Annales Geophysicae, 23, 3081–3088 (2005)

Russell, J.P., Shiells, G., Kerridge, D.J.: Reduction of wellbore position uncertainty through application of a new geomagnetic in-field referencing technique. Society of Petroleum Engineers, paper SPE 30452 (1995)

Torkildsen, T., Edvardsen, I., Fjogstad, A., Saasen, A., Amundsen, P.A., Omland, T.H.: Drilling fluid affects MWD magnetic azimuth and wellbore position. Society of Petroleum Engineers, paper IADC/SPE 87169 (2004)

Williamson, H.S.: Accuracy prediction for directional measurement while drilling. Society of Petroleum Engineers, paper SPE 67616 (2000)

Williamson, H.S., Gurden, P.A., Kerridge, D.J., Shiells, G.: Application of interpolation in-field referencing to remote offshore locations. Society of Petroleum Engineers, paper SPE 49061 (1998)

CHAPTER 5.3

SPATIOTEMPORAL CHARACTERISTICS OF THE GROUND ELECTROMAGNETIC FIELD FLUCTUATIONS IN THE AURORAL REGION AND IMPLICATIONS ON THE PREDICTABILITY OF GEOMAGNETICALLY INDUCED CURRENTS

A. PULKKINEN
NASA Goddard Space Flight Center, Code 612.2, Greenbelt, MD 20771, USA

INTRODUCTION

The central physical quantity from the geomagnetically induced current (GIC) viewpoint is the magnetic field and especially its variations at the surface of the Earth. This follows from Faraday's law of induction. Perhaps the simplest way to see this is to consider a conductor loop placed at the surface of the Earth. The electromotive force *emf* driving the electric current (GIC) in the conductor is given by

$$emf = -\frac{\mathrm{d}\phi}{\mathrm{d}t} \tag{1}$$

where ϕ is the magnetic flux through the loop defined as

$$\phi = \int_S \bar{B} \cdot \mathrm{d}\bar{S} \tag{2}$$

where $\bar{B}$ denotes the magnetic field. From Eqs. (1) and (2) it is quite obvious why the magnetic field $\bar{B}$, and its temporal variations have a special role in GIC studies; information about the spatiotemporal field fluctuations, or ionospheric equivalent currents producing those fluctuations are the starting point of any attempt to model GIC (for a review on GIC modeling, see e.g., Pulkkinen, 2003c, and references therein). Also, this is why the time derivative of the ground magnetic field (denoted hereafter as $\mathrm{d}B/\mathrm{d}t$) is often used as an indicator of the GIC activity (e.g., Viljanen et al., 2001). It is important to note that $\bar{B}$ in Eq. (2) is

J. Lilensten (ed.), Space Weather, 299–310.

actually a sum of fields from two different sources, the external field $\bar{B}_{ext}$ rising due to sources in the magnetosphere and ionosphere and the internal field $\bar{B}_{int}$ rising due to the geomagnetic induction that drives electric currents in the Earth. It follows, that not only the source characteristics but also the ground conductivity structure (dictating the behavior of $\bar{B}_{int}$) determine the GIC flow in the system.

The actual driving term on the left-hand side of Eq. (1) is the electric field integrated along the conductor (voltage). Thus, in fact, the primary field of interest in the context of GIC is the electric field (usually called the geoelectric field). As the actual measured geoelectric field typically represents very local ground structures ($\sim$1 km scale), the geoelectric field is usually determined by applying the so-called plane wave method, which requires knowledge about the ground conductivity structure and the magnetic field variations (Cagniard, 1953). Often the computation of the electric field is skipped and instead the $\mathrm{d}B/\mathrm{d}t$ field is analyzed. Although some counterexamples do exist (Trichtchenko and Boteler, 2006), typically $\mathrm{d}B/\mathrm{d}t$ is a reliable indicator for the GIC activity. Thus, when we talk below about "field" fluctuations, we always refer to the GIC-related fields, i.e. to the modeled geoelectric field and/or to the $\mathrm{d}B/\mathrm{d}t$ field. Also note that although the laterally varying conductivities may play a significant role in the spatiotemporal structure of the total magnetic field and GIC fluctuations (e.g., Thomson et al., 2005), it is assumed here that the most of the observed structure is of external origin. This assumption is discussed and validated for the data used here by Pulkkinen et al. (2006a).

GIC can be thought as an end link of the chain of physical interactions originating all the way from the surface of the Sun. The length and the complexity of the chain indicate that a thorough understanding of GIC and its origins may be a substantial challenge. This is particularly true for auroral regions and especially for geomagnetic storm periods when even the mean (or large spatiotemporal scale) properties of various processes in the near-space are not well understood (see e.g., Sharma, 2003). Note that often the *mean* properties of the processes are not adequate in the context of GIC as we are mainly interested in the *variations*, as was explained above. As the variations around the mean usually are more complex than the mean behavior itself, it is thus reasonable to anticipate that the relevant field fluctuations may pose quite complex spatiotemporal structure that may give some limitations for the predictability of GIC.

The aim of this paper is to briefly review the basic data-derived characteristics of the auroral region electromagnetic field fluctuations. It is hoped that the observed spatiotemporal field characteristics do not only give us hints about the nature of the physical processes driving these fluctuations but also that the characteristics could give us ideas about the predictability of GIC and possibly suggest new strategies for future GIC forecasts. The structure of the work is as follows. First, in Section 2 a basic qualitative "feeling" about the complexity of the field fluctuations is sought for by examining their spatiotemporal properties during some selected time periods. Then, in Section 3 the "feelings" of the Section 2 are put on more quantitative grounds by statistical analysis of the field

fluctuations. In Section 4, the implications of the observed characteristics are discussed and possible new strategies for GIC predictions are sketched. Finally, in Section 5 the findings of this work are summarized and few concluding remarks are given.

EVENT-BASED CHARACTERISTICS

Although the connection between the rapid variations of the ionospheric currents and GIC has been known for a while (e.g., Akasofu and Merrit, 1979), there still exist relatively few studies about the ground electromagnetic field characteristics driving GIC. Some rough estimates of the electrojet intensity and morphology during GIC events were carried out, for example, by Mäkinen, (1993); Bolduc et al. (1998, 2000) and Boteler (2001), but none of these earlier studies focused on the detailed spatiotemporal structure of auroral electromagnetic field fluctuations. Recently, more detailed event-based investigations focusing on the spatiotemporal structure have been carried out by Pulkkinen et al., 2003a, 2003b, 2005 (see also Viljanen et al., 2004). The basic observation of these studies has been, as was anticipated above and is seen from Fig. 1 (from Pulkkinen et al., 2005) where the ionospheric equivalent currents (interpreted in terms of the horizontal magnetic field) and modeled geoelectric field are shown for the October 29, 2003 storm period, that during geomagnetically active time periods the geoelecric fields (and dB/dt) have usually highly non-uniform and variable structure and involve a great variety of different spatiotemporal scales; the spatial scales vary within about 100–1000 km and temporal scales within about 10–1000 s. Note that the lower limits of these scales are given by the limitations of the observations, i.e. the 10-second sampling rate and about 100 km separation of the magnetometer stations.

Another important feature that can be observed from Fig. 1 is that large geoelectric fields tend to occur in the regions where the ionospheric current is enhanced. A similar behavior can be seen from increasing geoelectric field and GIC amplitudes as functions of local activity indices (see, e.g. Mäkinen, 1993;Trichtchenko and Boteler, 2004; Viljanen et al., 2006). These observations underscore the two-fold nature of GIC; although the local GIC-related electromagnetic field fluctuations can be very complex, large field amplitudes tend to occur in regions were the background geomagnetic activity has increased.

At this point it is emphasized that the present investigation focuses on the electromagnetic field fluctuations in the auroral region. It is reasonable to expect that the field structure at lower latitudes is simpler. Although a rigorous GIC-related study on this matter is still missing, indications of this has been seen in studies by Koen (2002) and Hejda and Bochnicek (2005) where the mid-latitude spatial field variations were observed to be clearly smoother (no significant difference in the field fluctuations observed hundreds of kilometers apart) compared to those in the auroral latitudes.

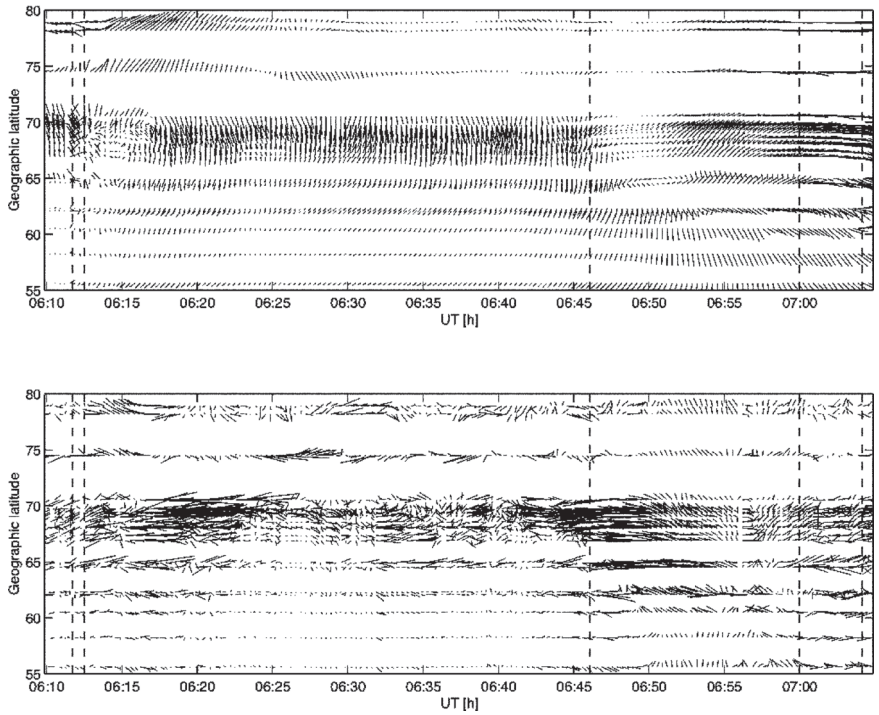


Figure 1. Time series of the horizontal geomagnetic field (top panel) and the modeled geoelectric field (bottom panel) at IMAGE magnetometer array stations on October 29, 2003. Only data for every 20 second is plotted. The dashed lines indicate the times of the failures in the Swedish high-voltage power transmission system. The maximum horizontal geoelectric field and geomagnetic field values of the depicted period are 6680 mV/km and 2580 nT, respectively. See Pulkkinen et al. (2005) for details

