LUND UNIVERSITY

Solar Wind and Geomagnetic Activity - Predictions Using Neural Networks

Gleisner, Hans

2000

Solar Wind and Geomagnetic Activity

Predictions Using Neural Networks

Hans Gleisner

Lund Observatory 2000

SOLAR WIND AND GEOMAGNETIC ACTIVITY

Predictions Using Neural Networks

Hans Gleisner

Thesis for the degree of Doctor of Philosophy
(Avhandling för doktorsexamen)
Lund Observatory
2000

LUND UNIVERSITY

Abstract

This thesis shows how artificial neural networks (ANNs) can be applied to predict geomagnetic activity from solar-wind data. It introduces and summarizes five papers where development of ANN models are reported, and where predictions of geomagnetic activity are discussed. The studies cover geomagnetic disturbances characterized by global-scale indices and geomagnetic variations that are locally observed. Several types of ANN are utilized: time-delay networks, radial-basis function networks, and partially recurrent networks. Methods and procedures that can be applied to forecasting based on real-time data are emphasized.

The first, introductory part of the thesis begins with an outline of the solar-terrestrial space environment, focusing on the processes and circumstances that play a role in the generation of geomagnetic disturbances. The ANN methods that have been used in the present studies are then briefly described, and put into a wider context of other modeling and prediction techniques. The introductory part of the thesis ends with short summaries of the five papers, which are reprinted in the second part of the thesis.

Paper I describes predictions of the ring-current index Dst from solar-wind data. It is shown that magnetic storms can be predicted with time-delay networks. The influence of the solar-wind input sequence length on the different storm phases is discussed.

Papers II, III, and V present studies of the solar wind-auroral electrojet relations using time-delay networks [II,III], and recurrent networks [V]. The relative importance of different solar-wind variables and coupling functions is studied in paper II. Paper III describes the influence of the Dst level on the modeled solar wind-auroral electrojet relations. In paper V, the capabilities of recurrent networks are evaluated, and compared to time-delay networks.

In **paper IV**, predictions of locally observed geomagnetic variations are studied. The daily, quiet-time variations are modeled with radial-basis function networks that account for annual and solar-cycle modulations. The horizontal magnetic disturbance field is modeled with gated, time-delay networks taking local time and solar-wind data as input.

Key words: solar wind, magnetosphere, space weather, geomagnetic activity, geomagnetism, ring current, auroral electrojet, neural networks, forecasting.

Hans Gleisner, Lund Observatory, Box 43, SE–221 00 Lund, Sweden

LUNFD6/(NFAS 1020)/1–92/(2000)

ISBN 91-7874-054-1
Printed by KFS AB, Lund 2000

To Katta and Theo

Research Articles

This thesis introduces and summarizes the following research articles:

I. Gleisner H., Lundstedt H., and Wintoft P.:
Predicting geomagnetic storms from solar-wind data using time-delay neural networks
Annales Geophysicae, **14**, 679-686, 1996.

II. Gleisner H. and Lundstedt H.:
Response of the auroral electrojets to the solar wind modeled with neural networks
Journal of Geophysical Research, **102**, 14269-14278, 1997.

III. Gleisner H. and Lundstedt H.:
Ring current influence on auroral electrojet predictions
Annales Geophysicae, **17**, 1268-1275, 1999.

IV. Gleisner H. and Lundstedt H.:
A neural network-based local model for prediction of geomagnetic disturbances
Journal of Geophysical Research, in press (8 pp), 2000.

V. Gleisner H. and Lundstedt H.:
Auroral electrojet predictions with dynamic neural networks
Journal of Geophysical Research, submitted (8 pp), 2000.

Contents

1 Introduction

On some days, the geomagnetic field at the Earth's surface undergoes smooth and regular variations, while on other days it is more or less disturbed. The irregular disturbances that are superposed on the regular, daily variations, are a consequence of interactions between the *solar wind* and the Earth's magnetic field. Large-scale electrical currents, flowing in near-Earth space and connecting to the upper atmosphere, are powered by the solar wind and react quickly to any variations in the solar-wind conditions. The resulting magnetic disturbances, or the *geomagnetic activity*, provide information on the global state of the near-Earth space which is not readily available by other means.

Although the most characteristic features of geomagnetic activity - the *magnetic storm* and the *magnetic substorm* - can be related to prior solar-wind conditions, the relations are seldom straightforward. The Earth's *magnetosphere* does not passively transform a solar-wind input into a geomagnetic-activity output. Powerful dynamical processes taking place in the near-Earth space add their own signatures to the geomagnetic disturbance field. Different methods have been used to explore the relations between the solar wind and magnetic storms and substorms: statistical correlative methods, linear filters, nonlinear filters, and low-dimensional, nonlinear systems. This thesis shows how *artificial neural networks* (ANNs) can be applied to predict geomagnetic activity from solar-wind data. It introduces and summarizes five papers where the development of ANN models are reported, and where predictions of geomagnetic activity are discussed.

That geomagnetic activity is somehow related to the Sun's activity has been known, or at least suspected, for nearly 150 years. This early understanding rested largely on common periodicities: the 11-year sunspot cycle and the 27-day solar rotation period were detected in the geomagnetic records.[1,2] What actually transmitted the influences from the Sun was at the time totally unknown. Over the years, several attempts were made to explain the solar influences by matter flowing out from the Sun, but the role played by the solar wind was not confirmed until the first in situ measurements were made beyond the Earth's protective magnetic fields. In fact, the very existence of a permanent solar wind was debated at the time of the first spacecraft travels.[3] It soon became apparent that the Sun's magnetic fields, extending far out into interplanetary space, play a fundamental role in controlling geomagnetic activity.

Today, 40 years after the first space probe encounters with the solar wind, our ideas about the solar-terrestrial space environment, the empirical data available, and the methods used to analyze them have developed considerably. A recent development is the establishment of real-time monitoring systems in space. The Sun is now being observed twenty-four hours a day with spaceborne telescopes, and the solar wind is continuously monitored from a location 1.5 million kilometers upstream in the solar wind. We can now watch how solar-wind disturbances are triggered by explosive

events on the Sun and then follow how they sweep past the Earth a few days later, generating disturbances in the near-Earth space environment, lighting up the skies with auroral displays, and sometimes causing problems to technical systems. A new term, *space weather*, has been coined to describe certain aspects of the solar-terrestrial space environment, particularly those that have consequences for technical systems on ground or in space. This focus on the consequences of the solar-terrestrial space conditions has made the development of forecasting capabilities an important concern, and has stressed the significance of real-time monitoring of the solar-terrestrial space environment. Solar monitoring provides an indication two or three days ahead that disturbances are on their way, whereas solar-wind monitoring gives us a more detailed view of arriving solar-wind disturbances up to an hour ahead. Translation of the observed solar-wind conditions into useful forecasts of upcoming disturbances requires predictive methods that are accurate and computationally efficient. This thesis will, hopefully, demonstrate the usefulness of neural networks, both as a means to map the solar wind-geomagnetic activity relations and as a tool for short-term forecasting of geomagnetic disturbances and related phenomena.

The thesis is based on five papers that have been, or will be, published in scientific journals. The first, introductory part of the thesis is intended to provide a background for the non-specialist. It begins in chapter 2 with an outline of the solar-terrestrial space environment, focusing on the processes and circumstances that play a role in the generation of geomagnetic disturbances. The neural networks that have been used in the present studies are briefly described in chapter 3, and in chapter 4 they are put into a wider context of other modeling and prediction techniques. The introductory part of the thesis ends with short summaries of the five papers on which the thesis is based, and the papers are then reprinted in the second part of the thesis.

2 The Solar-Terrestrial Space Environment

The continuous outflow of matter from the Sun's atmosphere fills interplanetary space with a thin, ionized gas that is traveling away from the Sun with velocities of several hundred kilometers per second. This is the *solar wind* to which all solar-system objects are exposed. The solar wind produces planetary aurorae and cometary plasma tails, and modulates the intensity of cosmic rays. This chapter describes the solar wind and the Earth's magnetic field, and some consequences of their interactions. One of the most important consequences is the formation of a magnetosphere where large-scale electrical currents are generated - currents that are the source of geomagnetic activity.

2.1 The magnetically active Sun

The Sun is an ordinary, hydrogen burning star with a mass of $2 \cdot 10^{30}$ kg, a radius of 700,000 km, and a total radiative output of $4 \cdot 10^{26}$ W. The bulk of this radiation leaves the Sun from the *photosphere*, which is a thin layer at the base of the solar atmosphere with a thickness of only a few hundred kilometers. Here, the Sun's convective interior is cooled by radiation into space. Above the photosphere lies the *chromosphere* and the extended solar *corona*. The temperature increases from a minimum around 4,500 K in the upper photosphere, to several tens of thousands of K in the chromosphere, and then rise rapidly to more than a million K in the corona.

To the unaided eye, the Sun appears to be essentially unchanging. The Sun is, however, a magnetically active star. Magnetic fields that rise from within the Sun become visible as strong surface fields in the photosphere and as complicated magnetic structures in the chromosphere and in the corona. All aspects of solar activity are related to the magnetic fields: *sunspots* are due to magnetic field concentrations visible in the photosphere, *prominences* are clouds of relatively cool gas supported by magnetic fields in the solar corona, and the violent *flares* and the huge *coronal mass ejections*, which are eruptive events on the Sun with important terrestrial consequences, are driven by conversion of magnetic energy into kinetic energy, heat and radiation.

The number of sunspots visible on the Sun varies regularly with a period around 11 years. This is the well-known sunspot cycle, which is just one of the manifestations of the Sun's magnetic cycle. The same periodicity is found in most measures of solar activity: the occurrence of flares and coronal mass ejections, energetic-particle fluxes, UV- and X-ray fluxes. In fact, the global structure of the corona changes radically from solar-activity minimum to maximum, in concert with the magnetic cycle.

The influence of the solar magnetic field reach far beyond the Sun itself. As the field is dragged out by the solar wind, it pervades the interplanetary space and interacts with all solar-system bodies, including the Earth. The Sun's magnetic field governs such terrestrial phenomena as aurorae and geomagnetic activity, and modulates the intensity of cosmic rays.

2.2 The solar wind

2.2.1 Basic properties

The high temperatures in the Sun's corona create a pressure which is not completely balanced by the Sun's gravity. The result is an acceleration of the solar atmosphere into a solar wind[4], which soon reaches velocities of several hundred kilometers per second. This velocity is preserved throughout the planetary system while the solar-wind density gradually decreases away from the Sun. Near the Earth's orbit, the average velocity and electron number density is 450 km/s and 7 cm^{-3}, respectively, and the temperature has dropped to roughly 10^5 K.

The solar wind gas is fully ionized, and consists of electrons and ions in numbers that make the gas electrically neutral on macroscopic scales. A gas in this state is referred to as a *plasma*. The electrical conductivity of a plasma is very high and, according to Faraday's law, the magnetic field and the plasma are forced to move together. This fact is often described in terms of the magnetic field being "frozen" to the plasma. As the solar wind is accelerated and flows out into interplanetary space, it drags the solar magnetic fields along with it. Near the Earth's orbit, the magnetic flux density is on average 6 nT, which means that the kinetic energy density is one or two orders of magnitude larger than the magnetic energy density. Extraction of only a fraction of the kinetic power of the solar wind plasma impinging on the magnetosphere (about 10^4 GW) is sufficient to drive the magnetospheric processes that generate geomagnetic activity and aurorae.

Averages alone do not express the fact that the solar-wind conditions are extremely variable. Steady, spatial solar-wind structures that rotate with the Sun past the nearly stationary Earth cause a 27-day recurrency of the solar-wind conditions (Fig. 1a). Episodic ejections of matter from the Sun cause transient, high-speed plasma flows that drive shock waves in the solar wind, and that can have unusual magnetic-field characteristics (Fig. 1b). Superposed on the co-rotating structures and the transient solar-wind flows, are a rich variety of smaller-scale inhomogeneities and wave motions.

2.2.2 Co-rotating structures

One of the earliest recognized features of the interplanetary magnetic field is its organization into *magnetic sectors*. For one or two weeks at a time the field has a predominant polarity, pointing either toward or away from the Sun. The solar-wind speed and density undergo systematic variations in association with the magnetic sector structure. At sector boundaries, the wind speed is usually low, 300-400 km/s. The boundary is followed by a compression of the solar-wind plasma and a steep rise in velocity, up to 600-700 km/s. The velocity and density structure can be described as a high-speed stream catching up a slower flowing plasma, thus creating an *interaction region* which coincides with the magnetic sector boundary. The structuring into sectors usually dominates the solar wind during the declining and minimum phases of the solar-activity

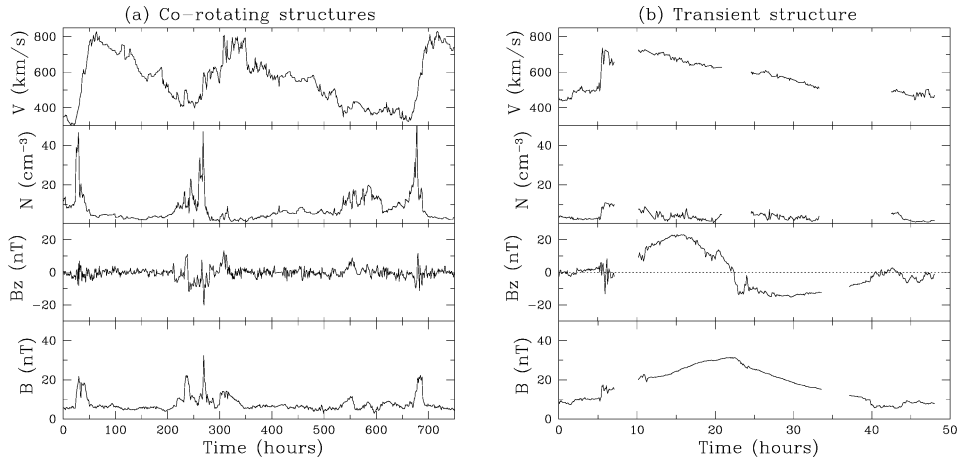

Figure 1. Solar wind speed, proton number density, and the north-south component (B_z) and magnitude (B) of the interplanetary magnetic field, during (a) 31 days in June-July 1974 when the solar wind was dominated by co-rotating structures, and (b) 2 days in January 1988 during the passage of a transient solar-wind disturbance caused by a CME.

cycle, when the solar magnetic fields have a relatively simple appearance.[5] A typical example is shown in Fig. 1a.

2.2.3 Transient structures

During solar-activity maximum, the solar magnetic fields are normally very complex. Often, no clear sector structure is developed and the solar wind is dominated by transient disturbances,[5] of which some are associated with coronal mass ejections (CMEs). Although our understanding of CMEs still is incomplete, it is clear that they originate in disruptions of coronal magnetic field structures not previously participating in the solar wind expansion.[6] A large-scale magnetic structure that has been in a closed, static equilibrium for days or weeks, suddenly loses equilibrium and rapidly expands outward into interplanetary space. The solar-wind signature of a CME is often a shock followed by a high, but monotonically decreasing speed, and a strong magnetic field that is steady or slowly rotating. An unusually well-behaved example of a transient solar-wind structure, sometimes referred to as a *magnetic cloud*, is shown in Fig. 1b.

2.2.4 Geoeffectiveness of the solar wind

Significant geomagnetic activity occurs only when the interplanetary magnetic field (IMF) has a southward component, anti-parallel to the dayside geomagnetic field. The only known explanation for this relationship is *magnetic reconnection*, i.e. the process

by which two plasmas with imbedded magnetic fields interact with each other. In the thin current sheet that develops between two volumes of plasma with different magnetic field direction and strength, the frozen-in field concept breaks down and magnetic field lines can merge across the boundary. Every time the IMF turns southward, efficient reconnection starts between the solar wind and the magnetosphere, driving the processes that generate geomagnetic activity and aurorae. The term "geoeffectiveness" is sometimes used to describe the efficiency of the solar wind in this respect. A strong, southward magnetic field together with a high wind speed and density, is the most geoeffective combination of solar-wind parameters.

On average, the interplanetary magnetic field lines are parallel to the ecliptic plane and the north-south component, B_z, is close to zero. The IMF is, however, extremely variable and deviations from the in-ecliptic field direction are a common occurrence. Although the north-south component often is significant, it is seldom steady and only rarely has the same sign for several hours. During the passage of a solar-wind transient associated with a CME (Fig. 1b), the solar-wind conditions often exhibit unusual characteristics, particularly the IMF which can be relatively steady, stronger than normal, and slowly rotating for a day or more. The most intense magnetic storms are almost invariably associated with CMEs.[7]

A substantial north-south IMF component, B_z, can also occur as a result of corotating solar-wind structures (Fig. 1a). The interaction regions are associated with high dynamic pressures and a fluctuating north-south IMF component of higher-than-average field strength. The frequently occurring weak and moderate magnetic storms are often coincident with the passage of sector boundaries.[7] Within the high-speed flows following a sector boundary one can often observe Alfvén waves that produce continuous substorm activity and that prolong the decay of magnetic storms back to quiet-time values.

2.3 The near-Earth space

2.3.1 The geomagnetic main field

The geomagnetic main field is generated in the Earth's fluid outer core with local contributions from magnetized regions in the Earth's crust. To a first approximation, the field can be described as a magnetic dipole. More refined models are derived from observational data, usually in the form of an Earth-centered spherical harmonic expansion of the magnetic scalar potential. At the Earth's surface, the first-order dipole approximation is on average able to represent about 90% of the field.

The present-day geomagnetic field strength is 31 μT (0.31 G) at the Earth's surface near the magnetic equator, and twice that at the poles. For epoch 2000 the northern pole of the best fitting geocentric dipole - the *geomagnetic pole* - is located near the northwestern tip of Greenland (79.5° N, 71.7° W). The location where the geomag-

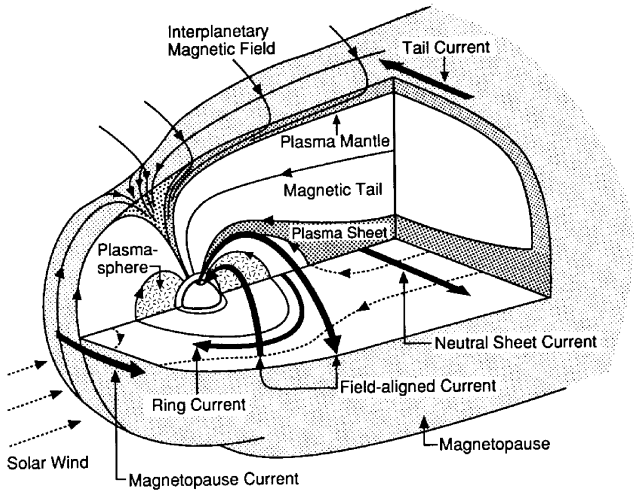

Figure 2. The magnetospheric structure showing magnetic fields, plasma regions, and large-scale electrical current systems. From *Introduction to Space Physics*, Kivelson and Russell (eds.), Cambridge University Press, 1995.

netic field is fully vertical - the *magnetic dip pole* - is currently about 700 km west of the dipole position.

The geomagnetic field exhibits variations on almost all time scales. An important distinction is made between *secular variations* caused by slow changes of the main field, and *transient variations* caused by electrical currents flowing in space. The term "geomagnetic activity" refers only to the latter. The secular variations must, however, often be considered as they slowly change the reference level used to define the transient field disturbances.

2.3.2 The Earth's magnetosphere

The Earth's magnetic field is constantly exposed to the solar wind. The field is strong enough to deflect the solar-wind plasma, but is at the same time significantly distorted by the interactions. A vast magnetic cavity, the magnetosphere, is formed. Only the inner part of the magnetosphere, out to 5 or 6 Earth radii (R_E), remains approximately dipolar. Beyond that, the magnetosphere more or less resemble a huge windsock (Fig. 2). It extends out to ten R_E in the sunward direction, and hundreds of R_E in the opposite direction, where magnetic field lines from the polar caps are dragged out into a long tail, usually referred to as the *magnetotail.*

The magnetospheric boundary to the solar wind, the *magnetopause*, consists of thin sheets of electrical currents with a typical thickness of only 500 to 1000 km.

The inner magnetospheric boundary has a very different appearance. It consists of a gradual transition down to the *ionosphere*, which is the partially ionized regions of the atmosphere at a height above 90 km. The ionosphere plays an important role for the magnetosphere. It provides an additional source of plasma, mainly protons and singly charged helium and oxygen, and it forms an electrically conducting shell at the base of the magnetosphere through which horizontal electrical currents can flow.

As a consequence of the magnetospheric structure, the geomagnetic disturbances have widely different characteristics at high and low latitudes. The *auroral oval* marks the boundary between the polar cap, which is magnetically connected to the magnetotail and the solar wind, and the low-latitude regions, where the magnetic fields are dipolar and the field lines close on Earth. Within the polar caps and the auroral ovals, the geomagnetic activity is predominantly caused by electrical currents flowing in the ionosphere, at a height of 90 to 130 km. Concentrations of horizontal ionospheric currents, commonly carrying more than 10^6 A, are referred to as *auroral electrojets*. The ionospheric currents are fed by *field-aligned currents* (Fig. 2) which transmit stresses from the solar wind-magnetosphere interactions. Unlike the low-latitude magnetic activity, the geomagnetic disturbances at high latitudes reflect physical processes in the magnetotail and in regions of space where the solar wind-magnetosphere interactions take place.

At low- and mid-latitudes, well equatorward of the auroral ovals, it is normally variations of the *ring current* (Fig. 2) that dominates the geomagnetic activity, although the other magnetospheric currents also contribute to the ground-level disturbances. The ring current is a permanent feature of the outer radiation belt.[8] It is formed by ions and electrons in the 20-200 keV range, located between 2 to 7 R_E, that are drifting under the influence of the inhomogeneous magnetic field.

2.3.3 Convection and substorms

The picture presented above is that of an "average" magnetosphere. However, the internal magnetospheric dynamics and the ever-changing solar wind only rarely, if at all, allow the magnetosphere to settle into a steady state. The dynamic behaviour of the magnetosphere is often described in terms of quasi-steady *magnetospheric convection* and transient *magnetospheric substorms*.

In the magnetosphere, any electric field $\mathbf{E}$ with a component orthogonal to the magnetic field $\mathbf{B}$ will drive a plasma flow. Two quasi-steady electric fields dominate the interior of the magnetosphere, the *co-rotation electric field* and the *cross-tail electric field*, corresponding to two modes of large-scale circulation: an inner region co-rotating with the Earth and an outer region circulating under the influence of the solar wind. The term magnetospheric convection refers to the latter. Unlike the co-rotation E-field, the cross-tail E-field is highly variable. The potential difference across the magnetotail is mapped along field lines down to ionospheric levels where it can be measured by satel-

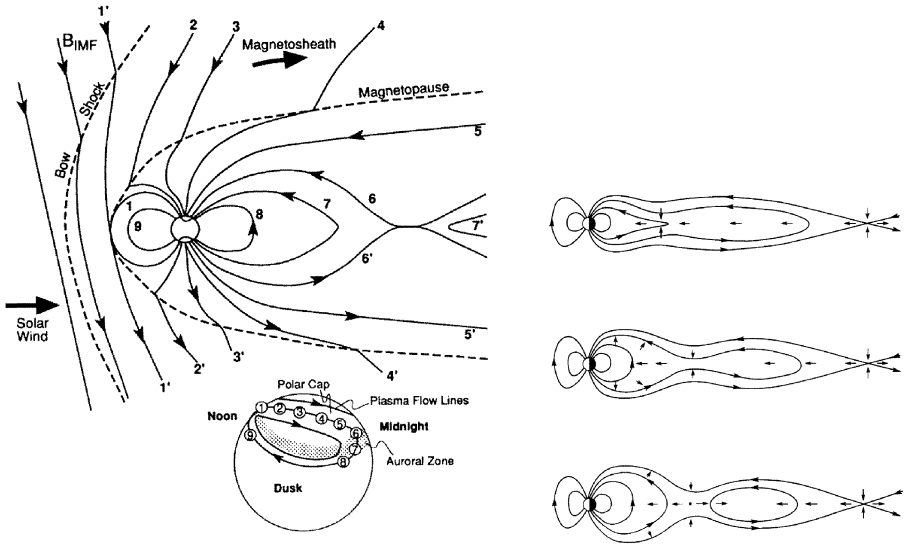

Figure 3. (a) Magnetospheric convection of magnetic field lines. Reconnection occurs between pairs of field lines 1-1' and 6-6'. The associated plasma flow in the polar ionosphere is also shown. From *Introduction to Space Physics*, Kivelson and Russell (eds.), Cambridge University Press, 1995. (b) A simplistic sketch of a magnetospheric substorm, showing the formation of a second site of reconnection in the magnetotail. The near-Earth magnetic fields becomes more dipolar, while a closed magnetic-field structure is cut off and released downtail.

lites in low-Earth orbit. Such measurements show cross-polar cap potentials between 20 kV and 150 kV, strongly depending on the solar-wind conditions.[9]

Applying the "frozen-in" field concept to the magnetospheric plasmas, we find that the convection must be associated with a circulation of magnetic flux. At two locations, the frozen-in field concept break down: at the dayside magnetopause, where flux is added to the magnetotail through reconnection, and within the tail plasma sheet, where flux is removed (field lines 1-1' and 6-6' in Fig. 3a). The newly merged field lines 1-1' in Fig. 3a are swept back over the poles and gradually sink into the tail, where the field lines 6-6' reconnect and flow back to the dayside of the magnetosphere. The associated plasma flow in the polar ionosphere is also shown in Fig. 3a.

The magnetospheric convection, as described above, can only be steady if the reconnection rates at the two sites are equal. Although they must be equal in a time-averaged sense, they are so only rarely on an instantaneous basis. During periods of southward IMF, when the energy transfer from the solar wind is enhanced, steady reconnection within the magnetotail cannot balance the inflow of magnetic flux. The magnetic flux content of the tail lobes increase until the flux is, suddenly, released by

explosive reconnection in the near-Earth plasma sheet, and the whole magnetosphere rapidly reorganizes into a more dipolar configuration. The whole sequence of events - build-up of energy in the magnetotail, explosive release of the energy together with a rapid dipolarization of the magnetic fields, followed by a slow recovery - that normally takes two or three hours, is referred to as a *magnetospheric substorm*. It is schematically described in Fig. 3b. The visual manifestations of the substorm are referred to as an *auroral substorm*, while the magnetic manifestations are referred to as a *magnetic substorm*. Much work has been devoted to understand the physical processes behind the substorm. Several phenomenological models exist, that have much in common but that also differ in important respects .[10]

During prolonged periods of strong magnetospheric convection, the ring current sometimes exhibit a rapid enhancement followed by a slow recovery back to normal levels. When this happens, we observe the classical signature of a *magnetic storm*: a depression of the field at low- and mid-latitudes up to several hundreds of nT lasting for around a day, followed by a slow recovery over several days. Magnetic storms are always accompanied by a frequent occurrence of substorms, which has led researchers to believe that substorms somehow are the cause of magnetic storms. More recently, this has been questioned and the exact role played by substorms in enhancing the ring current is still uncertain.[11]

2.3.4 Coordinate systems

The geomagnetic dipole organizes many auroral, ionospheric and magnetospheric phenomena. A simple Earth-centered dipole is therefore used to establish various coordinate systems. The *Geocentric Solar Magnetospheric* (GSM) system is a right-handed, Cartesian system centered at the Earth. The x-axis is defined by the direction from the Earth to the Sun, along the Sun-Earth line, and the z-axis is located in the plane defined by the x-axis and the dipole axis. With this definition of the z-axis, the coordinate system oscillates about the solar direction with a 24-hour period. If the z-axis instead is defined to be perpendicular to the ecliptic plane, the coordinate system is referred to as the *Geocentric Solar Ecliptic* (GSE) system.

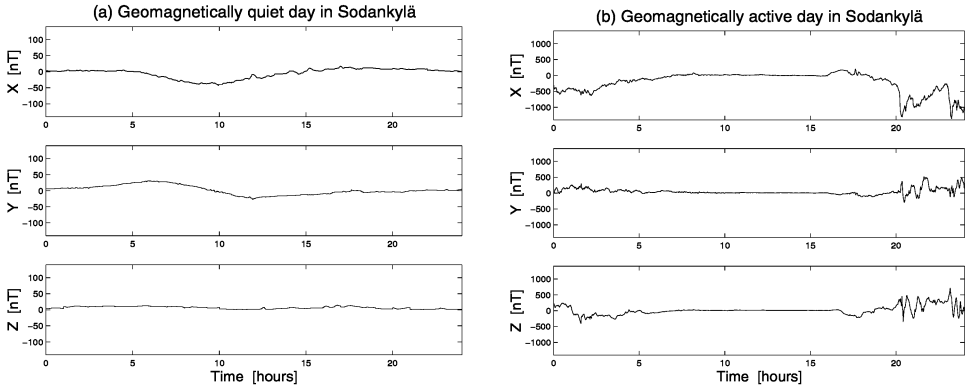

Figure 4. The northward (X), eastward (Y), and vertical (Z) geomagnetic field components in Sodankylä, Finland, during (a) a geomagnetically quiet day, and (b) a disturbed day. Note that the vertical scales are different: the quiet day record is enlarged 10 times compared to the disturbed day.

2.4 Geomagnetic activity

2.4.1 Regular geomagnetic variations

On geomagnetically quiet days, when the energy transfer from the solar wind is weak, the geomagnetic field is normally observed to undergo smooth and regular variations. These variations are caused by a global ionospheric current system fixed with respect to the Sun. Solar heating of the upper atmosphere drives winds that blow ionization across magnetic field lines. As the ions and electrons react differently to the winds of the neutral atmosphere, a dynamo is formed that drives electrical currents. The quiet-time, regular variations are mainly diurnal, with additional semi-diurnal components due to solar and lunar tides in the upper atmosphere. An example of a geomagnetically quiet day in Sodankylä, Finland, is shown in Fig. 4a.

2.4.2 Magnetic substorms

The magnetic substorm can be described as the result of two current systems with different spatial and temporal characteristics: one current system is associated with the magnetospheric convection during the substorm growth phase, and the other is associated with the transient events during the substorm expansion phase.

The growth phase of a substorm begins after a southward turning of the IMF. In the weakly collisional ionosphere, Hall currents are generated by the enhanced ionospheric convection. The currents flow roughly from midnight to noon across the polar cap, and then back to the nightside through the auroral ovals. The geomagnetic signature near the auroral oval is a positive northward disturbance before midnight and a negative

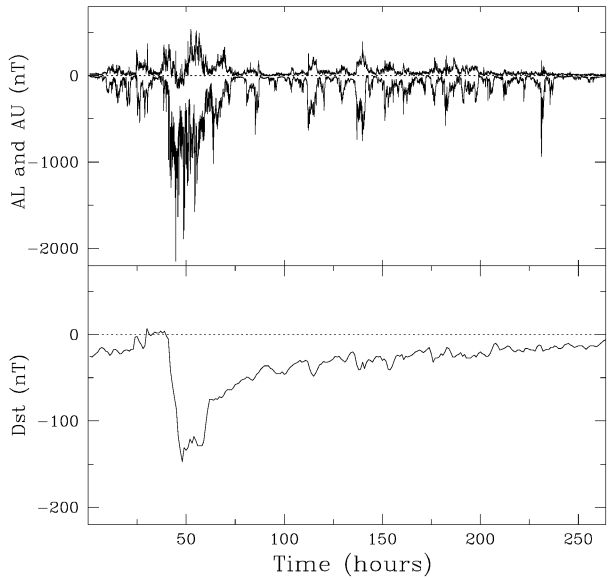

Figure 5. The development of a magnetic storm (Dst, lower panel) and the associated high-latitude geomagnetic activity (AU and AL, upper panel) during 11 days in January 1988. Note that the vertical scales of the two graphs are different: the lower graph span 300 nT, whereas the upper graph span 3000 nT.

disturbance after midnight, corresponding to an eastward electrojet across the dusk meridian and a westward electrojet across the dawn meridian.

During the growth phase, the magnetic energy in the magnetotail increases, and the cross-tail current becomes strongly enhanced and moves closer to the Earth. The substorm expansion phase starts when, due to some kind of plasma instability in the magnetotail, a part of the cross-tail current is diverted through the polar ionosphere along the midnight auroral oval in the form of a westward substorm electrojet. The geomagnetic field around local midnight show a rapid decrease of the horizontal component that can reach more than 1000 nT and that lasts some tens of minutes. In extreme cases the disturbances can reach several thousand nT.

An isolated substorm can be caused by a brief (~40 to 60 minutes) pulse of southward IMF. If the IMF remains southward for several hours, a whole sequence of substorms will be generated. These are the same conditions that enhance the ring current, and during prolonged periods of southward IMF the sequence of substorms will be accompanied by a magnetic storm as shown in Fig. 5.

2.4.3 Magnetic storms

The basic defining property of a magnetic storm is an enhanced ring current that cause a depression of the geomagnetic field at ground level. The classical signature observed at low- and mid-latitudes is a rapid decrease of the horizontal field component, followed by a slow recovery that normally takes several days. Sometimes the main phase of the storm is preceded by an increase of the geomagnetic field strength. This initial phase is no longer regarded as an essential feature of the magnetic storm.

The conditions that lead to magnetic storms are prolonged periods ($\sim$3 hours) of strong southward IMF and high solar-wind velocities. Shorter periods ($\sim$1 hour) of southward IMF do not usually lead to a buildup of the ring current, but they are normally sufficient for a substorm to develop. If the prolonged period of southward IMF is preceded by an increase of the solar-wind dynamic pressure, we also observe the typical initial phase of a magnetic storm. This phase of the storm is not caused by the ring current; it is the magnetopause currents that are strengthened and move closer to the Earth as a result of the increased dynamic pressure acting on the magnetosphere. A typical example of a major magnetic storm is shown in Fig. 5.

2.4.4 Geomagnetic activity indices

Geomagnetic observatories around the world are continuously monitoring the geomagnetic field. From this vast amount of data, various *geomagnetic indices* are derived.[12] Most indices are defined to be representative of a certain phenomenon, e.g., the strength of a specific current system. In practice, all indices are influenced by several current systems, and a separation of the different contributions can not be made from the index alone.

The Dst index, often referred to as the ring current index, has a relatively clear physical interpretation. It is a measure of the magnetic disturbance generated by a symmetric ring current. After correction for the magnetopause currents, Dst is roughly proportional to the total energy of the radiation belt particles that contribute to the ring current.[13] The Dst index is based on the horizontal field component, H, from a number of low-latitude observatories evenly distributed in longitude. The geomagnetic disturbance at each observatory is obtained by subtraction of a secularly varying base line, H_0, and the regular, daily variations, H_{Sq}, from the observed geomagnetic field, H. The geomagnetic disturbances at the observatories are averaged and divided by the average of the cosines of the geomagnetic dipole latitudes Λ

$$Dst = \frac{\langle H^i - H_0^i - H_{Sq}^i \rangle}{\langle \cos \Lambda_i \rangle} \,. \tag{1}$$

This definition of Dst allows the use of any time resolution and any number of geomagnetic observatories. However, the index presently derived at *World Data Center for Geomagnetism* in Kyoto is based on one-hour averages from four stations.

Dst is mainly influenced by the ring current and the magnetopause currents, but influences from other currents complicate the interpretation of Dst. Another uncertainty is the definition of a quiet-time reference level, since secular variations and regular daily variations can be as large as the ring-current disturbance itself.

The auroral-electrojet indices AU, AL, and AE measure the peak disturbances occurring in the auroral zone at any instant of time. They are based on a chain of geomagnetic observatories, spread around the auroral zone at magnetic latitudes between 65° and 70°, with a longitudinal spacing of 10° to 40°. At each observatory a monthly average quiet-time level, H_0, is defined. The disturbances, ΔH, are obtained as the difference between the observed field, H, and the reference level, H_0. The traces of ΔH from all the stations are then plotted with respect to a common baseline. The AU index is defined as the upper envelope of the traces, AL is defined as the lower envelope, and AE is defined as the difference between AU and AL. The AL index can thus be regarded as a measure of the maximum westward electrojet, while AU quantify the maximum eastward electrojet.

The most serious shortcoming of the AE indices is a very sparse grid of geomagnetic observatories. There are longitudinal gaps of more than 2 hours of local time, and the small latitudinal range (65° to 70° magnetic latitude) can also be a problem, particularly during low geomagnetic activity when the auroral oval contracts and the most intense electrojets tend to flow north of latitude 70°. Further, the definition of the magnetic disturbance, ΔH, does not eliminate the regular, daily variations.

3 Artificial Neural Networks

Artificial neural networks (ANNs) is the term used for a set of mathematical techniques applicable to problems which are inherently nonlinear and/or multi-dimensional, properties which often characterize "real-world" problems. Despite the somewhat strange designation, the methods are often relatively uncomplicated, with simple basic principles. During the last two decades, the ANN technique has developed until it now has become part of the standard toolbox for solving problems related to mapping between multi-dimensional spaces, time-series analysis, pattern recognition, or classification.

3.1 Modeling static systems

ANNs are composed of simple elements, or nodes, operating in parallel (Fig. 6). Each node receives several incoming signals and converts these into an outgoing activation that is passed on to other nodes in the network. Some nodes interface directly with the outside world, whereas others are internal or "hidden." An important distinction is made between feed-forward networks and recurrent networks, i.e. networks with feedback connections. The type of processing performed at the nodes and the connectivity of the network put limits to the range of possible behaviours of a network.

The basic feed-forward ANN (Fig. 6a) performs a nonlinear, static mapping from an input vector, $\{\xi_{\mathrm{k}}^{\mu}; k = 1, 2, \ldots, N_{in}\}$, with N_{in} components, to an output O^{μ}

$$O^{\mu} = \mathrm{g}_o(\sum_{\mathrm{j}=1}^{\mathrm{N_{hid}}} W_{\mathrm{j}} \mathrm{g}_h(\sum_{\mathrm{k}=1}^{\mathrm{N_{in}}} w_{\mathrm{jk}}\xi_{\mathrm{k}}^{\mu} + b_{\mathrm{j}}\xi_0) + B\xi_0) \; . \tag{2}$$

Each input-output sample $\{\xi_{\mathrm{k}}^{\mu}, O^{\mu}\}$ is labeled by superscript μ. Index j refers to a hidden node and index k refers to an input node. W_{j} and w_{jk} are weights associated with the connections. The bias input, ξ_0, is assigned a fixed value and is connected to all hidden and output nodes through the bias weights b_{j} and B. The network described by equation 2 has a single hidden layer. Most discussions in this chapter are, however, valid for any number of hidden layers.

Equation 2 describes what is sometimes referred to as a *nonlinear perceptron*. The processing performed at each node consists of a weighted summation of the incoming activations followed by a transformation, g, which is nonlinear at the hidden nodes and linear or nonlinear at the output node

$$y_{\mathrm{j}} = g(\sum_{\mathrm{k}} w_{\mathrm{jk}}\xi_{\mathrm{k}}) \; . \tag{3}$$

The function g_h, defining the transformation at the hidden nodes, should be nonlinear, continuous, and saturating (Fig. 6b). Two common choices are the hyperbolic tangent function and the logistic function. The function g_o, at the output, is often a purely linear function.

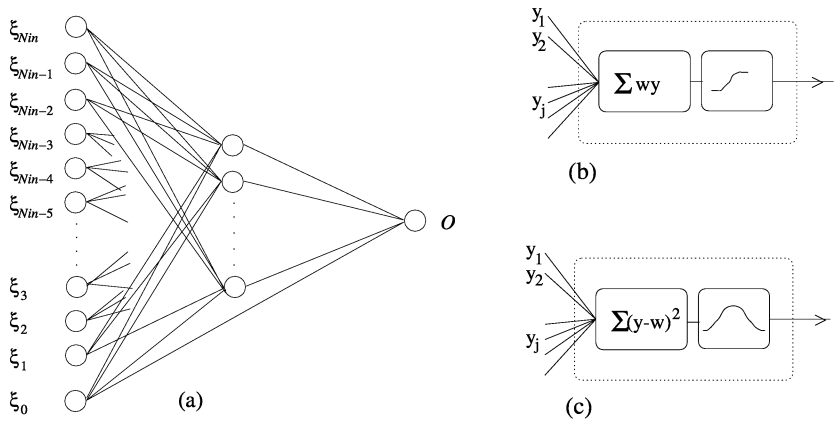

Figure 6. (a) Feed-forward network with input nodes to the left and a single output node to the right. (b) The processing performed at each hidden node for the nonlinear perceptron, and (c) for the radial-basis function network.

The *radial-basis function network* is a feed-forward network that offers an alternative to the nonlinear perceptron. It performs a different type of processing at the hidden nodes

$$y_{\mathrm{j}} = g(\sum_{\mathrm{k}} \frac{(\xi_{\mathrm{k}} - w_{\mathrm{jk}})^2}{\sigma^2}) \; . \tag{4}$$

The weights now define locations in the input space where the hidden-node activations are high. Each hidden node in a radial-basis function network produces large activations for a localized part of the input space, whereas a hidden node in a nonlinear perceptron produces large activations for all input values over a certain threshold. The function g is commonly an exponential, but any continuous function peaking at zero and with a monotonic fall-off away from zero can be used. A decisive advantage of the radial-basis function network compared to the nonlinear perceptron is that the training times can be significantly reduced.

3.2 Modeling dynamic systems

These two types of feed-forward ANNs work well for nonlinear, static mappings. However, in order to predict geomagnetic activity from solar-wind data, we need networks that can model a nonlinear, dynamic system driven by an external input. Two classes of neural networks have been used for this purpose: time-delay networks and partially recurrent networks. These two classes of networks have the advantage of being relatively simple and well-understood. They can be trained to approximate the input-

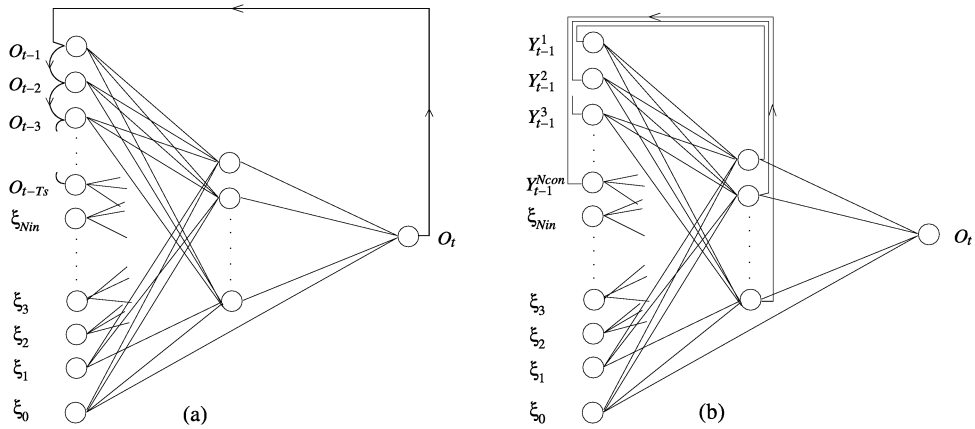

Figure 7. (a) Nonlinear ARMA network with a single feed-back connection from the output node to a time-delay line at the input. (b) Elman recurrent network with feed-back connections from the hidden nodes to a set of context nodes at the input.

output behaviour of a dynamic system, provided that sufficiently large sets of empirical input-output data are available.

The *time-delay network* (TDN) is simply a feed-forward network that is fed with a temporal sequence of time-lagged external inputs.[14] It is the organization of the input data that gives the TDN a dynamic behaviour; the mapping itself is static. The TDN is based on the assumption that the geomagnetic activity, O, can be described as a function of a vector, $\mathbf{I}$, of time-lagged solar-wind inputs

$$O_t = F(\mathbf{I}_{t-1}) \tag{5}$$

where

$$\mathbf{I}_{t-1} \equiv \{I_{t-1}, I_{t-2}, \ldots, I_{t-T_I}\} \tag{6}$$

and T_I is the temporal length of $\mathbf{I}$. If the solar-wind input at any instant of time is quantified by a single parameter, say the solar-wind electric field, then each element, I, of the input vector (6) is a single, scalar quantity and the number of input nodes, N_{in}, equals T_I. A TDN then performs the mapping

$$O_t = \mathrm{g}_o(\sum_{\mathrm{j}=1}^{\mathrm{N_{hid}}} W_{\mathrm{j}} \mathrm{g}_h(\sum_{\mathrm{k}=1}^{\mathrm{N_{in}}} w_{\mathrm{jk}} I_{t-\mathrm{k}} + b_{\mathrm{j}}\xi_0) + B\xi_0) \; . \tag{7}$$

In general, the solar-wind input at any instant of time is quantified by several parameters, and each element, I, of the input vector (6) must be a multiple quantity.

Partially recurrent networks have a limited set of fixed feed-back connections from either the output nodes,[15] or the hidden nodes.[16] The latter is referred to as an *Elman*

recurrent network (ERN). Unlike the TDN, the dynamical properties of an ERN is a result of the mapping itself being dynamic. The ERN incorporates the essential features of time-delay networks, but includes a set of special input nodes, or context nodes, that receive the hidden-node activations through a set of feed-back connections (Fig. 7b). No weights are associated with the feed-back connections which are kept fixed during the training process. This is an important property of a recurrent network. It allows the use of the same training algorithm as for the purely feed-forward networks.

The ERN is thus based on the assumption that the geomagnetic activity, O, can be described as a function of a vector, $\mathbf{I}$, of time-lagged solar-wind data together with a set of internal state variables, $\mathbf{Y}$,

$$O_t = F(\mathbf{I}_{t-1}, \mathbf{Y}_{t-1}) \tag{8}$$

where

$$\mathbf{Y}_{t-1} \equiv \{Y_{t-1}^1, Y_{t-1}^2, \ldots, Y_{t-1}^{N_{hid}}\} \tag{9}$$

are the hidden-node activations at time $t-1$, representing the internal state of the network. Under the same assumption as for the TDN in equation 7, the ERN performs the mapping

$$O_t = \mathrm{g}_o(\sum_{\mathrm{j}=1}^{\mathrm{N_{hid}}} W_{\mathrm{j}} \mathrm{g}_h(\sum_{\mathrm{k}=1}^{\mathrm{N_{in}}} w_{\mathrm{jk}} I_{t-\mathrm{k}} + \sum_{\mathrm{c}=1}^{\mathrm{N_{con}}} w_{\mathrm{jc}} Y_{t-1}^{\mathrm{c}} + b_{\mathrm{j}} \xi_0) + B\xi_0) \tag{10}$$

where index c denotes the context units. Similar to a TDN, the external input data to the ERN are organized as a temporal sequence of time-lagged data. We can, however, expect the required length, T_I, of the input sequence to be much smaller for an ERN than for a TDN. As the feedback structure of the ERN constitutes an implicit memory for past states, and thus indirectly for past inputs to the system, the importance of explicitly feeding the network with a long sequence of external inputs is reduced.