STATISTICAL CHARACTERISTICS

How to quantify "complexity"? Statistical methods are clearly needed as attempts to explain individual observations may be doomed by the complex dynamics of the system. The typical approach is to derive some statistical properties of the observed quantity and compare these properties with the properties of some perhaps theoretically better understood process whose known dynamics can then shed light on the dynamics of the studied phenomenon (see e.g., Sornette, 2004). One form of the complexity is the so-called self-similarity, which is familiar, for example, from fractals: the statistical properties of the object do not change under scale transformation. Such invariance under scale transformation is manifested, for example, in the *scaling* of the observables. Scaling means that a statistical property of the observable changes as some power of the scale (power-law):

$$y \propto x^{\alpha} \tag{3}$$

where y is the statistical property, x the scale and α the scaling exponent. In other words, there is no characteristic scale (the statistical property is scale-free). For example, the power spectrum of the *AE*-index scales in temporal scales of about 5–200 minutes roughly as $\alpha \approx 0.5$, which is an indication that the time series is similar to a Brownian process (Takalo et al., 1993). A Brownian process is one of the most basic classes of random processes. In general, the statistical scaling properties of various global geomagnetic indices, all of which are based on the ground magnetic field measurements, have been investigated by numerous authors (e.g., Tsurutani et al., 1990; Takalo et al., 1993; Uritsky et al., 2001; Hnat et al., 2005) and are relatively well-known. However, very few studies have addressed the local spatiotemporal scaling characteristics of the GIC-related electromagnetic field fluctuations.

In Pulkkinen et al. (2006a) the spatiotemporal scaling characteristics of the $\mathrm{d}B/\mathrm{d}t$ fluctuations were investigated by applying the structure function analysis to IMAGE magnetometer array observations. In Fig. 2, the second order moments of the spatial and temporal structure functions for different components of $\mathrm{d}B/\mathrm{d}t$ are shown. The analysis was carried out separately for all data over years 2002–2003 and for

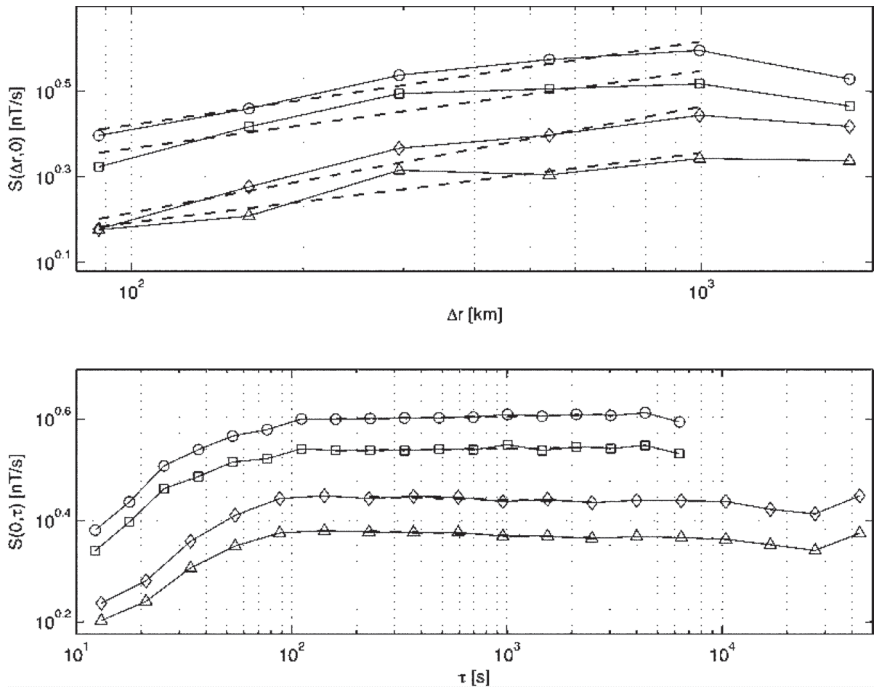


Figure 2. The structure functions $S(\Delta r, 0)$ (spatial) and $S(0, \tau)$ (temporal) and their least squares fits (dashed lines) for the time derivative of the magnetic field. Circles: $S(\Delta r, 0)$ (top panel) and $S(0, \tau)$ (bottom panel) for $\mathrm{d}B_x/\mathrm{d}t$ of substorm events, diamonds: $\mathrm{d}B_x/\mathrm{d}t$ of the full data set, squares: $\mathrm{d}B_y/\mathrm{d}t$ of substorm events, triangles: $\mathrm{d}B_y/\mathrm{d}t$ of the full data set. See Pulkkinen et al. (2006a) for details

detected substorms during the same time period. In agreement with the event-based view, the statistical view of Fig. 2 shows that the temporal behavior of dB/dt for periods longer than about 100 s (there is very clear break in the scaling around these periods) is very complex. By complex we refer here to the "flat" temporal scaling ($\alpha \approx 0$) of the structure function above about 100 s which is very closely to that of white noise. White noise can be viewed as one extreme end of complex processes as it is completely random and thus also completely unpredictable. Although the spatial scaling of the structure function shows a non-zero slope, before interpretations about the behavior of the field in the ionospheric level where the source is operating, one must take into account the low-pass-filtering effect of the field continuation from the ionospheric level to the ground level. When the effect of the field continuation is taken into account, it turns out that from the ionospheric viewpoint, also the spatial scaling is very similar to the scaling of the white noise. It was also found that the same spatiotemporal behavior was true for both horizontal field components and substorms and average overall field variations.

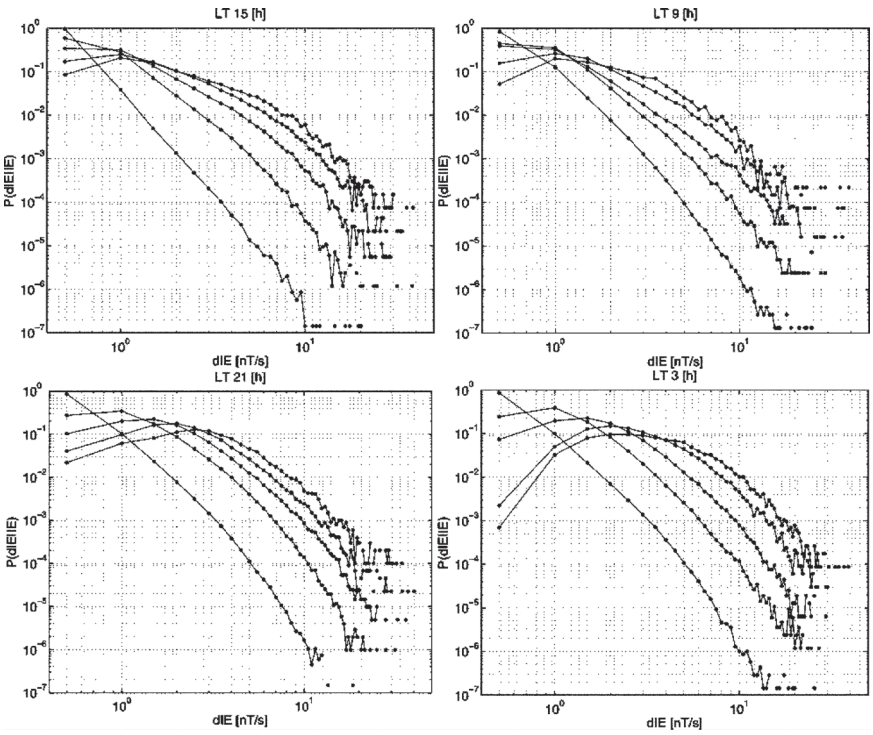


Figure 3. The conditional probability distributions $P(dIE \backslash IE)$ computed for four different local time sectors centered at the times indicated in the titles of the panels. The five curves moving from the left to the right correspond to the *IE*-index values 150, 450, 750, 1050 and 1350 nT, respectively. See Pulkkinen et al. (2006b) for details

The two-fold spatiotemporal behavior of dB/dt seen in the previous section has been found to be present also in a statistical sense. For example, studies by Trichtchenko and Boteler (2004); Pulkkinen et al. (2006b); Viljanen et al. (2006) have shown that there is a clear statistical coupling between GIC-related field variations and the local background level of geomagnetic activity, or as shown by Weigel and Baker (2003b) also, for example, to solar wind forcing. One way to see the coupling is via conditional probabilities that Pulkkinen et al. (2006b) calculated for the *IE*-index (local variant of the *AE*-index computed from IMAGE array observations) and d*IE*-index (the maximum of dB/dt taken over IMAGE stations). The result of this analysis that was carried out separately for different local time sectors is shown in Fig. 3. As is seen, the distributions change quite significantly as a function of the *IE*-index, which is indicating coupling between the two quantities. Also, signaling the very feasible possibility that different mechanisms generate the field fluctuations at different local time sectors, the statistics depend on the local time too.