3.3 Network training

Network training is the process of finding a set of weights that gives the network a response similar to the input-output samples in a set of training data. The ability to produce a "correct" output is monitored by the cost function

$$C(\mathbf{w}) \equiv \frac{1}{2} \sum_{\mu=1}^{\mathrm{Q_{trn}}} (O^\mu - T^\mu)^2 \tag{11}$$

where $\mathbf{w}$ is the set of weights, O^μ is the actual output of the network, T^μ is the "correct" output (or target), and Q_{trn} is the number of samples in the training set. If the

nodal transfer functions are nonlinear and differentiable, this cost function will also be nonlinear and differentiable.

Network training is a nonlinear optimization problem that can be solved with many of the methods from the standard toolbox: gradient-descent methods, second-order Newton-type methods, or various search methods. In practice, most of the standard methods are not well suited to a network implementation. One of the most widely used technique is a modified gradient-descent method referred to as error back-propagation.[17] The weights are iteratively updated according to the rule

$$\Delta w_i \leftarrow -\eta \left(\frac{\partial C}{\partial w} \right)_i + \alpha \Delta w_{i-1} \tag{12}$$

$$w_{i+1} \leftarrow w_i + \Delta w_i \tag{13}$$

where w is a single weight and subscript i denotes the iteration. Normally, it is only a subset of the Q_{trn} training samples that is used in each iteration, and the actual update is in an approximate gradient direction. The size, Q_{bat}, of this subset is a parameter that, along with η and α, controls the training process. The art of choosing these parameters is discussed in many of the text books available, e.g., *Haykin*.[18]

Error back-propagation can be used both for TDNs and ERNs. A widely used alternative is the Levenberg-Marquardt method.[19] At the cost of an increased complexity and memory requirement, this method is much faster than simple gradient descent. Second-order convergence is approached without actually having to compute the Hessian matrix. The training of radial-basis function networks differs somewhat from the other network types. The orthogonal least-squares method is one of several alternatives available for this kind of network.[20]

Much of the practical use of neural networks relies on their ability to make sensible generalizations. This ability can be defined as the average performance on a randomly chosen data sample. However, the cost function $C(\mathbf{w})$ measures a network's ability to memorize the training data rather than the ability to generalize to new data. For a network that has already attained a reasonable fit to the training set, further training could actually impair the generalization ability. At some point during training, the network starts to learn details that is due to random, non-representative features of the finite set of training data. The network performance on data that are not included in the training set will then start to deteriorate; the network becomes *overfitted*.

In order to achieve a good generalization ability, rather than a perfect fit to the training data, the training procedure need to be constrained. This can be done by excluding a part of the training set from the actual training, and instead use these data to determine when to stop the iteration. In this way the problem of overfitting is avoided, or at least lessened.

3.4 Neural networks and linear/nonlinear filters

Neural networks have much in common with the linear and nonlinear filters that have been applied to magnetospheric physics, but they also differ from them in important respects. The time-delay network is essentially a nonlinear generalization of the linear moving-average (MA) filter. The auto-regressive moving-average (ARMA) filter has much in common with the Elman recurrent network, but it also has a more direct analogue amongst the neural networks, namely the network shown in Fig. 7a.

The discrete, linear MA filter is given by

$$O_t = c + \sum_{\tau=1}^{T_I} (H_\tau I_{t-\tau}) \tag{14}$$

i.e. the impulse response function, H, of the magnetospheric system is convolved with a sequence of solar-wind inputs. For a filter to be linear, H must be time-invariant and independent of the solar-wind input. The ARMA filter is obtained by adding auto-regressive terms to the MA filter

$$O_t = c + \sum_{\tau=1}^{T_I} (H_\tau I_{t-\tau}) + \sum_{\tau=1}^{T_S} (G_\tau O_{t-\tau}) \tag{15}$$

where T_S is the temporal length of the memory for prior geomagnetic-activity states.

Linear filters can be generalized to handle nonlinearities. One approach is to approximate the nonlinear response F in equation 5 locally by linear MA filters. The filter coefficients are fixed for each given input sequence, but they are allowed to vary between different regions of the input space. ARMA filters can similarly be generalized by allowing some variation of the filter coefficients between different regions of the input-output space. If H and G are fixed in each small neighbourhood of the input-output space, but change between different neighbourhoods, then this filter will be locally linear but globally nonlinear. Such *local-linear, nonlinear filters*[21] have been used in several studies to describe the magnetospheric response to the solar wind.

Equation 7 describes the processing performed by a TDN. If the activation functions g_o and g_h are linear, i.e. $g(x) = x$, then we can write the output of the TDN

$$O_t = \sum_{\mathrm{k}} (\sum_{\mathrm{j}} W_{\mathrm{j}} w_{\mathrm{jk}} I_{t-\mathrm{k}}) + \sum_{\mathrm{j}} W_{\mathrm{j}} b_{\mathrm{j}} + B \; . \tag{16}$$

As the input vector I_t represents a time series of data, we can identify the inner sum of the first term of equation 16 with the impulse response coefficients of equation 14

$$H_\tau = \sum_{\mathrm{j}} W_{\mathrm{j}} w_{\mathrm{j}\tau} \; . \tag{17}$$

A TDN with linear activation functions is apparently identical to a linear MA filter. By the same kind of reasoning it can be shown that the partially recurrent network in Fig. 7a becomes identical to a linear ARMA filter if it is assigned linear activation functions. Hence the designation nonlinear ARMA filter for this neural network.

4 Predicting Geomagnetic Activity From the Solar Wind

The geomagnetic disturbances detected at the surface of the Earth can be interpreted as an output signal from a physical system driven by an external input, namely the solar wind. It is a complex output signal, with a low-latitude component tracing the conditions in the outer radiation belt, and a high-latitude component signaling both the occurrence of magnetospheric convection and the explosive release of magnetotail energy during substorms. The ANN technique is one of several methods that can be used to map the input-output relations of this system. The accuracies obtained depend partly on limitations intrinsic to the applied methods, and partly on the predictability of the underlying magnetospheric processes. Successful mapping techniques may also be used to forecast geomagnetic disturbances and related phenomena.

4.1 Solar-wind coupling functions

The dominating mechanism of energy transfer from the solar wind to the magnetosphere is magnetic reconnection. The rate of reconnection at the dayside magnetopause is largely controlled by the solar-wind electric field $E_y = -VB_z$, where subscripts y and z refer to the GSM system. However, magnetic reconnection is only effective for a southward IMF. It has therefore been common to use the rectified electric field $VB_s = -VB_z\Gamma(\theta)$, where $\Gamma(\theta)$ is a function of the IMF direction in the $y-z$ plane. Here, Γ is 0 for northward IMF and 1 for southward IMF. Most studies also find a modulating influence by the solar-wind velocity, V, density, n, and dynamic pressure, P_{dyn}. Higher velocities or higher dynamic pressures tend to enhance the rate of energy transfer to the magnetosphere.

Many attempts have been made to find a single combination of solar-wind variables, or a *coupling function*, that describes the rate at which energy is transferred from the solar wind to the magnetosphere.[22,23] A wide variety of functions have been defined and evaluated, e.g., VB_s, V^2B_s, and $P_{dyn}^{1/2}VB_s$. Some coupling functions allow a small amount of energy transfer also for northward IMF directions through the use of a "leaky" gating function Γ, e.g., $VB_T^2\Gamma(\theta)$ and $P_{dyn}^{1/6}VB_T\Gamma(\theta)$. Here, B_T is the magnitude of the IMF component in the $y-z$ plane. Several coupling functions appear in the papers included in this thesis. It is, however, also found that the use of separate solar-wind parameters n, V, and B_z, gives more accurate predictions than any of the most common coupling functions constructed from these parameters.

4.2 Magnetic storms and the ring current

The intensity of a magnetic storm is defined by the strength of the ring current, which is commonly measured by the geomagnetic index Dst. An ideal Dst index would be

proportional to the total energy of the radiation-belt particles that contribute to the ring current.[13] Although a rough measure, Dst is an important parameter for describing the state of the inner magnetosphere.

In 1975, *Burton et al.*[24] described the evolution of the ring current by a first-order differential equation

$$\frac{dDst^*(t)}{dt} = Q(t) - \frac{Dst^*}{\tau} \tag{18}$$

where $Q(t)$ is a source term representing the injection of particles to the ring current. The purpose of this simple model was primarily to explore to what extent $Q(t)$ is controlled by the solar wind. The designation Dst^* indicates that Dst has been corrected for the influence of magnetopause currents using the simple relation

$$Dst = Dst^* + b\sqrt{P_{dyn}} + c \tag{19}$$

where b and c are constants, and P_{dyn} is the solar-wind dynamic pressure. In the study by Burton et al., the source function Q was a linear function of VB_s and the decay rate τ was constant. Later studies have expanded on this basic model, e.g., *Feldstein et al.*[25], *Pudovkin et al.*[26], and more recently *O'Brien and McPherron*[27]. Here, the decay rates have been found to vary with the phase of the storm, such that τ must be described as a function of solar-wind parameters or Dst itself. If, in fact, τ is a function of Dst, the system described by equation 18 is nonlinear. The expressions for the source function, $Q(t)$, are commonly more complex than in the original study by Burton et al.

This work on the evolution of Dst established that the injection of particles to the ring current is well approximated by a function of solar-wind parameters. In 1979, linear MA filters were first used in a study by *Iyemori et al.*[28] That study together with others, e.g., *Fay et al.*[29] and *McPherron et al.*,[30] led to similar conclusions for the relations between the solar wind and Dst. More recently, *McPherron*[31] have used linear filters to study the role of substorms in the generation of magnetic storms, and *Detman et al.*[32] have explored the use of linear filters in real-time forecasting of Dst.

The possibility of using ANNs to predict Dst was first pointed out by *Lundstedt*[33]. In 1993, *Freeman et al.*[34] presented a neural network for prediction of Dst from solar-wind parameters and Dst itself. The fact that they used the observed, hourly Dst as input to make predictions one hour ahead, indicates that the ANN relied heavily on auto-correlation of Dst. Two subsequent studies, by *Lundstedt and Wintoft*[35] and *Gleisner et al.* [paper I], showed that time-delay neural networks can be used to predict Dst from solar-wind data alone. As no pressure correction was applied to the Dst index, it was the combined disturbance from the magnetopause currents and the ring current that was predicted. Using TDNs, the main phase of magnetic storms could be accurately predicted with as little as 4 hours of input data, whereas prediction of the recovery phase required 20 hours, or more, of solar-wind data. An example is shown in Fig. 8 where six different TDNs, all of them using n, V, and B_z as input, are

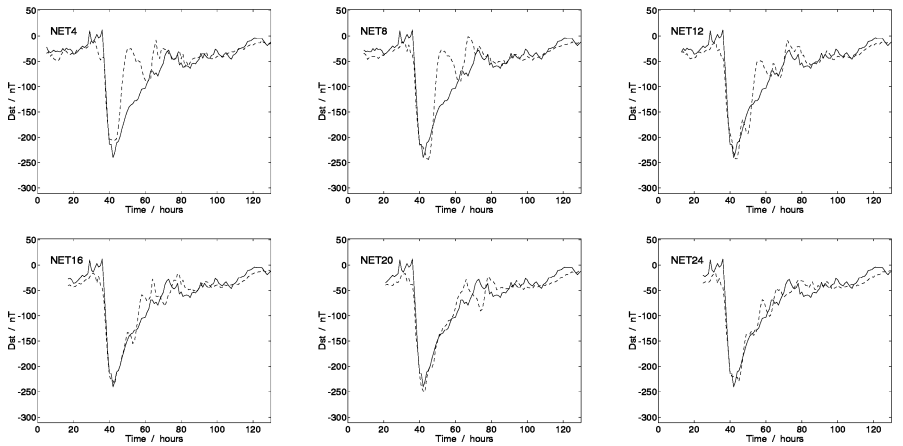

Figure 8. Comparison of observed and predicted Dst for different lengths of the solar-wind input sequence, from 4 hours (NET4) to 24 hours (NET24). The predicted recovery phase is systematically improved as the input sequence becomes longer. The predictions were made with time-delay neural networks using n, V, and B_z as input.

used to predict a major storm. Note that the predicted recovery phase is systematically improved up to the last panel of Fig. 8, where 24 hours of input data are used. The studies also showed that all three solar-wind parameters n, V and B_z are important. Removal of any of them impairs the network performance.

In 1996 and 1997, *Wu and Lundstedt*[36] published several studies on Dst predictions using Elman recurrent networks. Different network setups and input data sets were evaluated. A large number of coupling functions were tested as input to the ERNs. The one that performed best was $P_{dyn}^{1/6} V B_s$, which is very close to a function proposed by *Vasyliunas et al.*[23] on the grounds that it has dimension of power. It was, however, only marginally better than the more commonly used function $P_{dyn}^{1/2} V B_s$.

Recently, the standard model by *Burton et al.* was extended by *Klimas et al.*[37] and *Vassiliadis et al.*[38] to second order and time-varying coefficients

$$\frac{d^2 Dst(t)}{dt^2} + \nu \frac{dDst(t)}{dt} = Q(t) - \frac{Dst}{\tau} \tag{20}$$

The model was developed using local-linear, nonlinear filters,[46] related to the nonlinear filters briefly mentioned in chapter 3.4, together with a technique for deriving closed-form analytical models from the filters.[40] From such models, *Vassiliadis et al.*[39] were able to derive timescales governing ring current growth and decay.

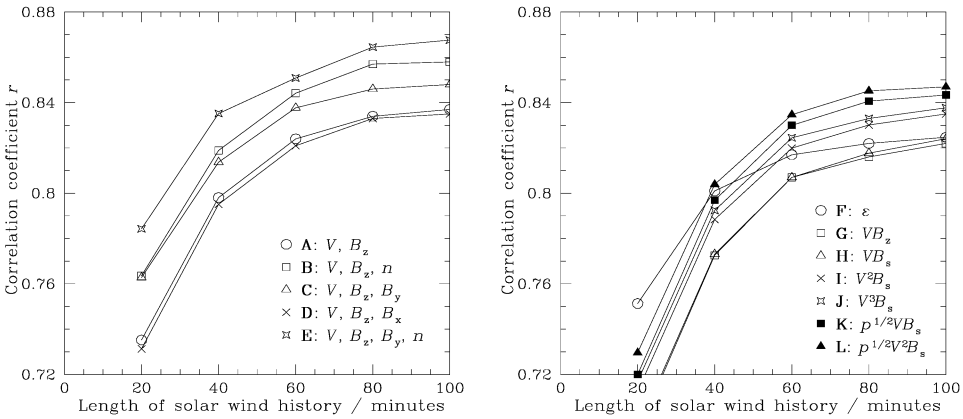

Figure 9. Predictions of the auroral electrojet index AE with time-delay networks. Network performances for different lengths of the solar-wind input sequence, (left) using individual solar-wind parameters as input, and (right) using coupling functions as input.

4.3 Magnetic substorms and the auroral electrojets

The effects of substorms can be seen in the records of the geomagnetic indices AU, AL, and AE, which are often used to quantify the strength of magnetic substorms. Ideally, these indices measure the peak geomagnetic disturbances occurring in the auroral zone at any instant of time.

In the years around 1970, a number of investigators began to publish studies on the relationships between the auroral-electrojet indices and the solar wind. These earliest studies were based on cross-correlation analysis, superposed epochs, or other statistical techniques. They all found a strong dependence on the rectified magnetic field B_s, and also confirmed that the correlation between geomagnetic activity and the solar wind is highest when the solar-wind magnetic field components are given in GSM coordinates.

In a study by *Iyemori et al.*[28] in 1979, linear MA filters were used to relate VB_s to hourly AL, AU, and AE indices. All three filters showed a peak around 1 hour lag time, but otherwise the filters had very different characteristics. This work was followed by several linear-filter studies using data of higher time resolutions, e.g., *Clauer et al.*[41], *Bargatze et al.*[42], and *McPherron et al.*[30] The study by *Bargatze et al.* clearly showed that the geomagnetic response is fundamentally nonlinear, with contributions both from magnetospheric convection and from the release of energy stored in the magnetotail. These two components of the high-latitude geomagnetic activity are commonly referred to as the *directly driven* and the *unloading* components, respectively.

In 1993, *Hernandez et al.*[43] published a study on AL predictions using two types of ANN: a nonlinear ARMA filter and a nonlinear MA filter, the latter being similar to a TDN. The ANNs were able to predict AL, but due to clipping of the high amplitude

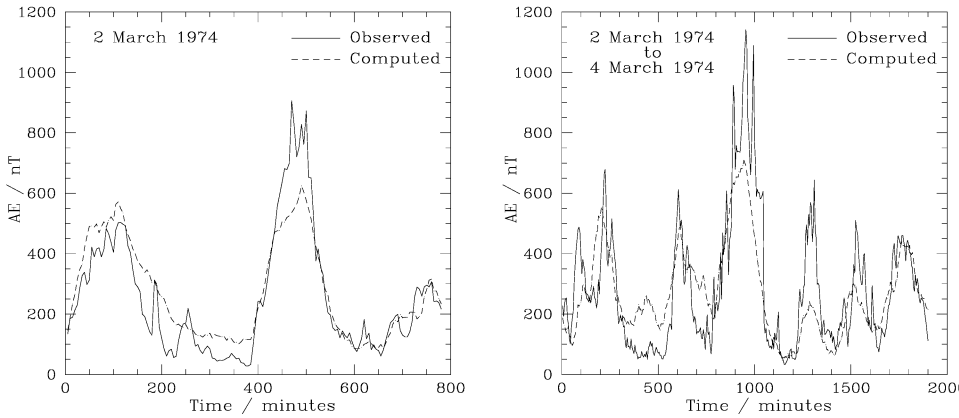

Figure 10. Observed and predicted AE during two days in March 1974. The two days are divided into two intervals. The first (left) begins in the morning of March 2 at 08.10 UT and the second (right) begins in the evening the same day at 21.35 UT. The predictions were made with a time-delay network using 100 minutes of n, V, B_y, and B_z as input.

variations they performed no better than linear filters. The question of how to deal with this clipping problem was addressed in 1999 by *Weigel et al.*[44], who suggested the use of a gating technique whereby different activity levels are handled by different networks, and then combined into a single predicted value using a gating network. In 1997, *Gleisner and Lundstedt* [paper II] used time-delay networks to predict AE from solar-wind data alone. Their study showed that the TDNs performed better when individual solar-wind variables were used as input instead of coupling functions. They also showed that the IMF component B_y has a modulating influence on the predicted AE, whereas no influence from B_x can be found. Both these results can be seen in Fig. 9. TDNs were also used to investigate whether the Dst state has any influence on the modeled solar wind-AE relation [paper III]. Recently, an AE prediction study based on Elman recurrent networks was presented by *Gleisner and Lundstedt* [paper V]. That study demonstrated that very simple ANN configurations can predict the AE index from solar-wind data with a relatively high accuracy, if the networks are provided with a simple feedback mechanism. Neural networks have further been employed by *Takalo and Timonen*[45] to study the dynamical behaviour of observed and nonlinearly predicted AE time series.

Parallel to this work on neural networks, there has been a continuous development of nonlinear filtering techniques. *Vassiliadis et al.*[46] developed a globally nonlinear, but locally linear filtering technique (see section 3.4), and used it to study AL predictions from solar-wind data. They found that the performances of the nonlinear filters were superior to the corresponding linear filters. The same method was later used by *Vas-*

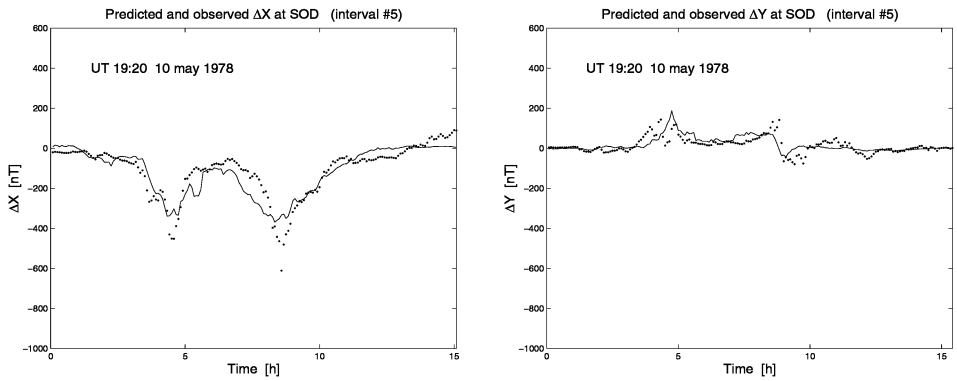

Figure 11. Observed and predicted northward (ΔX) and eastward (ΔY) geomagnetic disturbance components at Sodankylä Geomagnetic Observatory. The interval covers 15 hours, beginning on the evening on May 10, 1978.

siliadis et al.[47] to investigate the degree of nonlinearity of the geomagnetic response to the solar wind. This problem was also addressed by *Price et al.*[48] using a somewhat different technique, although based on the same general concepts.[49]

As pointed out above, the auroral-electrojet indices measure the peak disturbances occurring in the auroral zone. They do not provide any information on the location of the peak disturbances, or the disturbance levels are at other locations. For some purposes this is serious shortcoming. There are, however, a number of models, e.g., *Heppner and Maynard*[50], and *Weimer*[51], that describe the ionospheric convection patterns and the associated geomagnetic activity within the polar caps as a function of solar-wind parameters. These models are only valid for intervals dominated by quasi-steady convection effects, and the dependences on the solar wind are usually linear. Another type of model was recently reported by *Valdivia et al.*[52] They modified the nonlinear filter technique previously developed by *Vassiliadis et al.*[46] so that it can be applied to spatial distributions of geomagnetic disturbances. Using geomagnetic data from a chain of magnetic observatories, they obtained a model of the spatial pattern of disturbances at auroral-zone latitudes. Paper IV in this thesis demonstrates how a combination of time-delay networks and radial-basis function networks can predict the locally observed geomagnetic variations at one particular site, located near the peak of the auroral zone. These latter studies show an interesting development toward prediction models that take the spatial dimension of the geomagnetic disturbances into account.

5 Summary of Research Articles

The thesis is based on five journal articles that are published in *Annales Geophysicae* and in the space physics section of *Journal of Geophysical Research*. The common theme of the articles is prediction of geomagnetic activity from solar-wind data using neural networks.