IMPLICATIONS ON THE PREDICTABILITY OF GIC

We then move to interpret the results discussed above in terms of the predictability of GIC. Although there has been great effort to predict the behavior of the standard geomagnetic indices (see e.g., Boberg et al., 2000; Gleisner and Lundstedt, 2001; Siscoe et al., 2005) and spatially dependent large-scale field structures (e.g., Valdivia et al., 1999; Weimer, 2005), the number of attempts to predict GIC or field fluctuations associated directly with GIC is very small. Note, once again, that as GIC is related to the temporal *variations* of the magnetic field, predicting only the *mean* behavior of the field is not in general sufficient.

The attempts to predict the directly GIC-related field fluctuations have been made by Weigel et al. (2003a) and Wintoft (2005). Weigel et al. (2003a) showed that 30-minute averages of dB/dt can be predicted with some accuracy from the solar wind input for specific auroral regions where also the average overall amplitudes of the 30-minute averages are the largest. These regions are the vicinity of the local midnight and the local morning. The basic conclusion of the study by Wintoft (2005) in which the variance of 10-minute segments of dB/dt was predicted was essentially the same as that in Weigel et al. (2003a), i.e. although not all of the variability can be captured by the model run with the solar wind input, some overall aspects of the variability can be predicted.

Although it is certainly clear that much more understanding about the GIC-related electromagnetic field fluctuations needs to be obtained before anything more definitive can be stated, the investigations discussed above do indicate some aspects of the GIC phenomenon that can be utilized in future prediction efforts. First, it has been showed that some statistical properties of GIC-related field fluctuations (e.g. average or variance of dB/dt) have certain amount of predictability. Also, it has been shown that these field fluctuations are statistically coupled to

spatiotemporally larger scale and also more predictable characteristics of the near-space environment expressed, for example, in terms of geomagnetic indices. All this, of course, is an indication that the field fluctuations are predictable. However, secondly, on a more detailed level it has been seen that the field fluctuations are spatiotemporally very complex. In fact, it was seen that the scaling properties of the fluctuations resembled that of white noise. The behavior of white noise is completely unpredictable, which is, although not conclusive, still an indication that the detailed spatiotemporal behavior of the GIC-related field fluctuations may not be predictable.

The two aspects above are by no means contradictory. What the second aspect in fact says is that exact, *deterministic* predictions of GIC-related field fluctuations may not be possible. Thus, we need to move into a *statistical* description of the phenomenon. This is in agreement with the first aspect: there is a *statistical* coupling (i.e. no one-to-one coupling) between the field fluctuations and the larger scale characteristics of the near-space environment. Thus, both of these aspects suggest that although there clearly is predictability in the system, instead of pursuing strictly deterministic approach to auroral GIC forecasts, it may be beneficial to acknowledge the complex nature of the phenomenon and switch into a stochastic paradigm.

Let us briefly describe what it is meant here by a stochastic paradigm. Consider a state vector $\bar{X}_t$ describing the state of the system at time t. Different dimensions of the state vector could include, for example, magnitudes of GIC at different spatial locations and the state of the magnetosphere-ionosphere system expressed in terms of geomagnetic indices. Now, in the stochastic paradigm the future state of the system at time $t+\Delta t$ is given by relation

$$p(\bar{X}_{t+\Delta t}) = \int p(\bar{X}_{t+\Delta t}|\bar{X}_t)p(\bar{X}_t)\mathrm{d}\bar{X}_t \tag{4}$$

which is valid for all stochastic processes. Note now that in Eq. (4) we have replaced the deterministic mapping of the state $\bar{X}_t$ to $\bar{X}_{t+\Delta t}$ by introducing probability distribution p for observing the system in a certain state $\bar{X}_{t+\Delta t}$ given the known probability distribution of observing the system in the state $\bar{X}_t$. In the stochastic paradigm, the propagation from time t to $t+\Delta t$ is carried out in terms of transition probabilities $p(\bar{X}_{t+\Delta t}|\bar{X}_t)$. Now, the fundamental difference between the deterministic and stochastic views is that the transition probabilities of deterministic systems are delta functions, i.e. given some initial condition at time t, it is known exactly from the equations of motion what state of the system will follow at time $t+\Delta t$. In the stochastic paradigm, it is acknowledged that our knowledge about the evolution of the system is incomplete, i.e. given the state of the system at time t, there is a great number of different states, distribution, to which the system can end up at time $t+\Delta t$.

In its simplest form, a stochastic treatment of the GIC predictions could contain, for example, (1) a deterministic mean field prediction of the local geomagnetic activity which (2) could then be supplemented with a statistical prediction of the corresponding level of $\mathrm{d}B/\mathrm{d}t$. This approach has in fact been implemented already

in the *AL*-index predictions by Ukhorskiy et al. (2004). However, a more complete treatment requires the determination of appropriate variables to be used to describe the state $\bar{X}_t$ and a rigorous determination of the transition probabilities $p(\bar{X}_{t+\Delta t}|\bar{X}_t)$. How these goals can be achieved is clearly all but obvious. Although further speculations are out of the scope of this short article, it is pointed out that the Fokker-Planck formalism for Markovian processes applied already to geomagnetic indices by Hnat et al. (2005), could be one appealing approach to pursue the more complete stochastic description of the GIC-related field fluctuations.

CONCLUSIONS

Above both the event-based and the statistical spatiotemporal properties of auroral GIC-related geoelectric field and magnetic field fluctuations were reviewed. It was found that both views portrayed the two-fold nature of the fluctuations: (1) large fluctuations tend to occur at regions with increased background geomagnetic activity and (2) the detailed behavior of the fluctuations is very complex. The complexity is enhanced in the GIC view where we are interested in the magnetic field variations, not just the mean behavior. Also, it was noted that there has been some success in predicting some overall properties of $\mathrm{d}B/\mathrm{d}t$ fluctuations.

Based on this two-fold nature of the fluctuations, a stochastic approach to GIC predictions was proposed. The stochastic approach acknowledges that although some overall properties of the dynamics of the system can be understood (i.e. the phenomenon is not completely random), there are also important unknown factors that hinder the deterministic treatment of the problem. These unknown factors lead to the statistical treatment of the field fluctuations.

One natural question that arises from the considerations above is what in the solar wind-magnetosphere-ionosphere system gives rise to the randomness that necessitates the suggested stochastic approach. As discussed briefly in the introduction, our near-space system appears very complex especially during strong geomagnetic storms that are of main interest from the space weather viewpoint. The complexity may arise from the fact that the system includes a great number of interacting parts (without specifying what these "parts" may be), the dynamics of the interactions being possibly highly non-linear. These features are especially plausible for the auroral region coupled to the complex dynamics of the plasma sheet (e.g., Klimas et al., 2000). Now, although in principle the system is deterministic, the great number of these interacting parts hinders the treatment of all these parts in a "unified grand deterministic model". Instead, it is acknowledged that the lack of knowledge about the detailed behavior of the system induces randomness to the description of the evolution of the system. This lack of detailed knowledge then leads to the stochastic treatment. This is similar, for example, to the thermodynamic description of the gases where instead of tracking all the particles in the gas, only the statistical properties of the gas are treated

Although the auroral fluctuations were discussed here basically as a single phenomenon, it is pointed out that different processes operate at different parts

of the magnetosphere and thus also at different parts of the auroral ionosphere. These different processes may portray different stochastic properties and thus different degrees of deterministic behavior. For example, it is reasonable to assume that fluctuations associated with substorms are likely less deterministic than those associated with quasi-monochromatic geomagnetic pulsations. Accordingly, different prediction strategies may be desirable for different parts of the auroral region.

ACKNOWLEDGEMENTS

Numerous discussions on GIC and complex systems and fruitful collaboration with Drs. A. Viljanen, R. Pirjola, D. Vassiliadis, A. Klimas and V. Uritsky are greatly acknowledged. This work was performed while AP held National Research Council Associateship Awards at NASA/Goddard Space Flight Center. The work of AP was supported also by the Academy of Finland.

REFERENCES

Akasofu, S.-I., Merrit, R.P.: Electric currents in power transmission line induced by auroral activity, Nature, 279, 308 (1979)

Boberg, F., Wintoft, P., Lundstedt, H.: Real time Kp predictions from solar wind data using neural networks, Physics and Chemistry of the Earth, 25, 275–280 (2000)

Bolduc, L., Langlois, P., Boteler, D., Pirjola, R.: A Study of Geoelectromagnetic Disturbances in Québec, 1. General Results, IEEE Trans. Power Delivery, 13, 1251 (1998)

Bolduc, L., Langlois, P., Boteler, D., Pirjola, R.: A Study of Geoelectromagnetic Disturbances in Québec, 2. Detailed Analysis of a Large Event, IEEE Trans. Power Delivery, 15, 272 (2000)

Boteler, D.: Space Weather Effects on Power Systems, Space Weather, AGU Geophysical Monograph, 125, 347 (2001)

Cagniard, L.: Basic theory of the magneto-telluric method of geophysical prospecting, Geophysics, 18, 605 (1953)

Gleisner, H., Lundstedt, H.: Auroral electrojet predictions with dynamic neural networks, J. Geophys. Res. 106, 24 541–24 550 (2001)

Hejda, P., Bochnicek, J.: Geomagnetically induced pipe-to-soil voltages in the Czech oil pipelines during October–November 2003, Ann. Geophys. 23, 3089–3093, SRef-ID: 1432-0576/ag/2005-23-3089 (2005)

Hnat B., Chapman, S. C., Rowlands, G.: Scaling and a Fokker-Planck model for fluctuations in geomagnetic indices and comparison with solar wind epsilon as seen by Wind and ACE, J. Geophys. Res. 110, A08206, doi:10.1029/2004JA010824 (2005)

Klimas, A.J., Valdivia, J.A., Vassiliadis, D., Baker, D.N., Hesse, M., Takalo, J.: Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet, J. Geophys. Res. 105, 18765 (2000)

Koen, J.: Geomagnetically induced currents in the Southern African electricity transmission network, PhD thesis, University of Cape Town (2002)

Mäkinen, T.: Geomagnetically Induced Currents in the Finnish Power System, Geophysical Publications, No. 32, Finnish Meteorological Institute, 101 p (1993)

Pulkkinen, A., Amm, O., Viljanen, A., BEAR Working Group: Ionospheric equivalent current distributions determined with the method of spherical elementary current systems, J. Geophys. Res. 108, doi:10.1029/2001JA005085 (2003a)

Pulkkinen, A., Thomson, A., Clarke, E., McKay, A.: April 2000 geomagnetic storm: ionospheric drivers of large geomagnetically induced currents, Annales Geophysicae, 21 (709), (2003b)

Pulkkinen, A.: Geomagnetic induction during highly disturbed space weather conditions: Studies of ground effects", Finnish Meteorological Institute Contributions, No. 42. (available at http://ethesis.helsinki.fi/julkaisut/mat/fysik/vk/pulkkinen/geomagne.pdf) (2003c)

Pulkkinen, A., Lindahl, S., Viljanen, A., Pirjola, R.: Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system, Space Weather, 3, S08C03, doi:10.1029/2004SW000123 (2005)

Pulkkinen, A., Klimas, A., Vassiliadis, D., Uritsky, V., Tanskanen, E.: Spatiotemporal scaling properties of the ground geomagnetic field variations, J. Geophys. Res. 111, A03305, doi:10.1029/2005JA011294 (2006a)

Pulkkinen, A., Viljanen, A., Pirjola, R.: Estimation of geomagnetically induced current levels from different input data, submitted to Space Weather (2006b)

Sharma, A.: Disturbances in geospace: the storm-substorm relationship, AGU Geophysical Monograph, 268 p (2003)

Siscoe G., McPherron, R.L., Liemohn, M.W., Ridley, A.J., Lu, G.: Reconciling prediction algorithms for Dst, J. Geophys. Res. 110, A02215, doi:10.1029/2004JA010465 (2005)

Sornette, D. Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools, Springer, 2nd ed. (2004)

Takalo, J., Timonen, J. and Koskinen, H.: Correlation dimension and affinity of AE data and bicolored noise, Geophys. Res. Lett., 25, 1527-1530 (1993)

Thomson, A.W.P., McKay, A.J., Clarke, E., Reay S.J.: Surface electric fields and geomagnetically induced currents in the Scottish Power grid during the 30 October 2003 geomagnetic storm, Space Weather, 3(11), S11002, doi:10.1029/2005SW000156 (2005)

Trichtchenko, L. and Boteler D. H.: Modeling geomagnetically induced currents using geomagnetic indices and data, IEEE Trans. on Plasma Sci., v 32 N4, 1459–1467 (2004)

Trichtchenko, L. and Boteler D.: Response of Power Systems to the Temporal Characteristics of Geomagnetic Storms, Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering, Ottawa, May 2006 (2006)

Tsurutani, B.T., Sugiura, M., Iyemori, T., Goldsteing, B.E., Gonzalez, W.D., Akasofu, S.I., and Smith, E.J.: The nonlinear response of AE to the IMF Bs driver: A spectral break at 5 hours, Geophys. Res. Lett., 17, 279–282 (1990)

Ukhorskiy A.Y., Sitnov, M.I., Sharama, A.S., Papadopoulos K.: Global and multi-scale features of solar wind-magnetosphere coupling: From modeling to forecating, Geophys. Res. Lett., 31, L08802, doi:10.1029/2003GL018932 (2004)

Uritsky, V.M., Klimas, A.J., Vassiliadis, D.: Comparative study of dynamical critical scaling in the auroral electrojet index versus solar wind fluctuations, Geophys. Res. Lett., 28(19), 3809–3812, 10.1029/2001GL013026 (2001)

Valdivia, J.A., Vassiliadis, D., Klimas, A., Sharma, A.S., Papadopoulos K.: Spatiotempral activity of magnetic storms, J. Geophys. Res., 104(A6), 12239–12250, 10.1029/1999JA900152 (1999)

Viljanen, A., Nevanlinna, H., Pajunpää K. and Pulkkinen A.: Time derivative of the horizontal magnetic field as an activity indicator, Annales Gephysicae, 19, 1107–1118 (2001)

Viljanen, A., Pulkkinen, A., Amm, O., Pirjola, R., Korja, T. and BEAR Working Group, Fast computation of the geoelectric field using the method of elementary current systems, Annales Geophysicae, 22, 101–113 (2004)

Viljanen, A., Pulkkinen, A., Pirjola, R., Pajunpää, K., Posio, P. and Koistinen, A.: Recordings of geomagnetically induced currents and a nowcasting service of the Finnish natural gas pipeline system, Accepted for publicaiton in Space Weather (2006)

Weigel, R.S., Klimas, A.J., Vassiliadis, D.: Solar wind coupling to and predictablitiy of ground magnetic fields and their time derivatives, J. Geophys. Res. 108(A7), 1298, doi:10.1029/2002JA009627 (2003a)

Weigel, R.S., Baker, D.N.: Probability distribution invariance of 1-minute auroral-zone geomagnetic field fluctuations, Geophys. Res. Lett., 30(23), 2193, doi:10.1029/ 2003GL018470 (2003b)

Weimer D.R.: Predicting surface geomagnetic variations using ionospheric electrodynamic models, J. Geophys. Res., 110, A12307, doi:10.1029/2005JA011270 (2005)
Wintoft, P.: Study of the solar wind coupling to the time difference horizontal geomagnetic field, Ann. Geophys., 23, 1949–1957, SRef-ID:1432-0576/ag/2005-23-1949 (2005)

CHAPTER 5.4

FINNISH EXPERIENCES WITH GRID EFFECTS OF GIC'S

JARMO ELOVAARA
Fingrid Oyj, P.O.Box 530, FI-00101 Finland

INTRODUCTION

It is well known that geomagnetically induced currents (GIC's) can cause disturbances in electric power networks. In the most dramatic cases expensive equipment failures or long lasting areal black-outs are experienced. Especially the North American continent has suffered from the negative impact of such geomagnetic disturbances, but negative effects have also been noticed in Great Britain and lately in South Africa. The Nordic countries, Finland, Norway and Sweden, are located in an area where the geomagnetic disturbances reach large amplitudes. However, major negative effects have not been experienced in these countries although the city of Malmö faced a black-out caused by GIC's in the beginning of this millennium. A special case is Finland where the disturbances caused by GIC's were extremely rare although very high pseudo-stationary direct currents (DC's) have been measured since the 1970's when continuous monitoring of GIC's in the Finnish 400 kV grid started. We report here about the research made on this subject in Finland and on our attempts to explain the exceptionally good behavior of the Finnish grid during GIC events.

EFFECTS OF GIC'S IN AN ELECTRIC POWER NETWORK

Usually the following three conditions must be fulfilled before a geomagnetic disturbance can cause problems in electric power grids (networks):

- the network is located in an area where the magnitude of geomagnetic disturbances is usually substantial (e.g. at high magnetic latitudes)
- the power line network includes long overhead lines
- the overhead line system has electrically low-resistance connections to the ground.

A geomagnetic disturbance (geomagnetic storm) is associated with an intense ionospheric current which induces a horizontal electric field in the ground (telluric

J. Lilensten (ed.), Space Weather, 311–326.

field). The higher the specific resistivity of the substratum the higher is the telluric field strength. If an overhead line runs across the area, and if the line is connected to the ground at both ends, the telluric field causes a GIC to flow through the line. In power transmission and distribution systems three-phase lines are typically used, and especially in power transmission systems such lines can be hundreds of kilometres long. Because the losses in the transmission systems are minimized, the GIC's, if they have the possibility to flow, will normally meet low resistances along their paths so that even induced voltages of a few V/km can create significant pseudo-stationary DC-currents in the system.

The neutral point of a three-phase transformer is a common point where the voltages and currents of a three-phase system cancel each other under balanced and symmetrical conditions. A neutral point is formed by connecting the three phase windings of a three-phase transformer or generator together at one end. In order to limit equipment costs the neutral points of power transmission transformers are normally earthed. GIC's from each phase conductor can then flow through the transformer windings and via the neutral point to the earth. They thus magnetize the iron core in addition to the magnetization imposed by the regular AC current.

Figure 1 illustrates the basic flow of GIC's in a power transmission system and the distribution of the GIC between the phase-conductors and the neutral lead in a three-phase system.

Distribution systems, on the other hand, do not normally encounter problems due to geomagnetic disturbances because the lines are much shorter and the neutral points of the distribution transformers – if existing – are not earthed at all or are earthed only at the low voltage side. Usually GIC's cannot penetrate into the generators because they are not galvanically connected to the long overhead lines or they have a neutral-earthed generator-transformer between the grid and the generator. However, the side effects of GIC's might still be harmful for generators.