In a study by *Lundstedt and Wintoft*[35] it is demonstrated that time-delay networks are able to predict aspects of magnetic storms from an 8-hour sequence of solar-wind data. The main phases of the storms are well predicted, whereas the recovery phases are not accurately accounted for. **Paper I** describes an attempt to overcome the limitations found in that prior study. Using hourly Dst and solar-wind data, six time-delay networks were trained to predict Dst one hour ahead. All networks were fed with n, V, and B_z through a time-delay line with a temporal length from 4 hours to 24 hours. The results show that the trained networks perform better with a longer time-delay line, and that the improvements are significant up to an input-sequence length between 16 and 20 hours. A closer inspection of individual storms shows that the improvements are largely due to better predictions of the recovery phase. In many cases, the recovery phase is most accurate using a 24-hour delay line, while a 4-hour delay line is enough to predict the main phase. For the best performing network, the correlation between prediction and observation is 0.92 as calculated over the whole test set, corresponding to 84% of the observed Dst index variance.

In **Paper II**, time-delay networks are used to predict the auroral-electrojet index AE from solar-wind data at a 5-minute time resolution. The data were selected from essentially the same time period (Nov. 1973 to Dec. 1974) as in the seminal study by *Bargatze et al.*[42] In the first part of the study, various combinations of separate solar-wind parameters are used as input to the networks (Fig. 9a). For each combination, the temporal length of the input-data sequence vary from 20 to 100 minutes. It is found that the solar-wind parameters n, V, B_y, and B_z contribute to improved predictions. When any of these four solar-wind parameters is left out, the network performance decreases. It is also found that B_x does not have any influence on the network performance, and that the temporal size of the time-delay line should be at least 100 minutes for maximum prediction accuracy. A properly trained network with 100 minutes of n, V, B_y, and B_z as input, accounts for 76% of the AE variance.

In the second part of the study, the input to the networks consists of coupling functions. It is a selection of the most widely used functions that are evaluated (Fig. 9b). It is found that any coupling function constructed from V and B_s is improved by a simple scaling with $P_{dyn}^{1/2}$. Around 71% of the observed AE variance can be accounted for by a time-delay network taking 100 minutes of $P_{dyn}^{1/2}V^2B_s$ as input. This is less accurate than using the solar-wind parameters n, V, and B_z as separate inputs. Information on the solar wind that is relevant to the AE variations is obviously lost when the raw solar-wind parameters are combined into a coupling function.

It is also shown in paper II that much of the high-frequency variations are filtered out: the predicted AE is essentially a smoothed version of the observed AE. There are features in the geomagnetic data, such as the large and sudden excursions due to intensifications of the westward electrojet, that are not accurately reproduced by the networks. The amplitudes of the substorm disturbances also tend to be underestimated. Nevertheless, the gross features of individual substorms are actually reproduced, as shown in Fig. 10.

In **Paper III**, the previous work is extended to examine whether Dst has an influence on the AE predictions that could indicate a ring-current modulation of the modeled solar wind-AE relation. Predictions of AE based on both solar-wind data and Dst are compared with predictions from solar-wind data alone. Two conclusions are reached: (1) with an optimal set of solar-wind data available, the AE predictions are not markedly improved by the Dst input, but (2) the AE predictions are improved by Dst if less than, or other than, the optimum solar-wind data are available to the net. It appears that the solar wind-AE relation described by an optimized neural net is not significantly modified by the magnetosphere's Dst state. When the solar wind alone is used to predict AE, the correlation between predicted and observed AE is 0.86, and the prediction residual is virtually uncorrelated to Dst.

In **Paper IV** we do not use any magnetic index. Instead, it is the directly observed magnetic field variations at a particular site that are modeled. Secular variations are first removed from the observed geomagnetic records by a piecewise linear fit to quiet-time annual means. Then, the daily quiet-time variations are modeled by radial-basis function networks taking local time, day number, and solar 10.7 cm radio flux as input. The annual and solar-cycle modulations of the regular variations are thus accounted for. The remaining horizontal disturbance components ΔX and ΔY are modeled with gated TDNs taking local time and a sequence of solar-wind data as input.

This modeling procedure is used in paper IV to predict the geomagnetic variations at Sodankylä Geomagnetic Observatory, located near the peak of the auroral zone. It is shown that 73% of the ΔX variance, but only 34% of the ΔY variance, is predicted by the neural networks. The reason for this large difference is not clear. However, one potentially important factor is the different spatial scales of the source currents. Large-scale ionospheric electrojet currents have a predominant east-west flow direction. These are the currents that generate ΔX. The spatial scales of currents flowing in the north-south direction, which are the currents that generate ΔY, tend to be smaller.

The above figures refer to prediction of the irregular variations, or disturbances. The corresponding figures for prediction of *all* transient variations, including both the regular and the irregular components, are 74% and 51% for the northward and eastward components, respectively. Compared to predictions of the irregular variations alone, the prediction accuracy is significantly improved for the eastward component, whereas the accuracy is nearly the same for the northward component. This result reflects the fact that the regular variations contribute a larger part of the total variability

for the eastward component than for the northward component. The magnitudes of the northward and eastward regular variations are nearly equal, whereas the irregular variations ΔX are considerably larger than ΔY. It is also emphasized in paper IV that the prediction accuracies are subject to a strong local-time modulation. An example of observed and predicted horizontal field disturbance is shown in Fig. 11.

One motivation for **Paper V** was the apparent success of Elman recurrent networks to predict the storm-time Dst variations.[36] The evolution of Dst is partly governed by the solar-wind input and partly by internal magnetospheric processes, and ERNs are apparently able to describe at least a part of the dynamics of the solar wind-Dst relation. To what extent would ERNs be able to approximate the solar wind-AE relation, with its completely different dynamic characteristics?

Paper V describes predictions of AE using both ERNs and TDNs, with data at a 2.5-minute time resolution. The data are essentially those used by *Bargatze et al.*[42] It is shown that an ERN can predict 71% of the observed AE variance using only a single sample of solar wind n, V, and B_z as input, i.e. with no time-lagged external input data at all. A neural network with identical solar-wind input, but without a feedback mechanism, only predicts around 45% of the observed AE variance.

The ERNs are compared to TDNs taking a sequence of time-lagged solar-wind data as input. To reach comparable prediction accuracies as an ERN, a TDN needs up to 100 minutes of input data. The fact that it takes nearly 100 minutes of solar-wind data for a TDN to accomplish what an ERN can do with a single 2.5-minute sample of input data, shows that something of the solar wind-AE dynamics have been encoded into the feed-back structure of the ERN. The structure of the ERN can be very simple and still produce relatively accurate predictions: from 1 to 4 input nodes (depending on what solar-wind parameters are used), 4 hidden nodes, 4 context nodes, and a single output node. In fact, with only 2 hidden nodes, an thus 2 context nodes, the ERN produce predictions that are surprisingly accurate.

Acknowledgements

There are many people to whom I owe my gratitude. First of all Henrik Lundstedt, supervisor of this PhD project, whose enthusiasm for the Sun is an excellent indicator of the solar activity level, reaching a peak with every new coronal mass ejection or major flare. Peter Wintoft, Vicke Dovheden, and Ingmar Kronfeldt, working on related problems, are gratefully acknowledged for ideas, suggestions, and discussions. I also want to thank the whole staff at Lund Observatory for help and support in different respects.

Last, I want to thank my family. Katta could not have encouraged or helped me more, and Theo has given me many important new insights, such as the rich values of a spade and a bucket.

References

1. Introduction

[1] **Sabine, E.**, On periodical laws discoverable in the mean effects of the larger magnetic disturbances, *Phil. Trans.*, **142**, 103, 1852.

[2] **Broun, J.A.**, On certain results of magnetical observations, *Philos. Magazine*, **16**, 81, 1858.

[3] **Hufbauer, K.**, *Exploring the Sun*, pp. 213-39, The Johns Hopkins University Press, Baltimore, Maryland, 1991.

2. The Solar-Terrestrial Space Environment

[4] **Parker, E.N.**, Dynamics of the interplanetary gas and magnetic fields, *Astrophys. J.*, **128**, 664, 1958.

[5] **Gosling, J.T.**, Corotating and transient solar wind flows in three dimensions, *Ann. Rev. Astron. Astrophys.*, **34**, 35, 1996.

[6] **Crooker, N., J.A. Joselyn, and J. Feynman, (eds.)**, *Coronal Mass Ejections*, AGU Geophysical Monograph 99, AGU, Washington, DC, 1997.

[7] **Lindsay, G.M., C.T. Russell, and J.G. Luhmann**, Coronal mass ejections and stream interaction region characteristics and their potential geomagnetic effectiveness, *J. Geophys. Res.*, **100**, 16999, 1995.

[8] **Daglis, I., R.M. Thorne, W. Baumjohann, and S. Orsini**, The terrestrial ring current: origin, formation, and decay, *Rev. Geophys.*, **37**, 407, 1999.

[9] **Heppner, J.P., and N.C. Maynard**, Empirical high-latitude electric field models, *J. Geophys. Res.*, **92**, 4467, 1987.

[10] **Rostoker, G.**, Phenomenology and physics of magnetospheric substorms, *J. Geophys. Res.*, **101**, 12995, 1996.

[11] **Gonzalez, W.D., J.A. Joselyn, Y. Kamide, H.W. Kroehl, G. Rostoker, B.T. Tsurutani, and V.M. Vasyliunas**, What is a geomagnetic storm?, *J. Geophys. Res.*, **99**, 5771, 1994.

[12] **Mayaud, P.N.**, *Derivation, Meaning, and Use of Geomagnetic Indices*, AGU Geophysical Monograph 22, AGU, Washington, DC, 1980.

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

3. Artificial Neural Networks

[14] **Waibel, A.T., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang**, Phoneme recognition using time-delay neural networks, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **2**, 328, 1989.

[15] **Jordan, M.I.**, Serial order: a parallel, distributed processing approach, in *Advances in Connectionist Theory: Speech*, J.L. Elman and D.E. Rumelhart (Eds.), Erlbaum, Hillsdale, 1989.

[16] **Elman, J.L.**, Finding structure in time, *Cognitive Science*, **14**, 179, 1990.

[17] **Rumelhart, D.E., G.E. Hinton, and R.J. Williams**, Learning representations by back-propagating errors, *Nature*, **323**, 533, 1986.
[18] **Haykin, S.**, *Neural networks - a comprehensive foundation*, 2nd ed., Prentice Hall, Upper Saddle River, New Jersey, 1999.
[19] **Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery**, *Numerical Recipes in Fortran*, 2nd ed., Cambridge University Press, Cambridge, NJ, 1992.
[20] **Chen, S., F.N. Cowan, and P.M. Grant**, Orthogonal least squares algorithm for radial basis function networks, *IEEE Trans. on Neural Networks*, **2**, 302, 1991.
[21] **Vassiliadis, D.**, The input-state space approach to the prediction of auroral geomagnetic activity from solar wind variables, in proceedings of the *International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, Lund, Sweden, 22-24 Sept, 1993.

4. Predicting Geomagnetic Activity From the Solar Wind

[22] **Murayama, T.**, Coupling function between solar wind parameters and geomagnetic indices, *Rev. Geophys. Space Phys.*, **20**, 623, 1982.
[23] **Vasyliunas, V.M., J.R. Kan, G.L. Siscoe, and S.-I. Akasofu**, Scaling relations governing magnetospheric energy transfer, *Planet. Space Sci.*, **30**, 359, 1982.
[24] **Burton, R.K., R.L. McPherron, and C.T. Russell**, An empirical relationship between interplanetary conditions and Dst, *J. Geophys. Res.*, **80**, 4204, 1975.
[25] **Feldstein, Y.I., V.Y. Pisarsky, N.M. Rudneva, and A. Grafe**, Ring current simulation in connection with interplanetary space conditions, *Planet. Space Sci.*, **32**, 975, 1984.
[26] **Pudovkin, M.I., S.A. Zaitseva, and L.Z. Sizova**, Growth rate and decay of magnetospheric ring current, *Planet. Space Sci.*, **33**, 1097, 1985.
[27] **O'Brien, T.P., and R.L. McPherron**, An empirical phase-space analysis of ring-current analysis: solar wind control of injection and decay, *J. Geophys. Res.*, **105**, 7707, 2000.
[28] **Iyemori, T., H. Maeda, and T. Kamei**, Impulse response of geomagnetic indices to interplanetary magnetic fields, *J. Geomag. Geoelectr.*, **31**, 1, 1979.
[29] **Fay, R.A., C.R. Garrity, R.L. McPherron, and L.F. Bargatze**, Prediction filters for the *Dst* index and the polar cap potential, in *Solar wind-magnetosphere coupling*, edited by Y. Kamide and J.A. Slavin, p. 111, Terra Scientific Publishing, Japan, 1986.
[30] **McPherron, R.L., D.N. Baker, L.F. Bargatze, C.R. Clauer, and R.E. Holzer**, IMF control of geomagnetic activity, *Adv. Space Res.*, **8**(9), 71, 1988.
[31] **McPherron, R.L.**, The role of substorms in the generation of magnetic storms, in *Magnetic Storms*, AGU Geophysical Monograph 98, pp. 131-147, AGU, Washington, DC, 1997.
[32] **Detman, T.R., W.D. Gonzalez, and A.L. Cluá-Gonzalez**, Prediction of the geomagnetic (Dst) index by adaptive filtering of solar wind data, in proceedings of the *International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, Lund, Sweden, 22-24 Sept, 1993.
[33] **Lundstedt, H.**, A trained neural network, geomagnetic activity, and solar wind variation, in proceedings of the *Solar-Terrestrial Prediction IV Workshop in Ottawa, Canada, May 18-22, 1992*, edited by J. Hruska, M.A. Shea, D.F. Smart, and G. Hackman, NOAA, USA, 1993.
[34] **Freeman, J., A. Nagai, P. Reiff, W. Denig, S. Gussenhoven-Shea, M. Heinemann,**

F. Rich, and M. Hairston, The use of neural networks to predict magnetospheric parameters for input to a magnetospheric forecast model, in proceedings of the *International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, Lund, Sweden, 22-24 Sept, 1993.

[35] **Lundstedt, H., and P. Wintoft**, Prediction of geomagnetic storms from solar wind data with the use of a neural network, *Ann. Geophys.*, **12**, 19, 1994.

[36] **Wu, J.-G.**, *Dynamic neural network studies of solar wind-magnetosphere coupling*, PhD Thesis, Lund Observatory, Lund, Sweden, 1997.

[37] **Klimas, A.J., D. Vassiliadis, and D.N. Baker**, Dst index prediction using data-derived analogues of magnetospheric dynamics, *J. Geophys. Res.*, **103**, 20435, 1998.

[38] **Vassiliadis, D., A.J. Klimas, and D.N. Baker**, Models of Dst geomagnetic activity and of its coupling to solar wind parameters, *Phys. Chem. Earth*, **24**, 107, 1999.

[39] **Vassiliadis, D., A.J. Klimas, J.A. Valdivia, and D.N. Baker**, The Dst geomagnetic response as a function of storm phase and amplitude and the solar wind electric field, *J. Geophys. Res.*, **104**, 24957, 1999.

[40] **Klimas, A.J., D. Vassiliadis, and D.N. Baker**, Data-derived analogues of the magnetospheric dynamics, *J. Geophys. Res.*, **102**, 26993, 1997.

[41] **Clauer, C.R., R.L. McPherron, C. Searls, and M.G. Kivelson**, Solar-wind control of auroral zone geomagnetic activity, *Geophys. Res. Lett.*, **8**, 915, 1981.

[42] **Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones Jr.**, Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res.*, **90**, 6387, 1985.

[43] **Hernandez, J.V., T. Tajima, and W. Horton**, Neural net forecasting for geomagnetic activity, *Geophys. Res. Lett.*, **20**, 2707, 1993.

[44] **Weigel, R.S., W. Horton, and T. Tajima**, Forecasting auroral electrojet activity from solar wind input with neural networks, *Geophys. Res. Lett.*, **26**, 1353, 1999.

[45] **Takalo, J., and J. Timonen**, Neural network prediction of AE data, *Geophys. Res. Lett.*, **24**, 2403, 1997.

[46] **Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts**, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res.*, **100**, 3495, 1995.

[47] **Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts**, The nonlinearity of models of the vB_{South}-AL coupling, *J. Geophys. Res.*, **101**, 19779, 1996.

[48] **Price, C.P., D. Prichard, and J.E. Bischoff**, Nonlinear input/output analysis of the auroral electrojet index, *J. Geophys. Res.*, **99**, 13227, 1994.

[49] **Casdagli, M.**, A dynamical systems approach to modeling input-output systems, in *Nonlinear Modeling and Forecasting*, edited by M. Casdagli and S. Eubank, pp. 265, Addison-Wesley, 1992.

[50] **Heppner, J.P., and N.C. Maynard**, Empirical high-latitude electric field models, *J. Geophys. Res.*, **92**, 4467, 1987.

[51] **Weimer, D.R.**, A flexible, IMF dependent model of high-latitude electric potentials having "space weather" applications, *Geophys. Res. Lett.*, **23**, 2549, 1996.

[52] **Valdivia, J.A., D. Vassiliadis, A.J. Klimas, and A.S. Sharma**, Modelling the spatial structure of the high latitude perturbations and the related current systems, *Physics of Plasmas*, **6**, 4185, 1999.

I

Ann. Geophysicae 14, 679–686 (1996)

Annales Geophysicae

Predicting geomagnetic storms from solar-wind data using time-delay neural networks

H. Gleisner, H. Lundstedt, P. Wintoft

Lund Observatory, Box 43, S-22100 Lund, Sweden

Received: 9 October 1995/Revised: 9 February 1996/Accepted: 27 February 1996

Abstract. We have used time-delay feed-forward neural networks to compute the geomagnetic-activity index D_{st} one hour ahead from a temporal sequence of solar-wind data. The input data include solar-wind density n, velocity V and the southward component B_z of the interplanetary magnetic field. D_{st} is not included in the input data. The networks implement an explicit functional relationship between the solar wind and the geomagnetic disturbance, including both direct and time-delayed non-linear relations. In this study we especially consider the influence of varying the temporal size of the input-data sequence. The networks are trained on data covering 6600 h, and tested on data covering 2100 h. It is found that the initial and main phases of geomagnetic storms are well predicted, almost independent of the length of the input-data sequence. However, to predict the recovery phase, we have to use up to 20 h of solar-wind input data. The recovery phase is mainly governed by the ring-current loss processes, and is very much dependent on the ring-current history, and thus also the solar-wind history. With due consideration of the time history when optimizing the networks, we can reproduce 84% of the D_{st} variance.

1 Introduction

The earth's magnetosphere responds to the ever-changing solar-wind conditions in a variety of ways. Some of the resulting magnetospheric disturbances can be detected at the earth's surface as geomagnetic disturbances due to changes in the large-scale electrical current systems flowing in the magnetosphere and ionosphere. A widely used index for quantifying the disturbance level is D_{st}, which originally was introduced by Sugiura (1964) as a measure of the ring-currrent magnetic field. This index is defined as the reduction of the horizontal magnetic component at the geomagnetic dipole equator, and has often been used in studies of the solar wind-magnetosphere coupling.

A typical low-latitude disturbance, the *geomagnetic storm*, can be divided into three phases with different causes and characteristics. The *initial phase* is caused by an increased solar-wind dynamic pressure acting on the magnetopause as a result of the arrival of a solar-wind disturbance. The increased pressure compresses the day-side magnetosphere, forcing the magnetopause current closer to the earth while at the same time increasing it. It has been shown that the resulting D_{st} enhancement is proportional to the square root of the solar-wind dynamic pressure (Siscoe *et al.*, 1968; Ogilvie *et al.*, 1968).

The *main phase* is due to an increase in energetic ions and electrons in the inner magnetosphere, where they become trapped on closed magnetic field lines and drift around the earth, thus creating the ring current. This current creates a magnetic field opposing the geomagnetic field at the ground, and can be measured as a large decrease in the horizontal geomagnetic component (Akasofu and Chapman, 1961).

The ring current is subject to several loss processes. It will gradually lose particles to the upper atmosphere and the surrounding plasma populations. This can be seen in a ground-level magnetogram as the *recovery phase*, a slow recovery of the geomagnetic field back to its undisturbed strength (Williams, 1983).

This picture of the classical, low-latitude geomagnetic storm is very schematic. In reality, the storms show a considerable variety as a result of the diversity of interplanetary disturbances (Akasofu, 1981).

The development of D_{st} can (after correction for a varying dynamic pressure) be described in terms of source and loss mechanisms, following, for example, Akasofu (1981):

$$\frac{d(D_{st})}{dt} = Q - \frac{1}{\tau} D_{st}. \tag{1}$$

The build-up of the ring current depends on the efficiency of the coupling between the solar wind and the magnetosphere. This efficiency in turn depends on the magnitude and direction of the interplanetary magnetic field (IMF),

Correspondence to: H. Gleisner

the most efficient being an IMF with a large southward component (Rostoker and Fälthammar, 1967). The source term Q in Eq. 1 thus becomes a function of the solar-wind conditions, and controls the development of the main phase.

The ring-current particles are subject to several loss processes and the total decay rate τ varies considerably during a geomagnetic storm, mainly because different ion species have different lifetimes in the ring current (Williams, 1983). This decay is governed by the loss term in Eq. 1, which controls the development of the recovery phase.

Since in situ measurements of solar-wind properties became generally available, various methods have been used for studying the relationships between the solar wind and the magnetospheric response. Burton *et al.* (1975) developed a simple empirical model for predicting D_{st} solely from the solar-wind dynamic pressure and the dawn-to-dusk component of the interplanetary electric field. A similar study was presented by Feldstein *et al.*, in 1984. Another empirical model, though somewhat more complicated and including non-linear responses, was used by Goertz *et al.* (1993) to study the response of the auroral electrojet to the solar wind.

Amongst the general-purpose methods, the linear filters have caught most attention during the last 15 years. They were first applied to magnetospheric physics by Iyemori *et al.* (1979) and Iyemori and Maeda (1980), and have also been used by McPherron *et al.* (1986) to study the magnetospheric response, in terms of D_{st}, to the solar-wind input. A detailed description of the technique is given by Clauer (1986). Other studies of the solar wind-magnetosphere coupling have later been made using linear filtering methods (Bargatze *et al.*, 1985; Fay *et al.*, 1986; Detman *et al.*, 1993). By using non-linear filtering, the geomagnetic-activity predictions have been further improved (Vassiliadis *et al.*, 1995).

Prediction of the D_{st} index by means of artificial neural networks was introduced by Lundstedt (1992a, b). Neural networks have since been used by Freeman and Nagai (1992), and Lundstedt and Wintoft (1994). This technique has previously proved to be an efficient and rather simple way of finding complex non-linear relationships between two sets of interrelated data, something which is described in detail by Hertz *et al.* (1991). In this study we continue the work of Lundstedt and Wintoft aiming at predicting all phases of geomagnetic storms, including the recovery phase, using feed-forward neural networks.

2 Geomagnetic and solar-wind data

2.1 Data

Solar-wind plasma and IMF data have been available from measurements aboard many spacecraft since the beginning of the 1960s. Some of these data have been compiled by J. H. King (Couzen and King, 1986), and are distributed by the National Space Science Data Center. Solar-wind data, as measured from spacecraft outside the earth's bow shock either in earth orbit or in halo orbit around the sun-earth libration point L_1, have been selected and normalized. They are given as hourly averages, and the magnetic-field vectorial data are given in geocentric-solar-magnetospheric (GSM) coordinates.

The geomagnetic index D_{st} is also available in the King compilation.

2.2 Data preparation

We selected data from the 21-year period 1963–1983. These data consist of 75 storm-time periods and 9 relatively quiet periods ($-10 < D_{st} < 10$ nT), in total 84 periods covering 8800 h. The periods varied in length from 44 to 144 h, and no data gaps larger than 4 h occurred within each period. Missing data were replaced by linearly interpolated values. The 84 periods were divided into two groups: training data (62 periods covering 6600 h) and test data (22 periods covering 2100 h).

As input to the networks we used n, V and B_z a number of hours back in time, while the network output data were D_{st} one hour forward in time. To get a reasonable working range for the nodal transfer functions, all input data were scaled to the interval $[-1.0, +1.0]$ and the output data were scaled to $[-0.8, +0.8]$.

3 The time-delay feed-forward neural network

3.1 General

We have modelled the magnetospheric response to the temporally varying solar wind by artificial neural networks. These implement a functional relationship from a time series of solar-wind data

$$\xi(t), \xi(t-1), \xi(t-2), \ldots, \xi(t-(\tau_w-1))$$

to a magnetospheric response

$$O(t+1),$$

where the D_{st} index is used as a measure of the magnetospheric response. D_{st} is computed one hour ahead, which is approximately the same as the L_1-magnetopause travel time. The input data are independent of the output data, i.e. the functional relationship does not include any autocorrelation of D_{st}. This is of practical importance, since D_{st} is not available in real time.

A feed-forward network (Hertz *et al.*, 1991) is arranged in layers of nodes (Fig. 1). The input to the nodes in one layer is the sum of the weighted outputs from the nodes in the previous layer. The output from a node is given by the input to the node and the nodal transfer function, usually a sigmoidal function. Usually all nodes in one layer are connected to all nodes in the next layer, i.e. the networks are fully connected. There are no connections between nodes in the same layer. An additional node, the bias node, is set to 1 and connected to all hidden and output nodes in the network. The purpose of this is to adjust the nodal transfer functions. The key to network performance is the weights determining the strength of the connection between nodes. Since this network type belongs to the

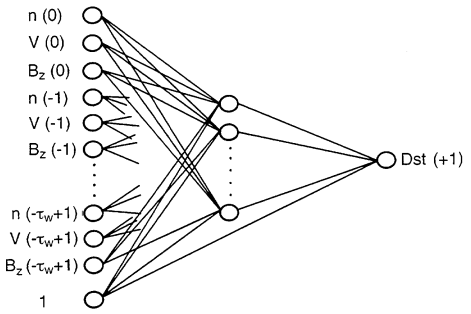

Fig. 1. The network with input nodes to the left and the single output node to the right. In the feed-forward phase the input is propagated through the net to the output. The error between actual output and desired output is propagated backwards through the net and the weights are updated accordingly. The bias node is always set to 1, as indicated. Input data sequence length τ_w varies from 4 to 24 h

class of supervised networks, it is trained by adjusting the weights until the average error on a set of known training examples is minimized. The most common training algorithm is a modified form of gradient descent called *error back-propagation* (Rumelhart *et al.*, 1986).

The neural networks used in this study were feed-forward networks with one hidden layer and one output layer. The input data to the networks are organized as a temporal sequence, where data sampled during a time window of length τ_w is shown to the network simultaneously. To get a time sequence of output data, this window is moved stepwise in time. The feed-forward neural network, together with this type of organization of the input data, is often referred to as a *time-delay neural network*.

For an input-output pair, or example, μ, the network output is given by

$$O_i^\mu = g_O\left(\sum_j W_{ij} g_H\left(\sum_k w_{jk}\xi_k^\mu\right)\right), \tag{2}$$

where $(\xi_k^\mu, k = 1 \ldots m)$ is the input-data vector. Here index i refers to a node in the output layer, index j to a hidden-layer node and index k to an input-layer node. Superscript μ denotes the examples. W_{ij} is thus a weight connecting two nodes between the hidden and output layers, while w_{jk} connects nodes between the input and hidden layers; g_H is the transfer function for nodes in the hidden layer and g_O is the transfer function for the output-layer nodes. These are hyperbolic tangent functions for hidden-layer nodes, and linear functions for output-layer nodes. The network output for an input vector $(\xi_k^\mu, k = 1 \ldots m)$ is then given by

$$O^\mu = \sum_j \left(W_j \cdot \tanh\left(\sum_k w_{jk}\xi_k^\mu\right)\right), \tag{3}$$

where index i has been omitted since the output vector consists of a single value, the predicted D_{st} index.

The network error is defined as the sum of the individual errors over a number of examples, an epoch. The network error is then given by

$$E = \frac{1}{2}\sum_\mu (O^\mu - D_{st}^\mu)^2. \tag{4}$$

Table 1. Architectures and data sets of the six networks. See text for the explanation of N_I, N_H, N_W, Q_{TRN} and Q_{TST}

Network	N_I	N_H	N_W	Q_{TRN}	Q_{TST}
NET4	12	45	631	6359	1997
NET8	24	23	599	6111	1909
NET12	36	15	571	5863	1821
NET16	48	11	551	5615	1733
NET20	60	8	497	5367	1645
NET24	72	7	519	5119	1557

3.2 Network setup – choice of network size

After choosing the type of network, in this case a feed-forward network with one hidden layer, one must decide the number of nodes in each layer. In the output layer there is only a single node, the predicted D_{st} index. The number of nodes in the input layer is determined by the number of input data. Since we used three hourly solar-wind parameters during τ_w hours as input data, each network had $3\tau_w$ input nodes. Six networks were created with varying size of input window: 4, 8, 12, 16, 20 and 24 h respectively. Their number of input nodes are shown in Table 1.

The size (i.e. number of weights) of each network is only determined by the number of hidden nodes, as the number of input and output nodes are given. The number of weights in the network has to be large enough to represent the full complexity of the problem, and it has to be small enough not to overfit and lose generalization ability. The minimum number of weights is thus determined by the complexity of the relationship we are trying to model, while the maximum number of weights is determined by the number of training data available. We chose to set the number of hidden nodes so that the number of weights in each network is approximately one tenth of the number of training data available. This rule was earlier used by Lundstedt and Wintoft (1994). The number of hidden nodes is however not a critical parameter, which is further discussed in Sect. 4.3.

In the discussions below, the six networks are referred to as NET4, NET8, NET12, NET16, NET20 and NET24. The number of input nodes (N_I), the number of hidden nodes (N_H), the number of weights (N_W), the number of training examples (Q_{TRN}) and the number of test examples (Q_{TST}) are shown in Table 1.

3.3 Network training

Training a network means finding a set of weights that minimizes the average error on the training set. The training is done iteratively, by showing the network known input-output pairs, calculating the error and updating the

weights accordingly. The weights are not updated after every input-output pair but after a number of examples, an epoch, which here was chosen as 500. The weight changes are given by the error derivatives and the weight changes in the preceding iteration,

$$\Delta w(i+1) = -\eta \cdot \frac{dE}{dw} + \alpha \cdot \Delta w(i), \tag{5}$$

where η and α are the learning rate and the momentum. The error derivatives are calculated according to the error back-propagation algorithm. The learning parameters are chosen according to a simple rule of thumb suggested by Lundstedt and Wintoft (1994):

$$\eta = \frac{1}{Q \cdot N}, \tag{6}$$

$$\alpha = 1 - \frac{1/N}{0.1}, \tag{7}$$

where Q is the epoch size and N is the fan-in, i.e. the number of connections going into a node.

The weights were initiated to random values in the interval

$$\left[-\frac{1}{\sqrt{N}}, +\frac{1}{\sqrt{N}}\right], \tag{8}$$

in order to keep the typical nodal input somewhat less than unity (Hertz *et al.*, 1991).

The total amount of training data includes 6607 h from 62 different periods. The length of each period is reduced by the size of the input-data window τ_w, i.e. there is less data available when using larger input-data windows, as shown in Table 1.

3.4 Network testing

The real test of a fully trained neural network is how well it can be expected to perform on inputs for which the output is not known in advance. We then need a statistically fair sample of input-output pairs which has not been shown to the network during training. For this purpose we used 22 periods covering 2085 h. These test data were not included in training the network.

The performance of the networks was checked according to three criteria: correlation coefficient (r) between measured and computed D_{st}, average relative variance (ARV) and the RMS error (RMSE). These are defined by

$$\mathrm{r} = \frac{1}{Q_{TST}} \cdot \frac{\sum_{\mu} (O^{\mu} - \langle O \rangle)(D_{st}^{\mu} - \langle D_{st} \rangle)}{\sigma_O \cdot \sigma_{Dst}}, \tag{9}$$

$$\mathrm{ARV} = \frac{\sum_{\mu} (O^{\mu} - D_{st}^{\mu})^2}{\sum_{\mu} (D_{st}^{\mu} - \langle D_{st} \rangle)^2}, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{Q_{TST}} \sum_{\mu} (O^{\mu} - D_{st}^{\mu})^2}, \tag{11}$$

where the sums include all the examples in the test set. $\langle O \rangle$ and $\langle D_{st} \rangle$ are the averages of the network output O and the desired output D_{st}, respectively; σ_O and σ_{Dst} are the corresponding standard deviations.

4 Results

4.1 General

Figure 4a–f shows correlation plots for the test data. Each of the plots includes the whole test set. The overall performance of the six trained networks is also shown in Table 2 and Fig. 2.

The most striking result is the better performance of networks with larger temporal size of the input-data sequence, τ_w. With τ_w large enough, we could reproduce 84% of the variance of the D_{st} index (i.e. $r^2 \approx 0.84$). The improvements with increased τ_w are significant in all three performance criteria. This is, however, only valid up to a certain level of τ_w, up to somewhere between 15 and 20 h as shown in Fig. 2. No further improvement is achieved by increasing τ_w above this level.

In order to gain more insight into the physical reason for the improved performance when increasing τ_w, we have to study predictions of individual geomagnetic storms (Fig. 5a–f).

All networks succeed in predicting the initial and main phases of the geomagnetic storms. The predicted onset and strength of the main phase is well correlated to the

Table 2. Overall performance of the six trained networks. See text for the explanation of r, ARV and RMSE

Network	r	ARV	RMSE/nT
NET4	0.84	0.30	22
NET8	0.87	0.24	19
NET12	0.91	0.17	16
NET16	0.92	0.15	16
NET20	0.92	0.15	15
NET24	0.92	0.15	15

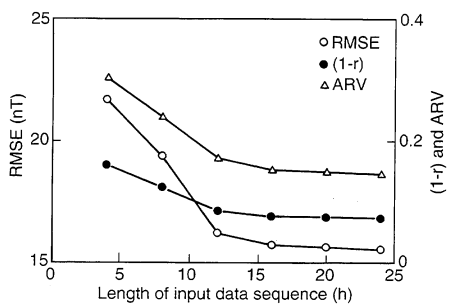

Fig. 2. Overall performance using different lengths of input-data sequence. The scale to the left measures the RMSE, while the scale to the right measures both ARV and 1 − r

measured D_{st}. The initial phases are predicted, but the predictions are in some cases smaller than the measured initial phases. The recovery phase is well predicted for some of the networks, but not for others, and here we find the main reason for the difference in performance between networks using different sizes of input-data sequences. A typical example of this behaviour can be seen in Fig. 5a–f showing the measured and computed D_{st} indices for a major geomagnetic storm. The storm starts by a density peak at the same time as a velocity increase, thus creating a peak in the dynamic pressure. B_z turns southward and the main phase starts. After a few hours B_z slowly starts to increase until at 46 h it turns northward. From now, and some hours forward in time, the geomagnetic disturbance level is mainly controlled by the slow decay of the ring current. When only a small part of the solar-wind history is available to the network, the predicted recovery phase ends abruptly after only a short time. Increasing the length of the time history makes the predictions more accurate. The improvements are noticeable up to a length of the time history somewhere between 15 and 20 h. This is the same conclusion as can be drawn from the overall performance results.

The initial and main phases, on the other hand, seem to be well predicted, almost independent of the length of the available solar-wind history. This is a consequence of the fact that the connection between the solar wind and the geomagnetic disturbance is more direct during these phases.

4.2 *Influence of the solar-wind history*

Most of the geomagnetic storm behaviour can be understood in terms of two concurrent mechanisms: (a) compression of the magnetosphere caused by an increased solar-wind dynamic pressure and (b) build-up and subsequent loss of the ring current. There is an important difference between these two mechanisms. The first gives a *direct* connection between solar-wind conditions and ground-level geomagnetic field disturbance; the second includes delays and time-dependent transport and dissipation processes, such that the geomagnetic disturbance becomes a function not only of the current solar-wind conditions, but also of the *solar-wind history*. These time dependences introduced by magnetospheric processes are particularly important during the recovery phase of geomagnetic storms.

To predict accurately D_{st} at a certain time, the network must have information from which it can calculate the amount of energy that has been injected into the ring current, and at what time it was injected. The length of the magnetospheric "memory" then determines the necessary length of the solar-wind history. This "memory" has a finite length due to dissipation. It is determined by the efficiency of the ring-current loss processes, which can be quantified by the ring-current decay time. The length of the solar-wind time sequence, used as input to the network, has to be a significant fraction of the decay time. This demand is a consequence of the underlying physics of the problem and not a limitation of the neural networks.

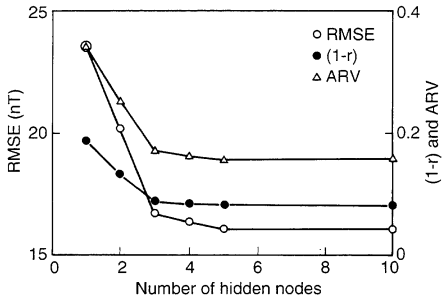

Fig. 3. Overall performance for networks with 16 h of input data and a varying number of hidden nodes. The scale to the left measures the RMSE, while the scale to the right measures both ARV and 1 − r. The number of hidden nodes has to be less than three to give a significant decrease in performance

If this outline of the necessary amount of input data is correct, then we would expect the predictions to be more accurate the larger input-data sequence we use, up to a certain limit. This is also what we saw in both the overall results and in studies of individual geomagnetic storms. The saturation of the performance measures (Fig. 2) at 15–20 h gives the approximate length of the magnetospheric "memory" as seen by the neural networks.

4.3 *Influence of the number of hidden nodes*

One disadvantage of using a large input-data sequence is that the number of input nodes, and thus also the number of weights, becomes large. If we want the number of weights to be a specific fraction of the number of training data, then we have to remove hidden nodes as the size of the input-data sequence increases. There could then be a risk that the network loses its ability to model the full complexity of the problem. In practice this seems not to be the case. Starting with NET16 and varying the number of hidden nodes from 10 down to 1 results in the RMSEs, ARVs and correlation coefficients shown in Fig. 3. The number of hidden nodes has to be less than three to give a significant decrease in performance, which is much less than the number used in any of the networks in this study.

5 Discussion and conclusions

Based on these results, we can now summarize the abilities of the networks:

- The initial and main phases of geomagnetic storms are well predicted with only 2 to 4 h of solar-wind data available.
- The recovery-phase predictions are improved by the availability of a larger part of the solar-wind history. The improvements are significant up to 15–20 h of solar-wind data, which is the approximate length of the magnetospheric "memory" as seen by the neural networks.

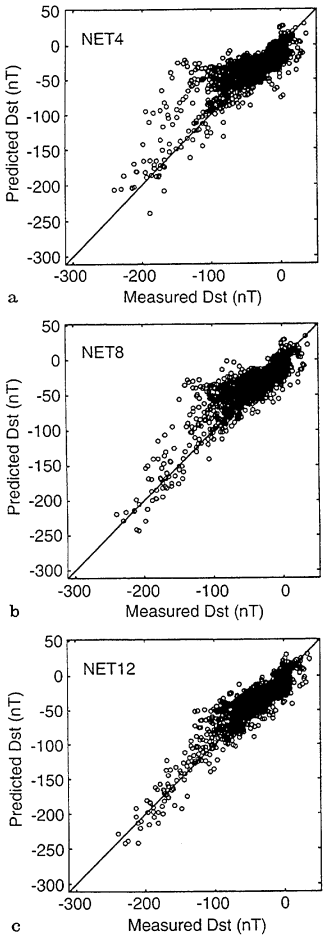

Fig. 4a–c. Test-data correlation plots for the networks NET4, NET8 and NET12. The plots include all test data

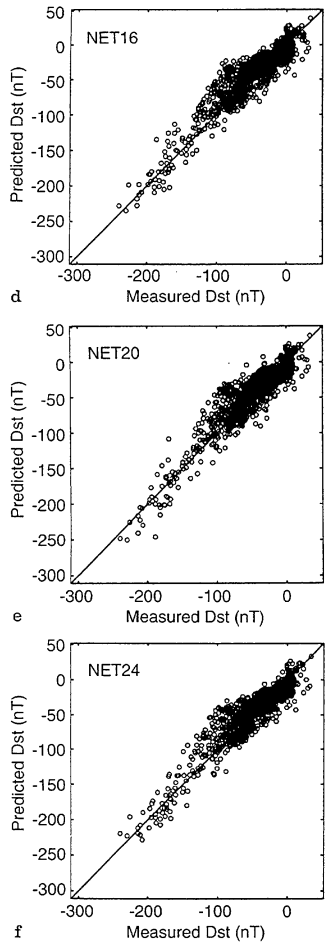

Fig. 4d–f. Test-data correlation plots for the networks NET16, NET20 and NET24. The plots include all test data

- Up to 84% of the variance of the D_{st} index was reproduced, using a large test set consisting of 2100 h of varied solar-wind and geomagnetic conditions.
- The neural networks that can make these excellent predictions are simple. The number of hidden nodes can be small, suggesting a fairly simple relationship between the solar wind and the D_{st} index (using 1-h averaged data). The size of the network is determined by the necessity of using 15 to 20 h of solar-wind data.

How do these results compare to previous studies using other methods?

Using linear filters with solar-wind dynamic pressure and the dawn-to-dusk component of the interplanetary electric field, McPherron *et al.* (1986) and Fay *et al.* (1986), both found that they could account for 70% of the D_{st} variance. When studying the two filters (i.e. the P_{dyn}- and the E filters) separately, McPherron *et al.* found that the dynamic-pressure filter had a width of only about 10 min, while the electric-field filter had a very long duration. They drew the conclusion that the characteristics of the electric-field filter are due to the long time constants associated with the decay of the ring current. This is in accordance with our finding that the recovery-phase predictions are very much dependent on the ring-current history, and thus also the solar-wind history.

Another method is to fit an analytical expression to solar-wind and geomagnetic-activity data. An example is Gonzalez *et al.* (1989), who systematically tested a number of analytical formulas for the prediction of the D_{st} index. They occasionally found correlation coefficients above 0.90 for individual storms, but the average over a more diverse set of storm-time periods was considerably lower. A major advantage of the analytical formulas of Gonzalez *et al.* is that they explicitly reveal the quantitative depen-

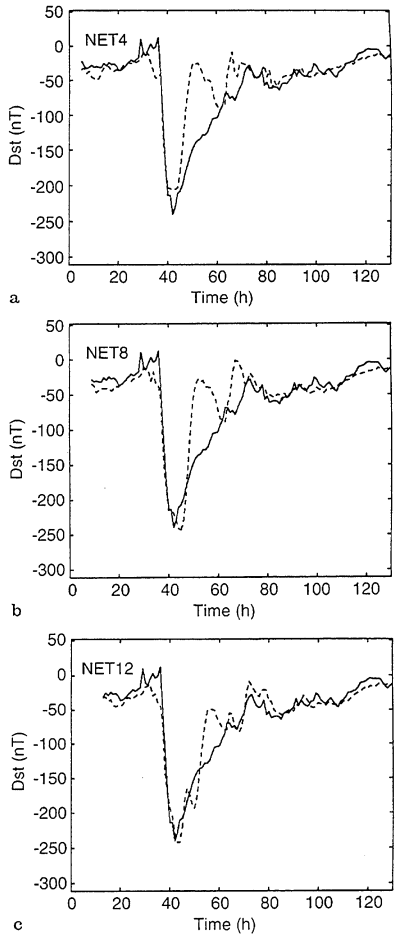

Fig. 5a–c. The measured (*filled line*) and predicted (*dotted line*) D_{st} indices for a major geomagnetic storm. The predictions are made with the networks NET4, NET8 and NET12

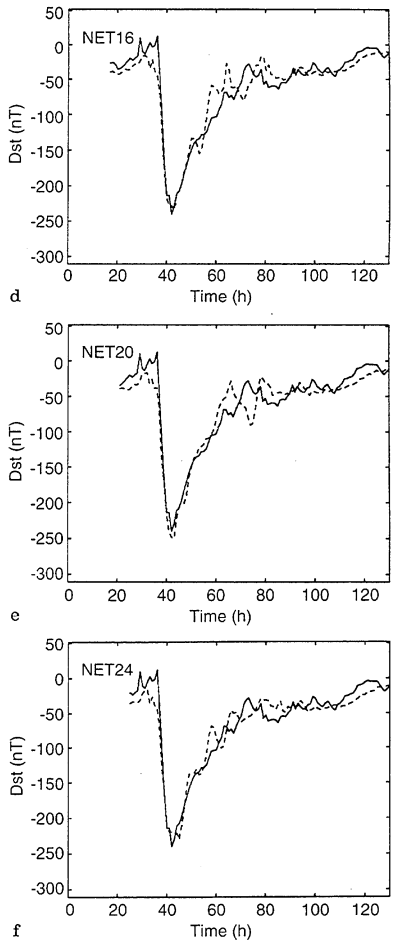

Fig. 5d–f. The measured (*filled line*) and predicted (*dotted line*) D_{st} indices for a major geomagnetic storm. The predictions are made with the networks NET16, NET20 and NET24

dence on the most important solar-wind parameters. The only previous paper claiming equal correlations to the present study is Goertz *et al.* (1993). They used an empirical non-linear model to predict the auroral electrojet index AE, and found a correlation coefficient of 0.92. The stated correlation was, however, criticized by McPherron and Rostoker (1993), based on an assumed biased selection of test data.

To compare different prediction methods is a difficult task. The correlation between measured and computed geomagnetic-activity indices depends on the type of index, averaging of data, as well as the statistical properties of the sample used for testing the methods. In the present study we have predicted 1-h averages, which is considerably less complicated than, for example, predictions of high time-resolution quantities such as AU, AL or AE. We have been very careful in selecting a large, varied and unbiased test set. The presented correlations should therefore be valid for continuous predictions made during a long time span.

A practical use of the presented neural-network technique would be real-time predictions of the geomagnetic activity one hour ahead. This will need a spacecraft continuously monitoring the solar wind at the sun-earth libration point L_1. To make accurate predictions during the recovery phase, 15 to 20 h of continuous solar-wind measurements are necessary. However, to predict the initial and main phases of a geomagnetic storm, 2 to 4 h of continuous measurements are enough.

The neural network, based only on measurements, can be seen as a purely empirical model of the whole chain of dynamical processes connecting the solar wind with the inner magnetosphere and the ring current. We can therefore use it to investigate the solar wind-magnetosphere

coupling and the magnetospheric dynamics controlling the energy flow from the solar wind to the ring current. Based on selected periods of well-behaved solar-wind data, it should be possible to use the networks for various sensitivity studies. The goal of such studies could be the variations of the injection rate Q and the decay rate τ, with the strength of the storms or the solar-wind conditions in general.