The reason for outages and catastrophic events in transmission grids is usually the half-cycle DC-magnetization of magnetic cores of different transformers. DC-magnetization can cause saturation of the transformer cores and increase the leakage fluxes. When harmonic currents and voltages present in saturated transformers add to the normal power frequency currents and voltages, the resulting quantities can be so high that the relay protection system disconnects lines and/or equipment from the system. In principle, this kind of events (overcurrents and voltages in static compensators) caused the black-out in Quebec (Canada) in March 1989

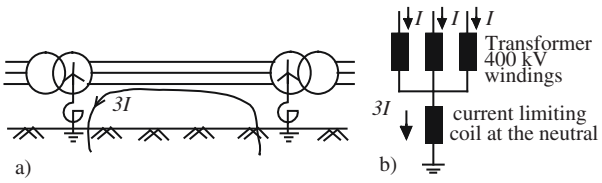


Figure 1. a) The principle of a GIC ($3 \cdot I$) in a power system which has two transformers earthed via current limiting coils. b) Flow of a GIC through the windings and via the neutral point to the ground

(Bolduc 2002). Dramatic effects have been experienced in certain types of large power transformers due to local overheating, because high leakage fluxes might produce local hot spots inside the transformer. Insulating oil may be dissolved and gas bubbles be formed. In the worst case electrical insulants may be burnt. In this case an equipment failure occurs the consequences of which become costly and for which a long repair time is needed. This mechanism is a possible explanation for equipment failures reported in the U.S.A. and U.K. (Molinski 2002). Finally, saturated power transformers consume much more reactive power than transformers in the normal state. If the grid has not enough reactive power resources, problems in keeping up the voltage levels may occur.

FINNISH 400 KV GRID AND GIC OBSERVATIONS

In Finland the solid bedrock reaches close to the surface which results in a high specific resistivity of the ground (average 2300 Ωm at 50 Hz). In the earth-faults of the electric system high ground potential rises occur which are dangerous to people and animals. Therefore the magnitude of the earth-fault currents has been limited in the Finnish system. In the 110–400 kV systems this is achieved through a current limiting coil placed between the transformer neutral and the earth. Further, in 400 and 220 kV systems, the sizes of the current limiting coils (i.e. the 50 Hz reactance of the coil) have been selected so that the magnitudes of the overvoltages created by the faults do not reach very high values. The resistance of the current limiting coils could have been selected freely, because no current is flowing through the coil under normal balanced conditions. In spite of this the coils have a resistance of typically only 1.5–2.5 Ω which is still high compared to the typical resistance of a 400 kV line (0.015–0.03 Ω/km per phase), especially when taking into account that the sum of the GIC's from all three phases flow through this coil. Fig. 1b shows schematically an example of a current limiting coil in transformers connecting 400 and 110 kV systems.

The ground lead of the current limiting coil offers simple and economic means to measure the GIC magnitudes in the grid. GIC's in transformer neutrals have been measured in Finland continuously since 1976. The recordings were started in Huutokoski in Eastern Finland and in Pirttikoski in Lapland. Thereafter the measurement points have been moved due to changes made to the configuration of the grid. The principle has been that the GIC's are measured at those points of the system where the occurrence probability of large GIC's is highest. At the moment the GIC's are continuously measured in Rauma, Yllikkälä and Pirttikoski, but recordings are stored only from intervals of large geomagnetic disturbances. Figure 2 shows the present extent of the 400 kV grid in Finland and the stations where GIC measurements are made. In one measurement campaign the Finnish Meteorological Institute (FMI) tested the measurement of the line-GIC with a magnetometer installed near a 400 kV overhead line.

The largest GIC's observed in Finland were recorded in Huutokoski in 1979 and in Rauma in 1991 (Elovaara et al. 1992). In Huutoskoski in eastern Finland the

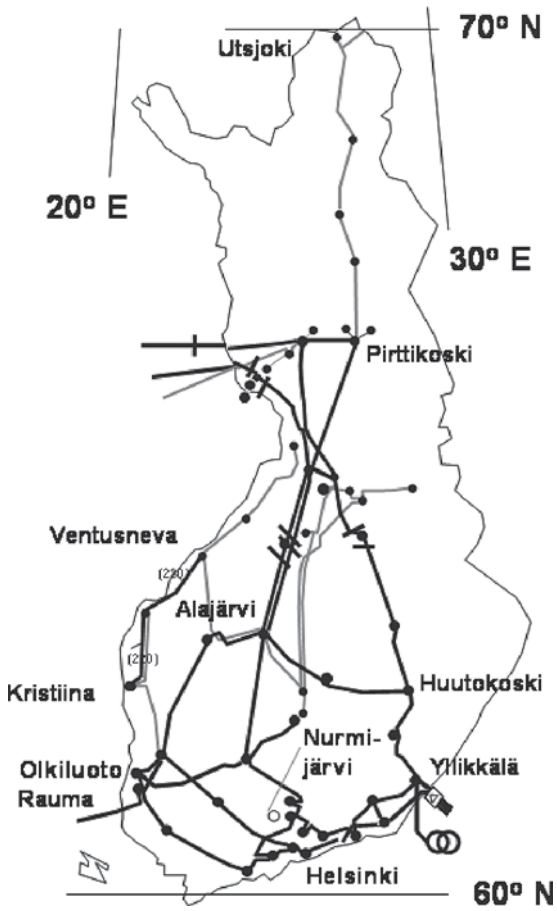


Figure 2. The Finnish 400 kV grid in 2005. The 400 kV lines are black, the 220 kV lines grey. The transversal short line segments on the 400 kV lines indicate the position of the series capacitor banks

10-s value of the GIC in the 400 kV transformer neutral reached the value 165 A on January 4, 1979, and in Rauma in western Finland the measured 1-minute average GIC-values were 200 A and 190 A on March 24 and 25, 1991, respectively.

Figure 3 shows samples of the current recordings in the Rauma and Pirttikoski transformer neutrals on March 24, 1991, and the variation of the geomagnetic field in Nurmijärvi, southern Finland, for the same time interval. The maximum peak value was recorded in Rauma at local midnight but relatively high GIC peaks were recorded also before and immediately after the maximum peak. Another high GIC peak was seen at Rauma about 6 hours later. An interesting detail is that at the time of the maximum peak in Pirttikoski, the GIC in Rauma had only a modest value, but when Rauma had the maximum, also the GIC in Pirttikoski was high. This illustrates qualitatively how the vorticity of the ionospheric currents varies

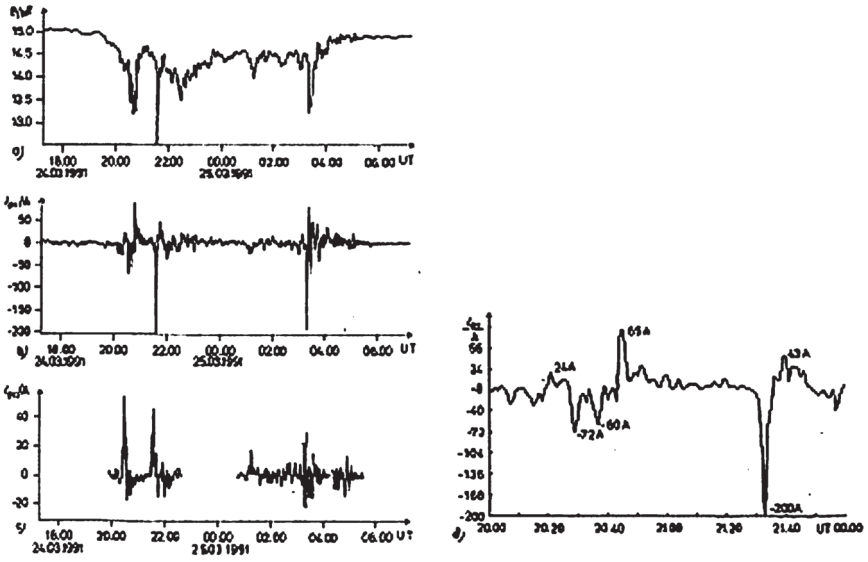


Figure 3. The GIC event on March 24–25, 1991 (Elovaara et al. 1992). a) Variation of the geomagnetic field in Nurmijärvi. b) GIC at Rauma 400 kV transformer neutral. c) GIC at Pirttikoski 400 kV transformer neutral. d) Enlargement of the GIC at Rauma at the time of the maximum value

between different events. Both these events also illustrate with real data the findings of simulation studies, namely that GIC's tend to reach their maxima in substations located at the geographical corner points.

According to Fig. 3 the duration of the individual peaks was relatively short, 3–5 minutes. This is much shorter than for instance the thermal time constant of a power transformer winding. However, also clearly longer GIC durations have been recorded. Figure 4 gives a sample measurement from the Huutoskoski transformer neutral in summer 1978. The maximum value of the GIC has in this case been about 55 A, but the current has exceeded the value of 30 A during 15 minutes, i.e. 900 s. However, a GIC having this duration is extremely rare according to FMI's statistics, Fig. 5.

The normal AC-magnetization current of a large transformer is roughly 1 A and the GIC can be two orders of magnitude larger than that. Therefore it is often claimed that the GIC is a potential risk to the transformers. In fact, the GIC and the AC-magnetizing current should not be directly compared, because the DC magnetizes the transformer core via the voltage drop the DC-component is able to cause in the winding. Besides, in determining the effects of the DC on the overall behavior of the transformer it is essential to take into account the hysteresis of the magnetic core material.