Another interesting application of time-delay neural networks is the prediction of high time-resolution magnetic data connected with the substorm phenomenon. The successful development of a method to predict such geomagnetic quantities could be a step toward an improved understanding of substorms and substorm triggering mechanisms.

Acknowledgements. Topical Editor K.-H. Glaßmeier thanks M. Paley and P. Newell for their help in evaluating this paper.

References

Akasofu, S.-I., Energy coupling between the solar wind and the magnetosphere, *Space Sci. Rev.*, **28,** 121–190, 1981.

Akasofu, S.-I., and S. Chapman, The ring current, geomagnetic disturbance and the van Allen radiation belts, *J. Geophys. Res.*, **66,** 1321–1350, 1961.

Bargatze, L. F., D. N. Baker, R. L. McPherron, and E. W. Hones Jr., Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res.*, **90,** 6387–6394, 1985.

Burton, R. K., R. L. McPherron, and C. T. Russel, An empirical relationship between interplanetary conditions and D_{st}, *J. Geophys. Res.*, **80,** 4204–4214, 1975.

Clauer, C. R., The technique of linear prediction filters applied to studies of solar wind-magnetosphere coupling, in *Solar wind-magnetosphere coupling*, eds. Y. Kamide and J. A. Slavin, Terra Scientific Publishing, Tokyo 150, Japan, 39–57, 1986.

Couzen, D. A., and J. H. King, *Interplanetary Medium Data Book – Supplement 3 1977–1985*, NSSDC/WDC-A-R&S, Goddard Space Flight Center, Greenbelt, MD, 1986.

Detman, T. R., W. D. Gonzalez, and A. G. C. Gonzalez, Prediction of the geomagnetic (Dst) index by adaptive filtering of solar-wind data, in *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, eds. J. Joselyn, H. Lundstedt and J. Trolinger, Lund, Sweden, 1993.

Fay, R. A., C. R. Garrity, R. L. McPherron, and L. F. Bargatze, Prediction filters for the Dst index and the polar-cap potential, in *Solar wind-magnetosphere coupling*, eds. Y. Kamide and J. A. Slavin, Terra Scientific Publishing, Tokyo 150, Japan, 111–117, 1986.

Feldstein, Y. I., V. YU. Pisarsky, N. M. Runeva, and A. Grafe, Ring-current simulation in connection with interplanetary space conditions, *Planet. Space Sci.*, **32,** 975–984, 1984.

Freeman, J., and A. Nagai, The magnetospheric specification and forecast model: moving from real-time to prediction, in *Proceedings of Solar-Terrestrial Workshop in Ottawa May 18–22, 1992*, ed. M. A. Shea, NOAA, 1992.

Goertz, C. K., L.-H. Shan, and R. A. Smith, Prediction of geomagnetic activity, *J. Geophys. Res.*, **98,** 7673–7684, 1993.

Gonzalez, W. D., B. T. Tsurutani, A. L. C. Gonzalez, E. J. Smith, F. Tang, and S.-I. Akasofu, Solar wind-magnetosphere coupling during intense magnetic storms (1978–1979), *J. Geophys. Res.*, **94,** 8835–8851, 1989.

Hertz, J., A. Krogh, and R. Palmer, *Introduction to the theory of neural computation*, Addision-Wesley, Redwood City, CA 94065, 1991.

Iyemori, T., and H. Maeda, Prediction of geomagnetic activities from solar-wind parameters based on the linear prediction theory, in *Solar-Terrestrial Predictions Proceedings, Vol. 4*, U.S. Dept. of Commerce, Boulder, CO, A-1-A-7, 1980.

Iyemori, T., H. Maeda, and T. Kamei, Impulse response of geomagnetic indices to interplanetary magnetic fields, *J. Geomagn. Geoelectr.*, **31,** 1–6, 1979.

Lundstedt, H., A trained neural network, geomagnetic activity and solar-wind variation, in *Proceeding of Solar-Terrestrial Workshop in Ottawa May 18–22, 1992*, ed. M. A. Shea, NOAA, 1992a.

Lundstedt, H., Neural networks and predictions of solar-terrestrial effects, *Planet. Space Sci.*, **40,** 457–464, 1992b.

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

McPherron, R. L., and G. Rostoker, Comment on "Prediction of geomagnetic activity" by C. K. Goertz, Lin-Hua Shan, and R. A. Smith, *J. Geophys. Res.*, **98,** 7685–7686, 1993.

McPherron, R. L., D. N. Baker, and R. F. Bargatze, Linear filters as a method of real time prediction of geomagnetic activity, in *Solar wind-magnetosphere coupling*, eds. Y. Kamide and J. A. Slavin, Terra Scientific Publishing, Tokyo 150, Japan, 85–92, 1986.

Ogilvie, K. W., L. F. Burlaga, and T. D. Wilkerson, Plasma observations on Explorer 34, *J. Geophys. Res.*, **73,** 6809–6824, 1968.

Rostoker, G., and C.-G. Fälthammar, Relationships between changes in the interplanetary magnetic field and variations in the magnetic field at the earth's surface, *J. Geophys. Res.*, **72,** 5835–5863, 1967.

Rumelhart, D. E., G. Hinton, and R. Williams, Learning representations by back-propagating errors, *Nature*, **323,** 533–536, 1986.

Siscoe, G. L., V. Formisano, and A. J. Lazarus, Relation between geomagnetic sudden impulses and solar-wind pressure changes – an experimental investigation, *J. Geophys. Res.*, **73,** 4869–4874, 1968.

Sugiura, M., Hourly values of equatorial Dst for the IGY, *Ann. Int. Geophys. Year*, **35,** 1964.

Vassiliadis, D., A. J. Klimas, D. N. Baker, and D. A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res.*, **100,** 3495–3512, 1995.

Williams, D. J., The Earth's ring current: causes, generation and decay, *Space Sci. Rev.*, **34,** 223–234, 1983.

II

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 102, NO. A7, PAGES 14,269–14,278, JULY 1, 1997

Response of the auroral electrojets to the solar wind modeled with neural networks

H. Gleisner and H. Lundstedt

Lund Observatory, Lund, Sweden

Abstract. The dissipative processes in the Earth's magnetosphere, such as the ring current and the auroral electrojets, depend on both the external solar wind forcing and factors internal to the magnetosphere. Previous studies have shown that artificial neural networks are able to compute the ring current index Dst very accurately from only solar wind data. In this study, we use neural networks to model the response of the auroral electrojets to the solar wind conditions. The solar wind input to the networks consist of 5-min averaged data from the Earth-orbiting spacecraft IMP 8, while the output is the auroral electrojet index AE. The relationships between the solar wind and the AE index, as modeled by the neural networks, are investigated in a parameter study. The relative importance of individual solar wind variables is studied, as well as the abilities of various coupling functions. It is shown that the use of individual solar wind variables as input to a neural network is superior to the use of corresponding coupling functions. The nonlinear neural networks are related to earlier linear techniques, and the abilities of linear networks (linear filters) are compared to those of nonlinear networks. It is found that a nonlinear network with n, V, B_y, and B_z as input during 100 min can account for 76% of the variance ($r \approx 0.87$) in the AE index. No influence of B_x is found. With the coupling function $p^{1/2}V^2B_s$ as input to a nonlinear network, 71% of the AE index variance is predicted. These results are averaged over a large test set ($\sim$ 330 hours) of data not used to train the networks. The test data are from 1973-1974 and include a diverse set of conditions, ranging from almost quiet to exceptionally disturbed.

1. Introduction

Efforts to predict geomagnetic activity have led to many correlation studies using a rich variety of techniques (e.g., the collection of papers edited by *Kamide and Slavin* [1986]). The geomagnetic activity has mostly been quantified by some global index that measures the effects of the major current systems in the magnetosphere and ionosphere. Two of the most widely used indices are Dst and AE [*Baumjohann*, 1986]. Dst measures the geomagnetic activity at low latitudes and responds most strongly to the ring current and the magnetopause currents. AE measures geomagnetic activity at auroral latitudes and responds to the convection electrojets (the DP 2 current system) and the substorm electrojets (the DP 1 current system). While Dst has been shown to be fairly easy to correlate to solar wind data [*Burton et al.*, 1975; *Iyemori et al.*, 1979; *Clauer*, 1986], the response of AE to the solar wind conditions has proven to be less easy to determine [*Holzer and Slavin*, 1982; *Clauer*, 1986]. One reason for these difficulties could be that there are two types of substorm-like magnetic signatures contributing to the AE index [*Pytte et al.*, 1978]: one is directly driven by the solar wind and is caused by modulations of the enhanced convection that occurs when the interplanetary magnetic field has a southward component, while the other includes current-wedge formation and near-midnight magnetic disturbances. Whether the latter type of geomagnetic disturbance is caused solely by some internal instability or triggered by external changes in the solar wind is a matter of much controversy.

Although the solar wind is known to be the primary source of energy that drives the dissipative processes in the magnetosphere, there still remains fundamental questions concerning how the energy is transferred from the solar wind and how it is further transformed into the various geomagnetic activity signatures. A widely used approach has been to combine a few relevant solar wind variables into a coupling function. The linear correlation between this coupling function and a geomagnetic activity index has then been calculated, after including a proper time delay. Following the demonstration by *Arnoldy* [1971] of a close relationship between AE and the rectified dawn-to-dusk component of the interplanetary electric field, many coupling functions have been investigated. With the introduction of the linear filter technique by *Iyemori et al.* [1979], the linear correlation studies were extended to take into account a whole time series of solar wind input, still in the form of coupling functions.

The abilities of the linear filters depend on the dynamical properties of the magnetosphere, particularly the linearity and time invariance of the magnetospheric response to the solar wind. In a linear filter study by *Bargatze et al.* [1985], it was shown that the magnetospheric response to the solar

Paper number 96JA03068.
0148-0227/97/96JA-03068$09.00

wind conditions varies with the level of geomagnetic activity. Further, as first shown by *Russell and McPherron* [1973], there is clear evidence for the existence of a component in the geomagnetic activity that is not directly driven by the solar wind. These findings imply a nonlinear and time-varying magnetospheric response that can not be properly modeled by linear filters.

During the last years, interest has turned toward nonlinear dynamical methods. Two approaches have emerged: analogue modeling and data-based phase space reconstruction. The recent development of analogue modeling started when *Baker et al.* [1990] adapted a dripping faucet analogue model to describe magnetospheric dynamics. This work was followed by *Goertz et al.* [1993] with a directly driven model of the AE index and by *Klimas et al.* [1992, 1994] with the so-called Faraday loop model. All these analogue models consist of low-dimensional systems of ordinary differential equations, which have the advantage that a physical interpretation is made possible. This is contrary to data-based input-output analysis methods, such as phase-space reconstruction, which lack an immediate physical interpretation but that instead have the advantage that system characteristics are determined directly from empirical data. Data-based phase-space reconstruction was used by *Price and Prichard* [1993] after they pointed out the inadequacy of treating the magnetospheric system as autonomous. Several input-output studies followed [*Price et al.*, 1994; *Vassiliadis et al.*, 1995]. Then in 1993, *Hernandez et al.* [1993] described a study where two types of neural networks were used to model the AL index, but the results were not conclusive. However, other studies have shown that artificial neural networks are able to compute the Dst index from solar wind data very accurately [*Lundstedt and Wintoft*, 1994; *Gleisner et al.*, 1996; *Wu and Lundstedt*, 1996]. It thus seems appropriate to continue a further exploration of neural network-based analysis methods applied to high time resolution auroral-zone geomagnetic activity.

The emphasis of this paper is on the abilities of fairly standard neural networks as an empirical model of the solar wind forcing of the auroral electrojets. The strength of the electrojets is quantified by the 5-min averaged AE index. After training the networks, they are evaluated in terms of the correlation between the observed and the computed AE indices. The relative importance of individual solar wind variables is studied, as well as the abilities of various coupling functions. Linear and nonlinear networks are compared and the qualitative agreement between the observed and the computed AE is studied during 2 days in March 1974. It is the purpose of the present study to show the usefulness of nonlinear neural network models and to point out some possible physical interpretations of the results.

2. Artificial Neural Networks

Some of the most widely used artificial neural network models have much in common with the various filters that have been applied to magnetospheric physics. The standard neural network techniques (multilayer feed-forward and partly recurrent networks) can be regarded as nonlinear generalizations of linear filters. In this paper, we use both linear and nonlinear feed-forward neural networks, where the linear neural networks correspond to linear filters.

2.1. Linear and Nonlinear Filters

The linear moving-average (MA) filter, and its nonlinear generalizations, is based on the assumption that the geomagnetic activity O can be described as a function of a time series of solar wind variables I,

$$O_t = F(I_{t-1}, I_{t-2}, \ldots, I_{t-T}), \tag{1}$$

where T is the length of the magnetospheric system memory for previous inputs. No geomagnetic activity variables are included among the independent variables. The discrete linear MA filter output is given by

$$O_t = \sum_{\tau=1}^{T} (H_\tau I_{t-\tau}) \tag{2}$$

that is, the impulse response function of the magnetospheric system, H_t, is convolved with a sequence of earlier solar wind inputs. For a filter to be linear, H_t must be time invariant and exhibit no dependence on the solar wind input. However, it has been shown [*Bargatze et al.*, 1985] that the empirical impulse response function depends on the level of geomagnetic activity, an indication that the real magnetospheric system is in fact nonlinear.

The nonlinear filter can be cast in many forms. One of the simplest forms of nonlinearity is to approximate the nonlinear response F locally by linear filters of the type given by equation 2. To give a nonlinear response, H_t must then depend on the solar wind input [*Vassiliadis et al.*, 1995]. As shown below, the multilayer feed-forward neural network can also be regarded as a nonlinear generalization of the linear MA filter [*Hertz et al.*, 1991; *Hernandez et al.*, 1993].

2.2. Feed-Forward Neural Networks

2.2.1. General. A feed-forward neural network [*Hertz et al.*, 1991] is a collection of processing nodes arranged in layers (Figure 1). The input to each node is the sum of the weighted outputs from all the nodes in the previous layer and the activity of each node is passed on to all the nodes in the following layer. The output from a node is given by the input to the node and the nodal activation function, which is a differentiable, saturating function. An additional node, the bias node, is set to 1 and connected to all hidden and output nodes in the network (Figure 1). The purpose of this is to adjust the nodal activation functions.

The key to network performance is the weights determining the strength of the connection between nodes. Since this network type belongs to the class of supervised networks, it is trained by adjusting the weights until the average error on a set of known training examples is minimized. The most common training algorithm, which is also used in this study, is a modified form of gradient descent called error back-propagation [*Rumelhart et al.*, 1986].

The neural networks used in the present study all have one hidden layer and one output layer. The input data to the networks are organized as a temporal sequence, where input data sampled during a time window of length L are shown to

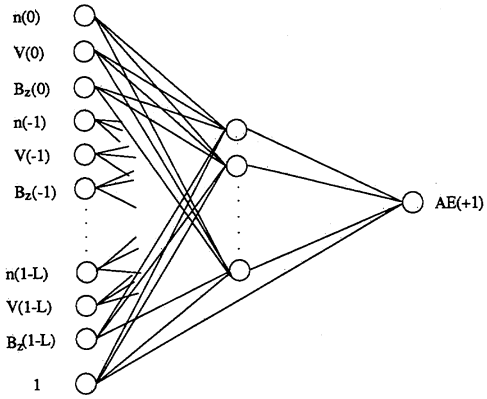

Figure 1. A network with (left) input nodes and (right) the single output node. In the feed-forward phase, the input is propagated through the network to the output. The error between the computed output and the observed AE is propagated backward through the network, and the weights are updated accordingly. The bias node is always set to 1 as indicated. The temporal length L of the input data sequence varies from 20 to 100 min.

the network simultaneously. To get a time sequence of output data, this window is moved stepwise in time. A feed-forward neural network, together with this type of organization of the input data, is often referred to as a time delay neural network.

For an input data vector, $\{\xi_k^\mu; k = 1, 2, \ldots, m\}$, with m components, the network output is given by

$$O_i^\mu = g_O \left[\sum_j W_{ij} g_H \left(\sum_k w_{jk} \xi_k^\mu \right) \right]. \quad (3)$$

Each input-output pair $\{\xi_k^\mu, O_i^\mu\}$ is labeled by superscript μ. Index i refers to a node in the output layer, index j refers to a hidden layer node, and index k refers to an input layer node. The weight W_{ij} thus connects a hidden layer node with an output layer node, while w_{jk} connects input and hidden layer nodes. Here g_H and g_O are the activation functions for nodes in the hidden layer and output layer, respectively. In the present study, g_H is the hyperbolic tangent function, and g_O is a linear function. The network output is then given by

$$O^\mu = \sum_j \left[W_j \tanh \left(\sum_k w_{jk} \xi_k^\mu \right) \right], \quad (4)$$

where index i has been omitted since the output vector consists of a single value, the predicted geomagnetic index. If both activation functions g_O and g_H are linear, then we can write the network output

$$O^\mu = \sum_j \left[W_j \left(\sum_k w_{jk} \xi_k^\mu \right) \right] = \sum_k \left(\sum_j W_j w_{jk} \xi_k^\mu \right). \quad (5)$$

As the input vector ξ_k^μ represents a time series of data, we can identify the inner sum of equation 5 with the impulse response coefficients of equation 2

$$H_t = \sum_j W_j w_{jt}. \quad (6)$$

The feed-forward neural network with linear activation functions is obviously identical to a linear filter. With nonlinear activation functions, the neural network can be regarded as a nonlinear generalization of the basic linear filter.

2.2.2. Network setup. After a certain network architecture has been specified, in this case a feed-forward network with one hidden layer, the number of nodes in each layer has to be determined. In the output layer, there is only a single node, the predicted geomagnetic index. The number of nodes in the input layer is determined by the number of input data. This in turn is determined by the temporal length of the input data time series, L, and the set of solar wind variables included in the time series. Using, for example, 5-min averages of n, V, and B_z during $L = 60$ min makes a total number of 36 nodes in the input layer.

The size (i.e., the number of weights) of a network is only determined by the number of hidden nodes as the number of input and output nodes are given. The number of weights in the network has to be large enough to represent the full complexity of the problem, and it has to be small enough not to overfit and lose generalization ability. As a rough rule of thumb, the number of weights in the network should be less than one tenth the number of training data [*Lundstedt and Wintoft*, 1994], which in this study means that the number of weights should be less than ~1690. All the networks we use here have eight hidden nodes. With this number of hidden nodes, we avoid the problems with too few or too many weights in the networks. This matter is further discussed in section 4.1.4 where we show eight hidden nodes to be a good choice.

2.2.3. Network training. Training a network means finding a set of weights that minimizes the average error on the training set. The training is done iteratively by showing the network known input-output pairs, calculating the network error and updating the weights accordingly. The weights are not updated after every input-output pair but after a number of examples, an epoch, which here is chosen as 1000 examples. The weight changes are given by the error derivatives and the weight changes in the preceding iteration,

$$w_{t+1} = w_t + \Delta w_t = w_t - \eta \frac{\partial E}{\partial w} + \alpha \Delta w_{t-1}, \quad (7)$$

where the constants η and α are referred to as the learning rate and the momentum. The network error is defined as the sum of the errors over an epoch,

$$E = \frac{1}{2} \sum_\mu (O^\mu - AE^\mu)^2, \quad (8)$$

where O^μ is the actual output of the network and AE^μ is the corresponding "correct" output. The error derivatives are calculated according to the error back-propagation algorithm [*Rumelhart et al.*, 1986], and the learning parameters are chosen as

$$\eta = \frac{1}{QN}, \quad (9)$$

$$\alpha = 0.90, \quad (10)$$

where Q is the epoch size and N is the fan-in, the number of connections going into a node. This choice of learning rate has been used earlier by *Gleisner et al.* [1996] in a similar study, and it is more thoroughly discussed by *Hertz et al.* [1991]. The weights are initiated to random values in the interval

$$\left[-\frac{1}{\sqrt{N}}, +\frac{1}{\sqrt{N}}\right] \tag{11}$$

in order to keep the typical nodal input somewhat less than unity [*Hertz et al.*, 1991].

2.2.4. Network testing. Much of the practical use of neural networks is based on their ability to make sensible generalizations. This ability can be formally defined as the average network performance on a randomly chosen new data point. The true generalization ability can not be known exactly, but it can be estimated by the network performance on the test set, a set of randomly chosen data not included in the training set. To get a good estimate of the generalization ability, the test set has to be large enough to be representative in a statistical sense.

The abilities of the networks are quantified with three diagnostics: correlation coefficient (r) between observed and computed AE, average relative variance (ARV), and the RMS test error ($RMSE$). These are defined by

$$r = \frac{1}{Q_{TST}} \left(\frac{\sum_{\mu} (O^{\mu} - \langle O \rangle)(AE^{\mu} - \langle AE \rangle)}{\sigma_O \sigma_{AE}} \right), \tag{12}$$

$$ARV = \frac{\sum_{\mu} (O^{\mu} - AE^{\mu})^2}{\sum_{\mu} (AE^{\mu} - \langle AE \rangle)^2}, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{Q_{TST}} \sum_{\mu} (O^{\mu} - AE^{\mu})^2}, \tag{14}$$

where the sums include the whole test set. The averages of the computed output O and the observed output AE are denoted $\langle O \rangle$ and $\langle AE \rangle$, respectively, while σ_O and σ_{AE} are the corresponding standard deviations.

3. Geomagnetic and Solar Wind Data

The data used in the present study span the interval from November 1973 to December 1974. The basic geomagnetic data consist of the 2.5-min averaged AE index obtained from World Data Center C1 for Solar-Terrestrial Physics in England. The AE database is complete and contains no data gaps. Each pair of neighboring 2.5-min values was combined into a 5-min average. The present study is based on these 5-min averages.

The solar wind data consist of 5-min averaged solar wind plasma and interplanetary magnetic field (IMF) parameters. These are from the Earth-orbiting spacecraft IMP 8 and are obtained from the National Space Science Data Center in Greenbelt, Maryland. The solar wind plasma data include the bulk velocity V, the proton number density n, and the fraction of He^{2+} ions in the solar wind. The IMF components B_x, B_y, and B_z are expressed in geocentric solar magnetospheric (GSM) coordinates. On the basis of these variables, we computed solar wind coupling functions such as VB_s, where B_s is defined as $B_s = -B_z$ when $B_z < 0$ and $B_s = 0$ when $B_z > 0$. Some coupling functions also included θ, the polar angle of the IMF vector projected onto the Y-Z plane in the GSM system.

Since the solar wind data contain numerous gaps, a selection based on data quality was made. The data were scanned to compile a list of intervals that were at least 24 hours long, contained less than 10% missing data, and contained no data gaps longer than 3 samples (i.e., 15 min). A search through the 14-month period gave 39 intervals covering 20,900 samples, 1740 hours. Thirtytwo of these intervals are used as training data (16,900 samples) and seven as test data (4000 samples).