The power transformers used in the Finnish system are usually also to certain degree different from those used in other countries as will be described in section 6. The transformers of the Finnish transmission grid are usually three-phase

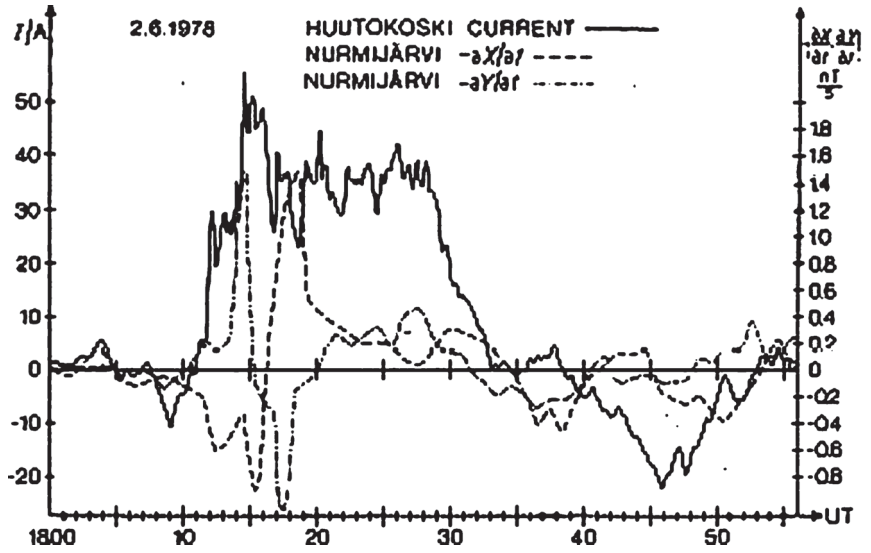


Figure 4. The GIC in Huutokoski and the variation of the time derivatives of the horizontal flux density components of the geomagnetic field at Nurmijärvi on June 2, 1978 (Elovaara et al. 1992)

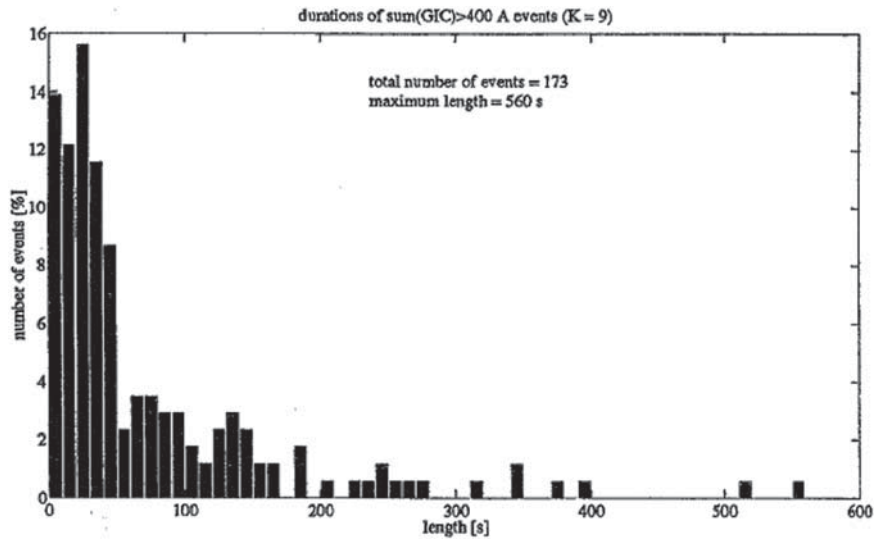


Figure 5. Distribution of the duration of those events during the years 1993–1999 where the sum of GIC's through all Finnish 400 kV transformer neutrals exceeded 400 A (only events K = 9 have been considered) (Pulkkinen et al. 2000)

three-winding full-wound five-legged core-form transformers in which the tertiary has the reactive power compensation reactors. However, the calculation of GIC-effects in such transformer is quite complicated and therefore the Finnish grid company has measured the behavior of its power transformers under DC-magnetization. This has been possible because the thermo elements used in the factory type tests of the transformers to measure the temperature rises in their constructional parts have been left in some of the transformers for the purpose of being available for potential future needs. Newer transformers have also fiberoptic temperature sensors in the windings.

The first DC-magnetization measurement was done for a 400 MVA transformer already in 1980 with following main findings (Pesonen 1982):

- the 10 minutes long magnetization with 100 A DC increased the no-load losses to the value 7 times the AC-value, i.e. 1400 kW, and the no-load apparent power to the value 48 times the AC-value (34 MVA); the reactive power consumption was 28 Mvar and a 50 A AC (nearly entirely third harmonic) was measured in the neutral
- even a 6 A DC increased the no-load losses by about 30% and amplified the no-load current by a factor of 4–6
- no gas formation was observed during and after the tests, and the thermal sensors did not indicate any large temperature rises
- the 400 kV phase-to-phase voltage of the DC-magnetized transformer was strongly distorted (total distortion 4 %) while the value without DC-magnetization was 0.5 %.

Because the constructional details of the power transformers used in Finland have changed over the years, this kind of test was repeated in 1999 with two power transformers connected in parallel in the 400 kV grid, (Lahtinen and Elovaara 2002). Figure 6 shows schematically the connection used in this test which was made under conditions which were favorable from a thermal point of view (ambient temperature −2 °C). In fact, the same DC was flowing through both transformers but because the DC-magnetization currents had opposite directions in individual transformers, the harmonic distortion of the 400 kV voltages in a real GIC-event had to be calculated analytically.

Figure 7 shows the temperatures in different parts of the transformer when the DC-magnetization current was increased step by step from 50 A to 200 A. It is obvious that the highest recorded internal temperature did not exceed 130 °C even though DC magnetization with a constant current had lasted more than 5 minutes. The corresponding temperature rise was less than 110 K and the time constant of individual temperature rises was about 10 minutes. The maximum winding temperature never reached critical values. No gas production was noticed during the test.

Under summer conditions and full load the initial temperature of the transformer would be higher than that applied in the tests, and internal temperatures of 170–180 °C could be reached. But this is not a problem because there is no cellulose type material in the vicinity of these spots. In the temperature range

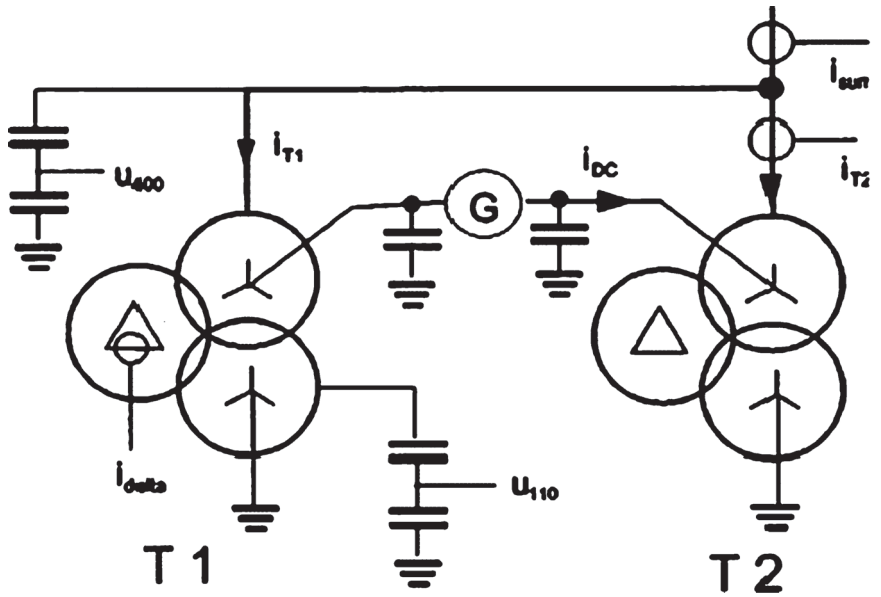


Figure 6. The half-cycle DC-magnetization tests in Toivila in 1999 for two 400/400/125 MVA, 410/120/21 kV, YNyn0d11-connected transformers (three-phase five-legged core-form constructions). (Lahtinen and Elovaara, 2002)

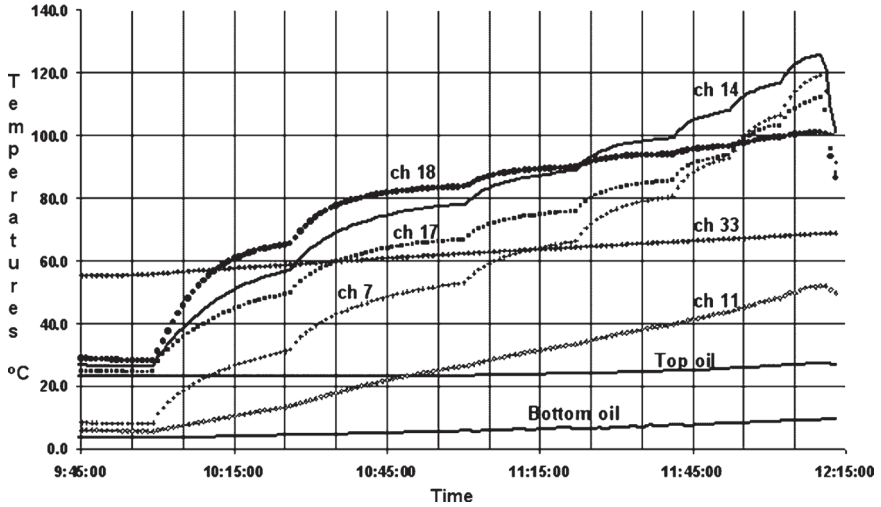


Figure 7. Temperature rises in one of the Toivila transformers under various DC-magnetizations in Toivila in 1999. The different curves correspond to the different recording points on and in the constructional parts of the transformer (ch 7: inside bottom yoke clamps; ch 11: winding support made of non-magnetizing material; ch 14: inside top-yoke clamp; chs 17 & 18: flitch plate; ch 33: mid-point of center phase iron core) (Lahtinen, 2002)

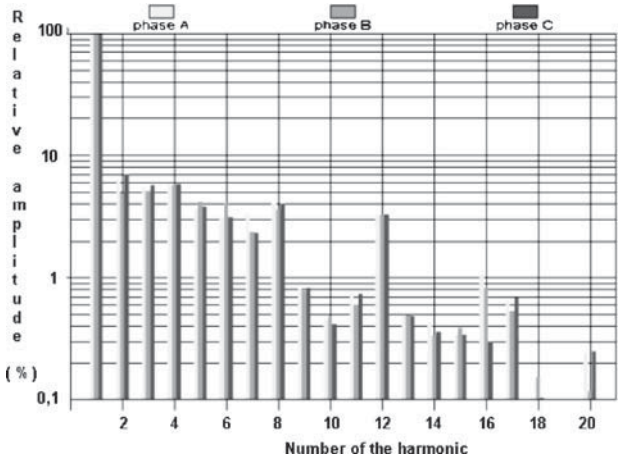


Figure 8. Simulated production of harmonic voltages in parallel operating 400 MVA transformers connected in the 400 kV grid when each transformer is magnetized with a continuous monopolar 100 A DC (total GIC to ground 200 A)

170–180 °C some transformer oil starts to dissolve so that the risk of gassing of oil exists, but the risk is small because a 200 A DC occurs approximately once in 50 years.