Most of the intervals are separated from each other by 4 days or more since IMP 8 is unable to measure the solar wind conditions during 4 to 8 days in each 12.5-day orbit. However, two of the test intervals are separated from the preceding training intervals by 3 and 4 hours, respectively. As pointed out by *Vassiliadis et al.* [1995], the separation between training and test periods must be larger than the autocorrelation length of the AE index to be certain that the training and test data are uncorrelated. For the AE index, this length is 1 to 3 hours. Considering that the separations are larger than the autocorrelation length and that the test data intervals are very long, the results should not be biased by such an effect.

4. Studies and Results

4.1. Parameter Studies

A neural network should do more than just be a good "predictor." It should also be a tool to improve our understanding of the physics that control the solar wind coupling to the auroral electrojets. Here we perform a few parameter studies to investigate the abilities of neural networks as predictors, while at the same time point out some possible physical interpretations of the results. Training a sequence of networks using different lengths of the solar wind history tells us something about the timescales of the dissipation processes that determine the magnetospheric system memory. Similarly, varying the solar wind parameters used as input to the networks tells us something about the physics of the energy transfer to the magnetosphere. Also, the question of linearity or nonlinearity of magnetospheric processes can be addressed by a suitable parameter study.

Such studies raise the question of the stability of the results. Will they still hold with another choice of network architecture? This question can also be addressed by a parameter study. Training a sequence of networks on one specific problem, while varying the network architecture, gives us an estimate of the influence our choice of network architecture has on the results. In this study, the network architecture is varied by the number of hidden nodes and thus also the number of weights in the network.

The performances of the networks referred to in this section are defined by equations 12, 13, and 14. They are calculated over the whole test set, and so they constitute re-

alistic performances averaged over both quiet and disturbed times.

4.1.1. Solar wind history and the magnetospheric system memory. Many correlation studies have used solar wind data at a single point in time to cross correlate with a geomagnetic activity index at a slightly later point in time. At some time delay Δt, mostly between 20 and 60 min, a maximum correlation of 0.50-0.70 was found [e.g., *Baker et al.*, 1981, 1983]. By applying filters or neural networks to the same problem, a whole time history of solar wind inputs can be used. This allows us to study the magnetospheric system memory of previous inputs, which in turn depends on the efficiency and timescales of the dissipation processes in the magnetosphere.

For each set of solar wind variables, we have trained a sequence of networks using different temporal lengths of the solar wind history. The length L varies from 20 to 100 min. All plots of a network performance measure versus L show the same general characteristics as in Figures 2 and 3. The network performance improves with increasing L until it saturates at $L \approx 100$ min. The network performance is not improved by including solar wind data older than 100 min. This is interpreted as the length of the magnetospheric system memory for previous inputs, as seen by the neural network. The same result is reached with all three network diagnostics: r, $RMSE$, and ARV. The empirical value of the length of the magnetospheric system memory found here is of the same order as the linear filter timescales ($\sim$ 2 hours; *Bargatze et al.* [1985]), the nonlinear filter timescales ($\sim$ 1.5 hours; *Vassiliadis et al.* [1995]), and also as estimates of the total duration of substorms ($\sim$ 2-4 hours; *Lui* [1991]).

4.1.2. Solar wind input parameters. The proper choice of variables to use as input to the networks is determined by the actual mechanisms of energy transfer from the solar wind to the magnetosphere. These are only partly known and several mechanisms may operate concurrently. After 3 decades of solar wind measurements, we know that n, V, and B_z control much of the geomagnetic activity [*Snyder et al.*, 1963; *Arnoldy*, 1971]. The IMF component B_y has also been shown to exert an influence on the magnetosphere and ionosphere [e.g., *Heppner*, 1972]. Other possible candidates are the IMF component B_x and the ionic composition of the solar wind plasma. As the solar wind undergoes a shock before arrival at the magnetopause, the plasma temperature is strongly determined by the solar wind bulk speed V. The temperature is thus considered unimportant.

The neural networks are first trained with five different sets of solar wind input variables according to Figure 2. The variables are shown to the networks separately. They are not combined into coupling functions. After training, we study the network diagnostics r, $RMSE$, and ARV. The results, in terms of the correlation coefficient r, are shown in Figure 2. With only V and B_z as input, the correlation is 0.83. Adding either n or B_y increase the correlation somewhat, while adding both n and B_y gives the highest correlation, 0.87 (network E in Figure 2). Using B_x, in addition to V and B_z, does not improve the correlation. The correlations referred to here are for 100 min of solar wind input. With a shorter solar wind history, the correlations are correspondingly lower.

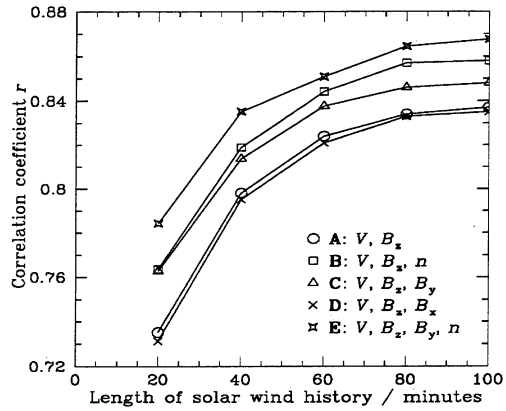

Figure 2. The correlation between observed and computed AE versus the length of the solar wind input sequence, for each set of solar wind variables listed in the figure. No assumptions were made about the combination of the solar wind variables used as input to the networks.

Can the results in Figure 2 tell us something about the coupling of the auroral electrojets to the solar wind? We have to remember that the neural networks reveal covariations rather than causal relationships, even though causality often underlies the covariations. The first network in Figure 2 use V and B_z, while the second use V, B_z, and n. Adding n gives an improvement. This suggests that the density n has a component that varies together with AE, independent of V and B_z. From networks A and C in Figure 2, the same is suggested for B_y. Similarly, networks C and E in Figure 2 tell us that there is a component of n that varies together with AE, independent of V, B_z, and B_y. The IMF component B_x does not improve the correlation. Altogether, this study suggests that all four variables, n, V, B_y, and B_z vary together with AE partly independent of the other variables, while B_x does not. Our interpretation is that all four variables, n, V, B_y, and B_z add some important information about the solar wind input. This is basically a confirmation of the generally accepted view put forward by many authors. However, the significant influence of the density found here, is not often stressed in discussions of the solar wind forcing of the auroral electrojets. In terms of an improved correlation, the effect of adding the density n is comparable to the effect of adding the IMF component B_y. Also, the absence of an influence from B_x is worth noting.

A standard approach in studies of the interaction between the solar wind and the magnetosphere has been to combine a few solar wind variables into a coupling function. The linear correlation between the coupling function and a geomagnetic activity index has then been calculated. Some of the most widely used coupling functions are VB_s [*Burton et al.*, 1975], V^2B_s [*Murayama et al.*, 1980], $\varepsilon \propto VB^2 sin^4(\theta/2)$ [*Perreault and Akasofu*, 1978], $p^{1/2}VB_s$ [*Murayama*, 1986], and $p^{1/6}VBsin^4(\theta/2)$ [*Bargatze et al.*, 1986]. Here p denotes the dynamical pressure. Most coupling functions have been developed based on an idea of which energy coupling mechanisms are most important. *Vasyliunas et al.* [1982]

argued that the coupling functions should have dimensions of power, and from dimensional analysis they showed that many of the previously used expressions were dimensionally incorrect. Of the coupling functions mentioned above, only the last one is dimensionally correct according to *Vasyliunas et al.* [1982], equation 12. However, it can be questioned whether this dimensional requirement is valid when the purpose is to compute a quantity such as AE that does not have dimensions of power and that certainly does not depend linearly on the magnetospheric energy input. Further, as was also pointed out by Vasyliunas et al., the dimensional analysis is not equally applicable to all timescales. The dimensional equality holds only if no energy is intermediately stored before dissipation. The timescale for energy storage in the magnetotail is of the order of 1 hour. It could then be argued that dimensional analysis does not impose any serious restrictions on coupling functions that connect the solar wind conditions with the high-time resolution AE index. In this study, we therefore investigate coupling functions regardless of their dimensions. Instead, we systematically test functions of the form $n^{\alpha}V^{\beta}B_s$ together with VB_z and the coupling function ε.

In the second part of the parameter study, the networks are trained with time series of solar wind coupling functions as input. The parameters and the results are shown in Figure 3. The networks with VB_s and VB_z as input perform equally well, an indication that AE is nearly independent of B_z when the IMF is directed northward. The solar wind parameter ε gives approximately the same correlation as VB_s when using 100 min of input data. As shown in Figure 3, the networks with ε as input are less sensitive to the length of the solar wind history than the networks that use other coupling functions. Such behavior would be seen for a parameter with a long-correlation length in the solar wind. Since ε includes B_y, which tends to have a longer-correlation length than B_z, we can speculate that this is indeed the reason for the relative insensitivity of ε to the history length L. This property gives ε a relative advantage when using a very short solar wind history. However, from a physical point of view, ε is not fully comparable to the other coupling functions since it also includes B_y.

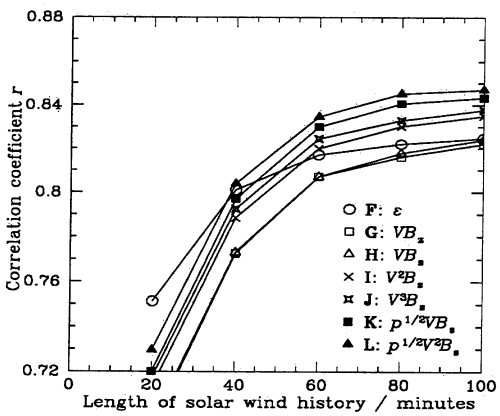

Figure 3. The correlation between observed and computed AE versus the length of the solar wind input sequence, for each solar wind variable listed in the figure. The basic solar wind variables were here combined to coupling functions which were used as input to the networks.

Among the coupling functions that do not contain any correction for the density variations, V^3B_s turns out to be the best, marginally better than V^2B_s. It is interesting to note that V^3B_s performs equally well to V and B_z given as separate inputs, when using 100 min of input data. However, when using a shorter solar wind history, there is a discrepancy between V^3B_s and the individual variables V and B_z. This discrepancy becomes systematically larger with a smaller L. The same behavior is also seen for V^2B_s, with just a marginally lower correlation than V^3B_s.

Two of the coupling functions include corrections for variations in the solar wind dynamic pressure. Simple scalings of VB_s and V^2B_s with $p^{1/2}$ give $n^{1/2}V^2B_s$ and $n^{1/2}V^3B_s$, respectively. Both of these are superior to the other functions, and $n^{1/2}V^3B_s$ provides the best coupling function in this study.

4.1.3. Linear versus nonlinear networks. The standard feed-forward neural network can be regarded as a nonlinear generalization of the linear filter. It is in fact a very general nonlinear model. Since the problem of computing the AE index from solar wind data is inherently nonlinear, we expect neural networks to perform better than linear filters. However, in a study by *Hernandez et al.* [1993], it was found that a linear filter actually performed slightly better than a neural network. The reason was that large amplitude variations were clipped by the nonlinear networks, while the linear networks showed no such tendencies. Hernandez et al. concluded that further exploration of this issue is necessary.

To address this question, two sequences of networks are trained, each with n, V, B_y, and B_z as input. The networks in one sequence have linear activation functions, while the other networks have the usual nonlinear activation functions in the hidden layer. The performances of the networks are shown in Figure 4. It is found that the nonlinear networks perform significantly better than the linear networks, both on individual substorms and as an overall result. This is true also for other choices of input data, such as the coupling functions in Figure 3. Further, we could not find that the nonlinear networks were more prone to cut large amplitude variations than the linear networks. Both types of networks fail to predict the largest variations of the highest frequencies, but the nonlinear networks are always better than the linear.

4.1.4. Number of hidden nodes. As the number of input data to a network is increased, the number of weights is also increased. With a large number of free parameters in the network, overfitting problems may arise with devastating effects on the generalization performance [*Hertz et al.*, 1991]. There is also a lower limit to the number of hidden nodes and the number of weights in a network. The number of free parameters of the network has to be large enough to represent the full complexity of the problem. In most neural network studies, it is essential to know these lower and upper limits to the number of hidden nodes and the number of weights.

All networks discussed above have eight hidden nodes. Since the number of input variables varies from 4 to 80, there is an accompanying large variation in the number of weights. Will this variation cause any systematic effects in

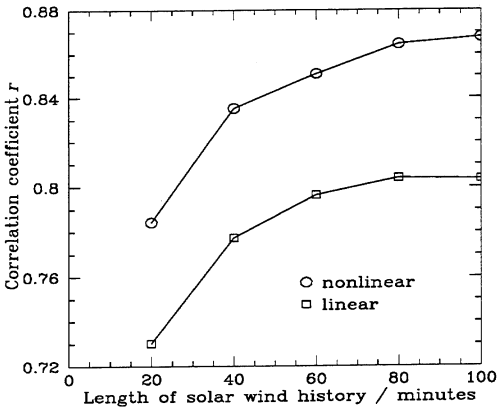

Figure 4. Comparison of linear and nonlinear networks. Both sequences of networks used n, V, B_y, and B_z as input. The networks with nonlinear activation functions performed significantly better than the linear networks. Other input data gave similar results.

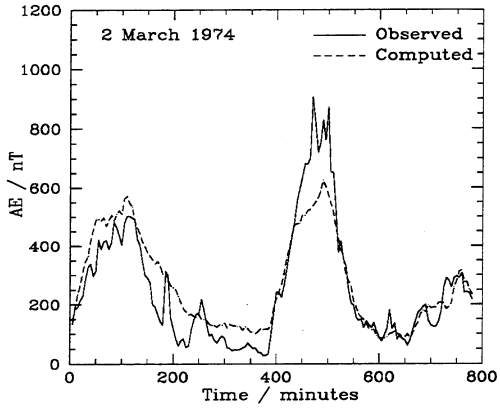

Figure 6. Observed and computed AE index using n, V, B_y, and B_z during 100 min as input. The interval started on the morning of March 2 (0810 UT), and ended in the evening the same day (21.05 UT).

the parameter studies above? Would the results and the general conclusions still hold with other choices of the number of hidden nodes and thus also the number of weights? A partial answer to these questions is given in Figure 5. Using the solar wind variables n, V, B_y, and B_z as input, two sequences of networks are trained, one with 20 min, and the other with 100 min of input data. For each of these sequences, the number of hidden nodes is varied from 1 to 16. In both cases, the number of hidden nodes has to be less than five to show any decrease in performance. With 100 min of input data, there is a tendency of overfitting when using more than 14 to 16 hidden nodes. There is such a tendency with 20 min of input data as well, although not so clear as in the 100 minute case. There has to be more than 16 to 20 hidden nodes to show clear signs of overfitting. By our choice of eight hidden nodes in all the networks, we avoid the problems of too few or too many weights, both for those networks with few input variables and those with many. We thus conclude that there are no systematic effects due to a varying number of weights in the networks that alter the conclusions made in the parameter studies.

4.2. Qualitative Abilities of the Networks

In addition to some general comments on the qualitative abilities of the networks, we have chosen 2 consecutive days in March 1974 to represent the results. The 2 days are divided into 2 intervals: the first begins on the morning of March 2 (0810 UT) and the second begins in the evening the same day (2135 UT).

4.2.1. General. As seen in Figures 6 and 7, the fit between observation and prediction is far from perfect. During

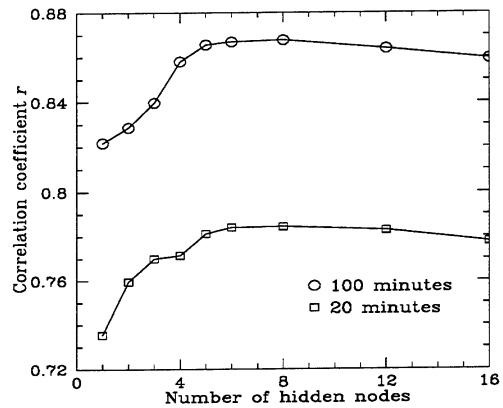

Figure 5. Network performance with 20 and 100 min of solar wind input data and a varying number of hidden nodes and thus also a varying number of weights in the networks. In both cases, the number of hidden nodes had to be less than five to show a decrease in performance.

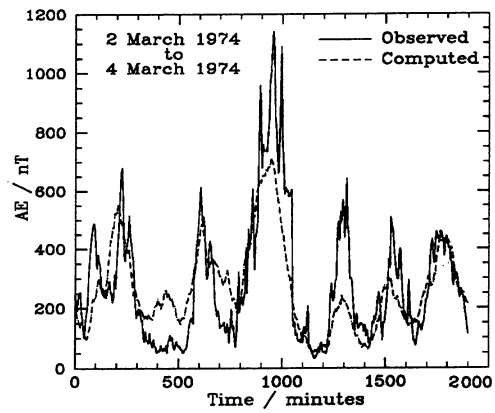

Figure 7. Observed and computed AE index using n, V, B_y, and B_z during 100 min as input. The interval started in the evening of March 2 (2135 UT) and ended in the early morning of March 4 (0510 UT).

disturbed times, the observed AE index often exceeds the predicted AE by quite a large factor. It is obvious that some structures in the AE index are missed by the networks. Such structures are the large and sudden excursions mostly attributed to intensifications of the westward electrojet caused by substorm expansions. However, the reverse relation between observation and prediction does not seem to occur. The networks very rarely predict an intensification where no such structures are seen in the observed AE index. There are occasions, such as the one in Figure 6, when the predictions somewhat overestimate AE, but the overestimates are always small and connected with broad features with durations longer than 1 hour. Only substorms that lack large and sudden excursions in the AE index, are overestimated.

4.2.2 March 2. The interval is shown in Figure 6. It stretches over 800 min and covers two substorms, one peaking at ~500 nT and the other at ~900 nT. The computed AE (dotted line) looks very much like a smoothed version of the observed AE. The network handles the broad features well, while narrow features are harder to predict. The first peak is one of the few that the network actually overpredicts, although not very much. Common features that are unpredicted are the large and sudden excursions, such as those seen at 190 min and at the top of the second peak. In the 5-min averaged AE data, the narrowest peaks consist of only one value rising above the other.

4.2.3 March 2 to March 4. The interval is shown in Figure 7. It stretches over 2000 min from the evening of March 2 to early morning of March 4. It covers six broad peaks. The highest of them reach ~1150 nT, while the others reach ~500 to 700 nT. This interval is somewhat more disturbed than the first, and the number of large and sudden excursions is correspondingly higher. It is even more obvious during this interval that the network acts as a low-pass filter and flattens out the high-frequency structures. For three of the six peaks, the predictions almost reach the same height as the observed AE, while the other three are flattened out below the observed AE.

5. Discussion and Conclusions

The work described here demonstrates the abilities of artificial neural networks as predictors of an auroral electrojet index of high-time resolution. The networks are tested on a large (~ 330 hours) and nonselected data set from 1973-1974, a period that includes both quiet and exceptionally disturbed conditions. Various solar wind inputs are used, and some are found superior. The results of this study can be summarized:

1. Nonlinear networks are superior to linear networks.

2. Individual solar wind variables as input are superior to composite variables, such as the commonly used coupling functions.

3. One hundred minutes of solar wind data are required as input to the networks. This is interpreted as the length of the magnetospheric system memory for previous inputs.

4. With the solar wind variables n, V, B_y, and B_z as input to the network, 76% of the AE index variance is accounted for. All four variables are needed. Removal of any of them impairs the network performance.

5. The IMF component B_x does not improve the network performance.

6. Among the coupling functions with no correction for the density variations, V^3B_s and V^2B_s give the highest correlations.

7. A simple scaling with $p^{1/2}$ improves the coupling functions. Consequently, $p^{1/2}VB_s$ is superior to VB_s, and $p^{1/2}V^2B_s$ is superior to V^2B_s. The best coupling function in this study is $p^{1/2}V^2B_s$.

8. The coupling function ε does not perform as well as V^3B_s or V^2B_s with 100 min of input data. Since it is relatively insensitive to the history length L, it has, however, a relative advantage over the other coupling parameters for short input data sequences.

That nonlinear networks are found to be superior to their linear counterparts should come as no surprise, as there have been many suggestions of a nonlinear component in the magnetospheric response to the solar wind conditions. At the same time, it has been suggested that nonlinear feed-forward neural networks do not offer any advantages over linear filters in terms of prediction accuracy [*Hernandez et al.*, 1993]. This is not what we found in the present study. Further explorations are necessary to find out why our conclusions differ.

An advantage that the neural network technique offers is that it allows the use of individual solar wind variables as input, rather than some coupling function. We do not have to assume very much about the energy-coupling mechanisms. When combining individual solar wind variables to a coupling function, some information on the solar wind is lost. If the coupling function does not fully describe the energy coupling between the solar wind and the magnetosphere, some of the lost information may turn out as essential. We would therefore expect networks using coupling functions never to perform better than networks with individual variables as input. This is also what we found here. We also found that the best coupling functions can be nearly as good as the corresponding individual variables when using a long history length L, while they are less successful for shorter L. The relation between the coupling function and the individual variables vary with the history length L, as is evident from a comparison of Figures 2 and 3.

While there are no theoretical reasons to use coupling functions as input to the neural networks, there are some important practical aspects to consider. As described in section 4.1.4, the number of weights in the network is not allowed to increase too much. It is often important to keep the number of input nodes small, while still give the network all essential data. The advantage of using coupling functions is that it reduces the number of input nodes and thus the number of weights in the network.

The three graphs that show a correlation plotted against the length of the solar wind history, show the same general trends; the correlation increases with L until it saturates at $L \approx 100$ min. This is consistent with results from previous studies using other techniques and is interpreted as the length of the magnetospheric system memory for previous inputs. The "magnetospheric system" in this context includes only the part of the magnetosphere that controls the solar wind-auroral electrojet coupling. There are also magnetospheric processes that develop on longer timescales, such as the growth and decay of the ring current and the associated geomagnetic storm. Influences from such longer timescale processes are not modeled by feed-forward neural networks with an integration time of only 100 min. There are, however, other types of neural network architectures that can take different timescales into account.

With the solar wind variables n, V, B_y, and B_z as input to the network during 100 min, a correlation coefficient $r \approx 0.87$ was found. This means that 76% of the variance of the AE index is accounted for. We also found that removal of any of the four variables makes the network less accurate. Our interpretation is that all four variables, n, V, B_y, and B_z add some important information about the solar wind input. We could, however, not find any influence from B_x. The present study also shows that V^2B_s, and rather surprisingly also V^3B_s, is superior to both VB_s and ε. As discussed in section 4.1.2, the dependence of the correlation on history length L is weaker for ε than for the other coupling functions, possibly as a result of a longer-correlation length in the solar wind. The coupling functions can be further improved by including a correction for the solar wind dynamic pressure. A simple scaling of V^2B_s with $p^{1/2}$ gives the parameter $n^{1/2}V^3B_s$ which provides the best coupling function in this study. VB_s is also improved by such a scaling. With $n^{1/2}V^3B_s$ as input to a network during 100 min, 71% of the AE index variance is accounted for.

How do these results compare to previous studies using other techniques? Using linear filters with various coupling functions as input and AE or AL as output, the prediction accuracy reported is around 40% [*Clauer*, 1986; *McPherron et al.*, 1988]. In a recent linear filter study, *Blanchard and McPherron* [1995] were able to predict 47% of the variance in the AL index from a time series of VB_s. These prediction accuracies refer to data for which the linear filters have not been specially fitted. The filters have generally been found to vary with the level of geomagnetic activity, and the response is bimodal for moderate levels of activity and unimodal for high levels of activity [*Bargatze et al.*, 1985; *McPherron et al.*, 1988].