Figure 8 demonstrates that the voltage distortion would be very large in case of two half-cycle saturated transformers although the parallel transformers together tolerate a DC-magnetization of 200 A. This high harmonic content has effects also in the lower voltage networks. The effects are, besides extra losses, malfunctioning of some devices which are designed to work only under purely sinusoidal voltage conditions. However, such effects were not experienced during the test.

CO-OPERATION WITH THE FINNISH METEOROLOGICAL INSTITUTE

The Finnish transmission grid company Fingrid Oyj and earlier Imatran Voima Oy have over last 25 years carried out three GIC-study projects in co-operation with the Finnish Meteorological Institute (FMI) with the following objectives:

- development of a GIC-distribution calculation model based on the time-variations of the magnetic and geoelectric fields in the ground
- estimation of the statistical distribution of the GIC magnitudes in the 400–220 kV grid in Finland
- determination of the effects of different kinds of space current distributions on GIC distributions in the "man-made grids" on the earth

In connection with these modeling studies special measurement campaigns were organized.

As a result we have now GIC-models based on realistic ionospheric current models and on spatially inhomogeneous geoelectric fields available in Finland. Geoelectric fields produced by realistic ionospheric model currents were applied to realistic inhomogeneous multi-layer Earth models (Mäkinen 1993, Pulkkinen et al. 2000). In the latest co-operation project the geoelectric field was calculated from the magnetic data of the BEAR and IMAGE magnetometer arrays. Although this study included 161 cases, 11 of which during $K = 9$ and 20 during $K = 8$, the worst case has probably not yet been seen. In spite of large and extended magnetometer arrays, the registration network is not dense enough to cover well all parts of Finland. It was concluded that the geoelectric field distributions vary greatly from event to event and so does the magnitude of the GIC's through the transformers.

As a consequence Fingrid possesses quite a representative data set of the induced electric field at the surface of the Earth. Further, Fingrid has special software to determine the GIC distribution in the transmission grid on the basis of the geoelectric field strengths at the surface of the ground. This information has been used to determine the risks GIC's might cause to the Finnish transmission grid. It is also possible to study the effect of different GIC reduction means if this should become necessary.

Figure 9 shows occurrence statistics of different GIC magnitudes in the Finnish 400 kV transformer neutrals. The bars indicate how many times per year the sum of the absolute values of GIC's in all transformer neutrals exceeds a given number. For example, the sum of 1000 A was exceeded about four times per year when the number of transformers was about 60 and there where no series capacitors

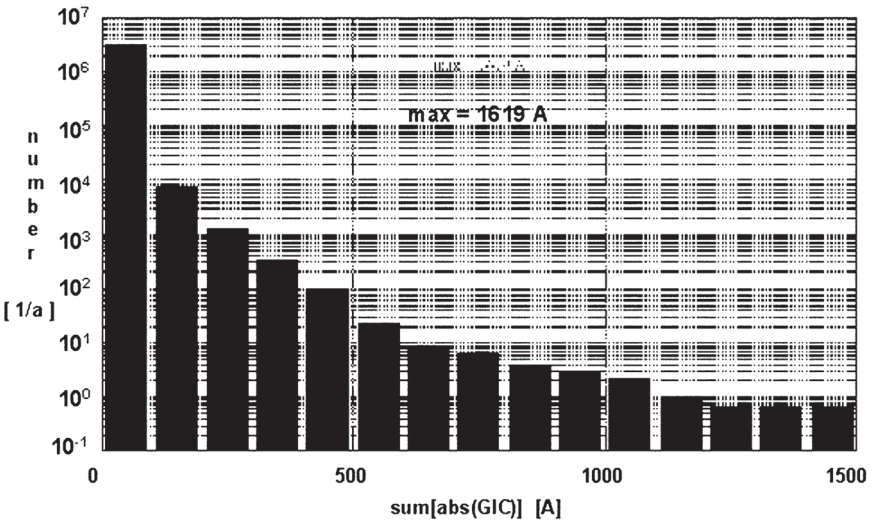


Figure 9. Annual occurrence probability of different GIC-magnitudes in all 400 kV transformer neutrals in Finland on the basis of theoretical simulations. Grid as in year 1999. (Pulkkinen et al. 2000)

Table 1. GIC-values (in A) exceeded *n* times per year at certain substations (Mäkinen 1993)

Identification	*n* [1/a]	Substation			
		Pirttikoski	Huutokoski	Rauma	Yllikkälä
Normal year	100	20	3	18	~ 5
	10	41	19	39	11
	1	79	32	71	18
	0,1	138	54	119	28
Very active year	100	63	27	53	15
	10	112	46	97	24
	1	195	73	168	37
	0,1	> 200	114	> 200	53

in the north-south-directed 400 kV lines. Corresponding statistics have also been calculated for various substations on the basis of older and less accurate data. Table 1 gives results from some substations of the grid of 1991 (Mäkinen, 1993).

FINNISH VIEW ON THE EFFECTS OF GIC IN GRID OPERATION

Although very high GIC's have occurred in Finland, prior to 2003 no GIC-related grid problems were noted in Finland. Even the severe geomagnetic storms of 1978 and 1991 did not have negative consequences for the Finnish grid. No disturbances, outages or equipment failure were reported nor has been noted that the Finnish power transformers had started to produce more gas during or after the GIC events. However, in Sweden some unintended and unnecessary trippings of protection. relays took place (Elovaara et al. 1992). The erratic tripping has usually been caused by old sensitive earth-fault relays, which are of electro-mechanical type in Sweden, with constant tripping time characteristics and without harmonic blocking features. In Sweden tripping problems have been encountered also in connection with some static relays. The ultimate reason for erratic trippings has been the harmonic content of the current and not the saturation of the iron core of the current transformers, because the induction in older current transformers is low due to the small burden. The latest of this kind of tripping took place in city of Malmö in October 2003 causing a black-out in Malmö (Pulkkinen et al. 2005).

It is also reported that GIC's caused "unstable" operation of Swedish generator voltage controllers. On certain occasions series capacitors were blocked because of the incorrect operation of their special protection systems. However, these events did not lead to system outages.

In Finland the protection is arranged partly in a different way. The current transformers feeding the overcurrent relays are usually linearized, i.e. they have a small gap filled with insulating oil which prohibits the saturation of the current transformers. Moreover, the earth-fault relays are not requested to be able to disconnect a fault having a fault resistance higher than 500 Ω (corresponds to a current 460 A),

and the relays have a harmonic filter which is able to filter out the low-frequency even and odd harmonics. However, during the geomagnetic storm of October 30, 2003, an erroneous relay tripping occurred in Northern Lapland. It was the first time that GIC's had negative consequences for the Finnish transmission grid. Northern Norway was then fed from Finland, and the tripping caused a local areal black-out there when the feeding line was disconnected. The reason for the tripping was a mistake in the configuration of a new type of digital relays: the relay reacted in an undesired way with respect to the harmonic content of the load current.

REASONS FOR THE POSITIVE FINNISH EXPERIENCES

The internal construction of the power transformer is critical for its reaction to the DC magnetization. The basic questions are: how easy is it for the DC flux created by the GIC to flow in the transformer, and where does it flow. The concepts used to describe the properties of a transformer are selected on the basis of the AC operation, they are not appropriate to describe the transformer's behavior under DC magnetization. If AC concepts are used, the zero-sequence impedance of the transformer is important, and this impedance has a completely different magnitude at power frequency than at DC. The short-circuit impedance plays a role, too.

Zero-sequence impedance means the impedance met by the current in a three phase AC system, when a similar current (with similar amplitude and phase angle) is fed in all three phases. In an AC system in symmetrical conditions, the zero-sequence impedance does not play a role, but in faults where ground is involved, also a zero-sequence circuit has to be taken into account. If direct current is fed in the transformer, the DC-flux takes in principle the same path as the zero-sequence AC-flux but when the core or part of it is saturated, also the leakage flux path starts to describe the behavior. The zero-sequence impedance of a transformer (z_0) depends on the construction of the iron core, on the connection of the windings and on the earthing of the neutral points of the windings. A low z_0-value usually means that the zero-sequence flux has a return path in the transformer which consists of magnetic material.

Figure 10 illustrates the basic constructional features of different types of transformers (winding types, construction of the core). One observes that in core-form transformers the zero-sequence flux like the flux component caused by the direct current cannot get a closed low-reluctance path (or it has only a path with small area) and this flux remains relatively small. However, if the core has a shell-form, a low-reluctance path is available, and the zero-sequence flux assumes a high value. The core of a single-phase transformer is more often a shell-form than a core-form type.

Leakage flux (stray flux) denotes the part of the magnetic flux in a transformer which does not flow in the core. Its magnitude is determined by the distance between the windings wound around the same leg and by the width of the windings. Part of this flux penetrates into or out of the core near the places where the windings end. The leakage flux creates extra losses and temperature increase in places where

it penetrates into or out of the iron core or where it meets conductive parts. The magnitude of the leakage flux is given by a quantity called short-circuit impedance (z_{sc}). If a transformer has separate primary and secondary windings, its short-circuit impedance varies – depending on the rating of the transformer – from a few percent up to 20 percent of the value calculated from the rated voltage and current of the transformer.

However, one can construct a transformer such that it is not full-wound; the low voltage winding is instead part of the high voltage winding (Fig. 10a,b). In this case the leakage flux is very low and the short-circuit impedance is practically zero, $z_{sc} \approx 0$. This construction is called auto-transformer. An auto-transformer is lighter, lower and cheaper than its full-wound alternative, but the auto-transformer cannot limit the fault currents because of its low z_{sc}-value. Cheaper auto-transformers are normally used, but in conditions where fault-current related problems are very difficult to solve, full-wound transformers can be used to keep the fault-current levels low. From a GIC-point of view it is important that the flux of a saturated transformer takes the path of the leakage flux. If the transformer designer has

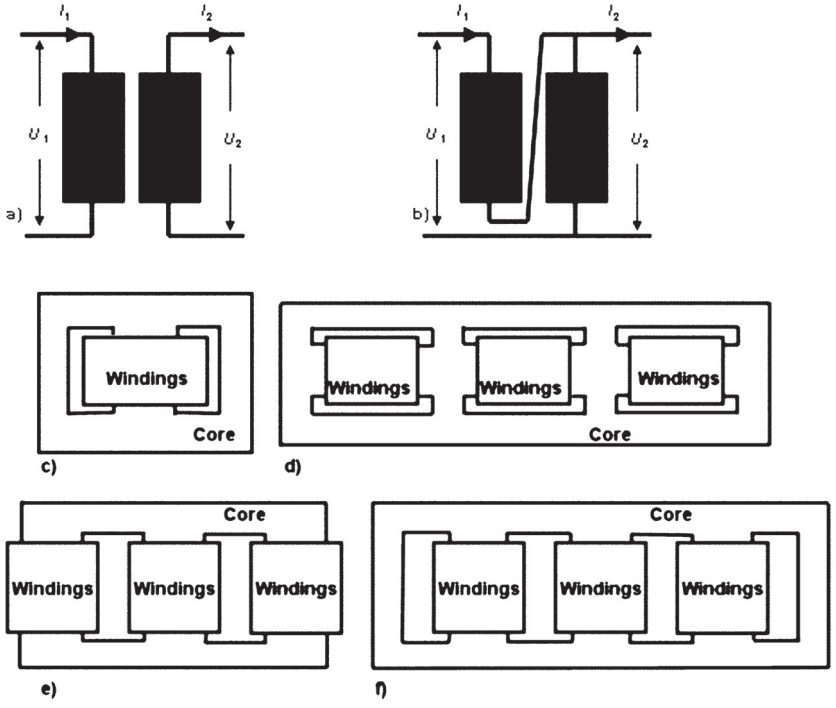


Figure 10. Different power transformer constructions. a) Full-wound transformer. b) Auto-transformer. c) Single-phase transformer with shell-form core. d) Three-phase transformer with shell-form core. e) Three-phase transformer with core-form core (three legs). f) Three-phase transformer with core-form core (five legs, the outmost legs have only a small area)

not paid attention to stray flux effects high GIC situations might result in bad surprises. (The assumed GIC-related transformer failures have most often occurred in auto-transformers made of single-phase units.)

The sensitivity of different transformer types to the effects of GIC's is sometimes compared and ranked according to the iron area which is available to the DC-flux generated by the windings (McNutt 1990). Table 2 ranks the different transformer types in this way. A value of 0 means the most insensitive and a value of 1 the most sensitive construction. Clearly, the transformer types used in Finland have a very low GIC sensitivity.

The reasons for the excellent Finnish experiences (compared to other countries) with the reaction of power transmission grids to GIC conditions are probably the following.

- When the 400–110 kV transformer neutral points are earthed, the magnitudes of the earth fault currents are actively limited by inserting current limiting coils in the neutral branches of transformers. The resistance of the current limiting coil limits the magnitudes of the GIC's.
- The Finnish power transformers were specified and manufactured so that the magnitudes of the fault currents are limited. Both their short-circuit and zero-sequence impedances are high. In practice only three-phase single unit full-wound transformers are used. Moreover, their short-circuit impedance is 20%, i.e. the transformers are designed for relatively large leakage fluxes. The core form is chosen such that the zero-sequence flux meets a high reluctance (three-phase YNyn0d11-connected 5-legged or YNyn0-connected 3-legged core type).
- The heat-run tests made for the Finnish transformers include special tests where the magnitude of the leakage flux exceeds the rated levels. The temperature rises at critical points in the core and windings are monitored and recorded. Dangerous hot spots in the constructional parts can thus be eliminated.
- The sensitive earth-fault protection relay system is set to operate at the fault-resistance of 500 Ω, and harmonic blocking is used due to the inrush-currents of transformers.
- To improve the transmission capacity series capacitors are installed on the lines going to the north and towards Sweden. This has also decreased the magnitude of GIC's.

Table 2. Different transformer types and their sensitivity to the effects of GIC's (McNutt 1990). p.u. = per unit, i.e. 1 p.u. = 100 per cent

Type of the transformer	Nm of phases	Nm of legs	z_{sc} [p.u.]	z_0/z_{sc} [p.u.]	GIC sensitivity
Full-wound, core-form	3	3	free	5	0
Full-wound, core-form	3	5	free	1–∞	0.24–0.33
Auto-transform., shell-form	3	3	~ 0	−1	0,5–0,67
Full-wound, shell-form	1/3	2/7	free	−1	1
Auto-transform., shell-form	1	2/3	~ 0	1	1

GIC'S AND RISKS IN GRID OPERATION

During a heavy geomagnetic storm the distortion level of the voltages can be very severe also in Finland. The possibility of a component failure cannot be completely excluded, but the risk is low. Voltage and reactive power balance control problems constitute a risk. A 400 MVA transformer consumes at 200 A DC-magnetization about 55 Mvar reactive power. Once per year 1200 A altogether is flowing via the transformer neutrals to and from the earth which would require that the grid can supply under this GIC event 330 Mvar extra reactive power. For extreme conditions the uncertainties of the GIC estimation have to be taken into account. A safety factor of 2 would mean 660 Mvar reactive power capacity. At present there are enough reactive power resources available for this during about 95% of the time. This means from the system point of view that a voltage collapse and system black-out could occur roughly once in 20 years if an extreme case takes place. However, on the same time the new line investments decreasing the line lengths and the new series capacitors to be installed on the longest uncompensated lines will reduce the GIC's.

More attention should be paid to the properties of the modern digital relays. The specifications and factory acceptance tests shall be designed so that unexpected trip commands are not caused by the relatively rare and abnormal system conditions like those created by GIC-flows in the system.

As a whole the risk caused by GIC's is acceptable in the Finnish 400–220 kV grid, and warning systems or other mitigation methods are not necessary. A necessary condition is, of course, that the components to be used or connected in the Fingrid's future system have basic features similar to those in Fingrid's present system. If for example large single phase transformers will be connected in the system, the GIC-risk of these transformers has to be studied separately and suitable mitigation methods like a neutral point capacitor or a neutral point resistor have to be used, if the GIC-risk is unacceptable (Bolduc et al. 2005). In the design of the reactive power reserves the effect of the GIC's shall be remembered, and these reserves have to be maintained also in future conditions at an appropriate level. More attention should perhaps be paid to withstand the capability of high voltage equipment concerning high harmonic currents and voltages.

REFERENCES

Bolduc, L.: GIC observations and studies in the Hydro-Quebec power system. Journal of Atmospheric and Solar-Terrestrial Physics, 64, 1793–1802 (2002)

Bolduc, L., Granger, M., et al.: Development of a DC current-blocking device for transformer neutrals. IEEE Transactions on power delivery, 20, 163–168 (2005)

Elovaara, J., Lindblad, P., et al.: Geomagnetically induced currents in the Nordic power system and their effects on equipment, control, protection and operation. Paper 36–301 of CIGRE Session 1992, Paris, 10 p (1992)

Lahtinen, M., Elovaara, J., GIC occurrences and GIC tests for 400 kV system transformer. IEEE Transactions on Power Delivery, 17, 555–561 (2002)

McNutt, W.: The effect of GIC on power transformers. Paper presented in IEEE PES Summer meeting July 17, 1990 in Minneapolis during the panel session "geomagnetic storm cycle 22: Power system problems on the horizon, pp. 32–37 (1990)

Molinski, T.S.: Why utilities respect geomagnetically induced currents. Journal of Atmospheric and Solar-Terrestrial Physics, 64, 1765–1778 (2002)

Mäkinen, T.: Geomagnetically induced currents in the Finnish power transmission system. Geophysical Publications 32, Finnish Meteorological Institute, Helsinki, 101 p (1993)

Pesonen, A.J.: Discussion in IEEE Transactions on Power Apparatus and Systems, 101, pp. 1453–1454 1982)

Pulkkinen, A., Viljanen, A., et al.: Large geomagnetically induced currents in the Finnish high-voltage power system. Reports, 2, Finnish Meteorological Institute, Helsinki, 99 p (2000)

Pulkkinen, A., Lindahl, S., et al.: Geomagnetic storm of 29–31 October 2003: Geomagnetically induced currents and their relation to problems in the Swedish high-voltage power transmission system. Space Weather, 3, S08C03,doi:10.1029/2004SW000123 (2005)

INDEX