The nonlinear studies that followed have led to an increased prediction accuracy. *Vassiliadis et al.* [1995] used both nonlinear moving-average filters and nonlinear state-input models to describe the response of AL to the solar wind input at a 2.5-min resolution. For out-of-sample predictions, the nonlinear moving-average filter accounted for 67% of the AL index variance, while the single-step state-input model had a prediction accuracy of 86%. For multiple-step predictions, this accuracy decreased to around 70%. *Hernandez et al.* [1993] used neural networks to predict AL at a 2.5-min resolution. They reported a prediction accuracy of 76% for a nonlinear feed-forward network and slightly higher for a linear network. The only previous paper claiming significantly higher correlations than the present study is *Goertz et al.* [1993]. They used a low-dimensional analogue model to predict the auroral electrojet index AE at a time resolution of 2.5 min and found a correlation coefficient of 0.92 (i.e., $r^2 \approx 85\%$). However, the stated correlation was criticized by *McPherron and Rostoker* [1993], based on the fact that all data had been filtered to remove high-frequency components. Together with a partly biased selection of test data and a nearly constant solar wind velocity, this made *McPherron and Rostoker* [1993] conclude that the equations and parameters would not do well for other intervals.

To compare different prediction methods is a difficult task. The correlations between observed and computed geomagnetic activity indices depend on the type of activity index, averaging of data, and the statistical properties of the samples used for training and testing. In the present study we have used the 5-min averaged AE index. We have been cautious to use a large and nonselected test set, not used in training the networks. The presented correlations should therefore be valid for continuous predictions made during a long time span.

In summary, we have shown the usefulness of artificial neural networks as predictors of the AE index. We have investigated some properties of these predictors and the abilities of various solar wind input data, including some of the frequently used coupling functions. The importance of a long enough solar wind history has been stressed. Two applications of the neural networks suggest themselves. The first is forecasting of the geomagnetic activity 30 to 70 min ahead. A prerequisite for this is a spacecraft continuously monitoring the solar wind at the Sun-Earth libration point L_1. With the use of an Earth-orbiting spacecraft, the lead time is correspondingly shorter. A major advantage of the feed-forward neural network in real-time applications is the very fast processing of data. The filtering of the solar wind input through the network will not be a limiting part of a future real-time forecast system. The second application is to study the solar wind forcing of the auroral electrojets by simulated input data. The neural network can be regarded as an empirical description of the dynamical processes connecting the solar wind with the auroral electrojets. The dynamical properties of the neural network should thus be able to reveal something of the dynamical properties of the real magnetosphere.

Acknowledgments. The editor would like to thank D. Vassiliadis and the other referee for their assistance in evaluating this paper.

References

Arnoldy, R.L., Signature in the interplanetary medium for substorms, *J. Geophys. Res., 76*, 5189, 1971.

Baker, D.N., E.W. Hones Jr., J.B. Payne, and W.C. Feldman, A high time resolution study of the interplanetary parameter correlations with AE, *Geophys. Res. Lett., 8*, 179, 1981.

Baker, D.N., R.D. Zwickl, S.J. Bame, E.W. Hones Jr., B.T. Tsurutani, E.J. Smith, and S.-I. Akasofu, An ISEE-3 high time resolution study of interplanetary parameter correlations with magnetospheric activity, *J. Geophys. Res., 88*, 6230, 1983.

Baker, D.N., A.J. Klimas, R.L. McPherron, and J. Büchner, The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos, *Geophys. Res. Lett., 17,* 41, 1990.

Baker, D.N., A.J. Klimas, and D. Vassiliadis, Global convection, low-dimensional magnetospheric dynamics, and deterministic chaos, in *Proceedings of the International Conference on Substorms II,* edited by S.-I. Akasofu and J.R. Kan, Univ. of Alaska, Fairbanks, 1994.

Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones Jr., Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res., 90,* 6387, 1985.

Bargatze, L.F., R.L. McPherron, and D.N. Baker, Solar wind-magnetosphere energy input functions, in *Solar Wind-Magnetosphere Coupling,* edited by Y. Kamide and J. A. Slavin, pp. 101-109, Terra Sci., Tokyo, 1986.

Baumjohann, W., Merits and limitations of the use of geomagnetic indices in solar wind-magnetosphere coupling studies, in *Solar Wind-Magnetosphere Coupling,* edited by Y. Kamide and J. A. Slavin, pp. 3-15, Terra Sci., Tokyo, 1986.

Blanchard, G.T., and R.L. McPherron, Analysis of the linear response function relating AL to VB_s for individual substorms, *J. Geophys. Res. 100,* 19155, 1995.

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

Clauer, C.R., The technique of linear prediction filters applied to studies of solar wind-magnetosphere coupling, in *Solar Wind-Magnetosphere Coupling,* edited by Y. Kamide and J. A. Slavin, pp. 39-57, Terra Sci., Tokyo, 1986.

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

Goertz, C.K., L.-H. Shan, and R.A. Smith, Prediction of geomagnetic activity, *J. Geophys. Res., 98,* 7673, 1993.

Heppner, J.P., Polar cap electric field distributions related to the interplanetary magnetic field direction, *J. Geophys. Res., 77,* 4877, 1972.

Hernandez, J.V., T. Tajima, and W. Horton, Neural net forecasting for geomagnetic activity, *Geophys. Res. Lett., 98,* 7673, 1993.

Hertz, J., A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation,* Addison-Wesley, Reading, Mass., 1991.

Holzer, R.E., and J.A. Slavin, An evaluation of three predictors of geomagnetic activity, *J. Geophys. Res., 87,* 2558, 1982.

Iyemori, T., H. Maeda, and T. Kamei, Impulse response of geomagnetic indices to interplanetary magnetic fields, *J. Geomagn. Geoelectr., 31,* 1, 1979.

Kamide, Y., and J.A. Slavin, *Solar Wind-Magnetosphere Coupling,* Terra Sci., Tokyo, 1986.

Klimas, A.J., D.N. Baker, D.A. Roberts, D.H. Fairfield, and J. Büchner, A nonlinear dynamical analogue model of geomagnetic activity, *J. Geophys. Res., 97,* 12253, 1992.

Klimas, A.J., D.N. Baker, D. Vassiliadis, and D.A. Roberts, Substorm recurrence during steady and variable solar wind driving: Evidence for a normal mode in the unloading dynamics of the magnetosphere, *J. Geophys. Res., 99,* 14855, 1994.

Lui, A.T., A synthesis of magnetospheric substorm models, *J. Geophys. Res., 96,* 1849, 1991.

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

McPherron, R.L., D.N. Baker, L.F. Bargatze, C.R. Clauer, and R.E. Holzer, IMF control of geomagnetic activity, *Adv. Space Res., 8(9),* 71, 1988.

McPherron, R.L., and G. Rostoker, Comment on "Prediction of geomagnetic activity" by C.K. Goertz, L.-H. Shan, and R.A. Smith, *J. Geophys. Res., 98,* 7685, 1993.

Murayama, T., T. Aoki, H. Nakai, and K. Hakamada, Empirical formula to relate the auroral electrojet intensity with interplanetary parameters, *Planet. Space Sci., 28,* 803, 1980.

Murayama, T., Coupling functions between solar wind and the Dst index, in *Solar Wind-Magnetosphere Coupling,* edited by Y. Kamide and J. A. Slavin, pp. 119-126, Terra Sci., Tokyo, 1986.

Perreault, P., and S.-I. Akasofu, A study of geomagnetic storms, *Geophys. J. R. Astron. Soc., 54,* 547, 1978.

Price, C.P., and D. Pritchard, The nonlinear response of the magnetosphere, *Geophys. Res. Lett., 20,* 771, 1993.

Price, C.P., D. Pritchard, and J.E. Bischoff, Nonlinear input/output analysis of the auroral electrojet index, *J. Geophys. Res., 99,* 13277, 1994.

Pytte, T., R.L. McPherron, E.W. Hones Jr., and H.I. West Jr., Multiple-satellite studies of magnetospheric substorms: Distinction between polar magnetic substorms and convection-driven magnetic bays, *J. Geophys. Res., 83,* 663, 1978.

Rumelhart, D.E., G. Hinton, and R. Williams, Learning representations by back-propagating errors, *Nature, 323,* 533, 1986.

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

Snyder, C.W., M. Neugebauer and U.R. Rao, The solar wind velocity and its correlation with cosmic-ray variations and with solar and geomagnetic activity, *J. Geophys. Res., 68,* 6361, 1963.

Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res., 100,* 3495, 1995.

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

Wu, J.-G., and H. Lundstedt, Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks, *Geophys. Res. Lett., 23,* 319, 1996.

H. Gleisner and H. Lundstedt, Lund Observatory, Box 43, S-22100 Lund, Sweden. (e-mail:hansg@astro.lu.se)

(Received April 29, 1996; revised July 15, 1996; accepted August 5, 1996.)

III

Ann. Geophysicae 17, 1268–1275 (1999) © EGS – Springer-Verlag 1999

Annales Geophysicae

Ring current influence on auroral electrojet predictions

H. Gleisner and H. Lundstedt

Lund Observatory, Box 43, S-22100 Lund, Sweden

Received: 25 August 1998 / Revised: 8 March 1999 / Accepted: 17 March 1999

Abstract. Geomagnetic storms and substorms develop under strong control of the solar wind. This is demonstrated by the fact that the geomagnetic activity indices *Dst* and *AE* can be predicted from the solar wind alone. A consequence of the strong control by a common source is that substorm and storm indices tend to be highly correlated. However, a part of this correlation is likely to be an effect of internal magnetospheric processes, such as a ring-current modulation of the solar wind-*AE* relation.

The present work extends previous studies of nonlinear *AE* predictions from the solar wind. It is examined whether the *AE* predictions are modulated by the *Dst* index.This is accomplished by comparing neural network predictions from *Dst* and the solar wind, with predictions from the solar wind alone. Two conclusions are reached: (1) with an optimal set of solar-wind data available, the *AE* predictions are not markedly improved by the *Dst* input, but (2) the *AE* predictions are improved by *Dst* if less than, or other than, the optimum solar-wind data are available to the net. It appears that the solar wind–*AE* relation described by an optimized neural net is not significantly modified by the magnetosphere's *Dst* state. When the solar wind alone is used to predict *AE*, the correlation between predicted and observed *AE* is 0.86, while the prediction residual is nearly uncorrelated to *Dst*. Further, the finding that *Dst* can partly compensate for missing information on the solar wind, is of potential importance in operational forecasting where gaps in the stream of real time solar-wind data are a common occurrence.

Key words. Magnetospheric physics (solar wind – magnetosphere interactions; storms and substorms).

Correspondence to: H. Gleisner
e-mail: hansg@astro.lu.se

1 Introduction

Variations in the solar wind can be detected at the Earth's surface as small disturbances of the main geomagnetic field. The disturbances are caused by variations in the strength and location of electrical currents flowing in the ionosphere and magnetosphere. These currents are energized by the solar-wind interaction with the magnetosphere and respond dynamically to variations of the solar-wind forcing.

At middle and low latitudes the *ring current* and the *magnetopause currents* dominate the geomagnetic records. At higher latitudes, a system of ionospheric *electrojet currents* and *field-aligned currents* is more pronounced. The complicated time-varying pattern of geomagnetic disturbances generated by these currents, is transformed into a number of *geomagnetic indices* that quantify the global level of geomagnetic activity (e.g., Mayaud, 1980; Baumjohann, 1986). Geomagnetic disturbances at low and middle latitudes are monitored by the *Dst* index at a relatively coarse 1-h resolution. At higher latitudes, transient disturbances are monitored by the *AL*, *AU*, and *AE* indices at a time resolution from one to a few minutes.

The modern definition of a *magnetic storm* is based on the strength of the ring current, as quantified by the *Dst* index (Gonzalez *et al.*, 1994). The *magnetic substorm* is defined from transient geomagnetic disturbances in the auroral zone (Rostoker *et al.*, 1980). Observations show that major storms are always accompanied by intense and frequent substorms, but that substorms can occur in the absence of a magnetic storm. The most intense substorms are usually found within the main phase of storms. In agreement with these observed storm/substorm relations, the *Dst* and *AE* indices tend to be correlated (e.g., Davis and Parthasarathy, 1967; Akasofu, 1981; Cade *et al.*, 1995).

The correlation between substorm and storm indices is largely a consequence of the fact that both processes

are controlled by the solar wind (McPherron, 1997), as demonstrated by the many studies of geomagnetic-activity prediction from the solar wind. A part of the correlation is, however, likely to be an effect of internal magnetospheric processes, such as a ring-current modulation of the solar wind–*AE* relation. In particular, the observed relations between *Dst* and the location of the maximum electrojet currents (Feldstein, 1992; Feldstein *et al.*, 1997) could play a role.

Predictions of *Dst* and *AE* from the solar wind alone demonstrate that magnetic storms and substorms are dynamically controlled by the solar wind. The first of many prediction studies were based on linear techniques: linear cross-correlations between solar-wind parameters and geomagnetic-activity indices (e.g., Arnoldy, 1971; Murayama, 1986) and linear moving-average filters (e.g., Bargatze *et al.*, 1985; Clauer, 1986; McPherron *et al.*, 1988). More recently, nonlinear input-state space reconstructions have been employed (Vassiliadis, 1993; Vassiliadis *et al.*, 1995).

During the last few years, prediction schemes based on artificial neural networks (ANNs) have been developed. The first ANN studies dealt with predictions of the *Dst* index from hourly averaged solar-wind data (Freeman *et al.*, 1993; Lundstedt and Wintoft, 1994; Wu and Lundstedt, 1996; Gleisner *et al.*, 1996). Correlations between observed and predicted *Dst* were as high as 0.91 over a large and varied test set covering all phases of the solar-activity cycle (Wu and Lundstedt, 1997). ANNs were found to give better predictions of the *Dst* index than other techniques.

ANNs were also applied to predictions of the auroral electrojet index *AL* at 2.5-min resolution (Hernandez *et al.*, 1993), but the results were not conclusive due to a serious clipping problem. Recently, Gleisner and Lundstedt (1997) showed that ANNs can be used to predict the 5-min *AE* index from solar-wind data, though not with the same high correlation as the hourly *Dst* index. They also identified an optimal set of solar-wind data (n, V, B_y, and B_z during 100 min) for use in *AE* predictions. Less information on the solar wind than the optimum, led to a decrease of prediction accuracy.

An advantage of the neural network technique is that a very diversified set of input data can be handled simultaneously. Any parameters that contribute information on the solar wind or on the magnetospheric state can be included in the input. The *Dst* index, either measured or predicted, can be used along with a sequence of solar-wind data as input to *AE* predictions. If, in fact, the ring current modulates the solar wind-*AE* relation, we can expect the networks based on both *Dst* and the solar wind to be superior to networks based on the solar wind alone.

The present paper address two aspects of nonlinear *AE* predictions. Firstly, we examine whether the *AE* predictions are improved by *Dst* when an optimal set of solar-wind data are available. A clear improvement would indicate that the solar wind-*AE* relation is significantly modified by the magnetosphere's *Dst* state. Secondly, we examine to what extent *Dst* can improve predictions when less than the optimum solar-wind data are available. As *Dst* indirectly contain information on recent solar-wind conditions, it is not unlikely that the *Dst* index can partly compensate for a loss of solar-wind data. Practical experiences of short-term forecasting based on real-time data (Gleisner and Lundstedt, not yet published) show that loss of information on the solar wind are a common occurrence. One often has to use less than, or other than, the optimum set of solar-wind parameters. Methods to make the networks more tolerant to loss of input information are therefore of potential importance in operational forecasting.

2 Artificial neural networks

The following description focuses on the particulars of the ANN models that are used in the present study. A broader view of artificial neural networks can be found in, e.g., Hertz *et al.* (1991).

2.1 Network setup

An artificial neural network is an assembly of interconnected nodes where the strength of the connection between any two nodes is determined by a modifiable weight (Fig. 1). Each node is fed by the sum of the weighted outputs from all the nodes in the previous layer, and pass on the output to all the nodes in the following layer. An additional node, the bias node, is set to 1.0 and connected to all hidden and output nodes in the network. The incoming signal at a node is processed by an activation function, usually a nonlinear, saturating function for a hidden node and a linear function for the output node.

The ANNs used in the present study all have one hidden layer and one output layer. For an input data

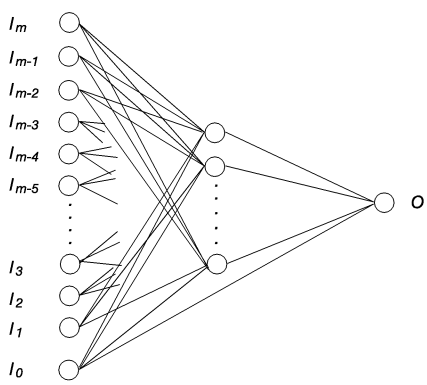

Fig. 1. Network with (*left*) input nodes and (*right*) the single output node. The input signal, $\{I_k; k = 1, 2, \ldots, m\}$, is propagated to the output, through the hidden nodes where the signal is transformed by nonlinear functions. The input I_0 is a bias that is set to 1.0 and connected to all hidden and output nodes in the network

vector, $\{I_k^\mu; k = 1, 2, \ldots, m\}$, with m components, the network output is given by

$$O^\mu = g_O\left[\sum_j W_j g_H\left(\sum_k w_{jk} I_k^\mu + \theta_j\right) + \theta\right] , \quad (1)$$

where

$$g_H(x) = \tanh(x); \quad g_O(x) = x . \quad (2)$$

Each input-output sample $\{I_k^\mu, O^\mu\}$ is labeled by superscript μ. Index j refers to a hidden layer node, index k refers to an input layer node, and in the output layer there is only a single node. The weight W_j thus connects a hidden layer node with an output layer node, while w_{jk} connects input and hidden layer nodes. The terms θ_j and θ are the weights associated with the bias input I_0.

2.2 Network training

Network training is the process of adjusting the weights until the network produces a response similar to the input-output samples in the *training set*. The network's ability to produce a correct output is monitored by the cost function

$$C(\mathbf{w}) \equiv \frac{1}{2Q_{\text{trn}}} \sum_{\mu=1}^{Q_{\text{trn}}} (O^\mu - T^\mu)^2 , \quad (3)$$

where O^μ is the actual output of the network, T^μ is the correct output (or "target"), and Q_{trn} is the number of samples in the training set.

This nonlinear optimization problem is solved using a modified gradient-descent method referred to as *error back-propagation* (Rumelhart *et al.*, 1986). The weights are iteratively adjusted according to the gradient-descent rule

$$w_{t+1} \leftarrow w_t + \Delta w_t; \quad \Delta w_t = -\eta\left(\frac{\partial C}{\partial w}\right)_t + \alpha \Delta w_{t-1} , \quad (4)$$

where η and α are constant parameters and t denote the iteration. In each iteration only a subset of the training set is used, and the weights are updated in an approximate gradient direction. This subset consists of Q_{bat} samples that in each iteration are randomly selected from the set of training data. The three parameters that control the training process have here been assigned the values

$$Q_{\text{bat}} = 1000 ,$$

$$\eta = 0.015 ,$$

$$\alpha = 0.90 .$$

Variations on this basic training procedure are more thoroughly discussed by Hertz *et al.* (1991).

2.3 Generalizing with a trained network

Much of the practical use of neural networks is based on their ability to make sensible generalizations. This ability can be formally defined as the average network performance on a randomly chosen new data sample (Hertz, 1993). The generalization ability can be estimated by the network performance on a *test set* which contain data that are not used during training.

The training procedure described above optimizes the network's ability to memorize the training data. In order to optimize the generalization ability, the training procedure needs to be constrained. This is done by excluding a small part of the training set from the actual training, and using these data (the *validation set*) to determine when to stop the iteration. In this way the problem of *overfitting* is avoided, or at least lessened.

In the present study, a network's generalization ability is quantified by two measures: the correlation coefficient between observed and predicted AE,

$$r = \frac{\frac{1}{Q_{\text{tst}}} \sum_{\mu=1}^{Q_{\text{tst}}} (O^\mu - \langle O \rangle)(AE^\mu - \langle AE \rangle)}{\sigma_O \sigma_{AE}} , \quad (5)$$

and the mean-squared error normalized by the variance of the observed AE data,

$$V_{rel} = \frac{\frac{1}{Q_{\text{tst}}} \sum_{\mu=1}^{Q_{\text{tst}}} (O^\mu - AE^\mu)^2}{\sigma_{AE}^2} . \quad (6)$$

Here, the averages of the computed output O (i.e., predicted AE) and the observed AE are denoted $\langle O \rangle$ and $\langle AE \rangle$, respectively, while σ_O and σ_{AE} are the corresponding standard deviations.

3 Data

3.1 Data sources and selection

The 5-min averaged solar-wind data were obtained from the IMP 8 database at NSSDC. In the present study, we used all intervals of data from a 14-month period (Nov. 1973 to Dec. 1974) that were at least 24 h long, contained less than 10% missing data, and contained no data gaps longer than 3 samples (i.e., 15 min). This selection gave 40 intervals covering 21600 samples: 32 intervals (1400 h) were used to train the networks and 8 intervals (400 h) were used as an independent test of network performance.

The solar-wind data included the proton number density n, the wind speed V, the three components of the interplanetary magnetic field, B_x, B_y, and B_z, given in the Geocentric Solar Magnetospheric (GSM) reference system. We also used the southward component of the magnetic field, B_s, defined as $B_s = -B_z$ when $B_z < 0$ and $B_s = 0$ when $B_z > 0$.

The AE data were obtained from World Data Center C1 in England. The original 2.5-min averages were averaged over 5 min to be consistent with the solar-wind data. The hourly Dst data, uncorrected for the solar-wind dynamic pressure, were provided by NSSDC through the OMNIweb database.

3.2 Data characteristics

The period of study (Nov. 1973 to Dec. 1974) was a geomagnetically very active period. The solar wind was largely dominated by long-lasting high-speed streams associated with coronal holes. Within the body of the high-speed streams, periods of large-amplitude Alfvén waves occurred, generating sustained substorm activity (Tsurutani *et al.*, 1995). The period of study also includes one major ($Dst = -204$ nT) geomagnetic storm, and several moderate (-100 nT $\leq Dst \leq -50$ nT) and weak (-50 nT $\leq Dst \leq -25$ nT) storms. The occurrences of *AE* and *Dst* during the 1800 h of study are shown in Fig. 2.

The tendency of substorms to become more frequent and intense during magnetic-storm conditions, is demonstrated by the correlation between *AE* and *Dst* over the 1800 h of data: $r = 0.58$ based on the 5-min *AE* index and $r = 0.62$ based on the hourly averaged *AE*.

4 *AE* prediction studies

All ANNs in the present study were trained with 16700 samples of the "correct" input-output relation and tested on 4700 input-output samples that were not used during training. One set of networks was trained with the solar-wind quantities used as individual input variables (the solid symbols in Fig. 4). Another set of networks was fed with coupling functions (the solid symbols in Fig. 5). A third set of networks was fed with both *Dst* and a sequence of solar-wind data (the open symbols in Figs. 4 and 5).

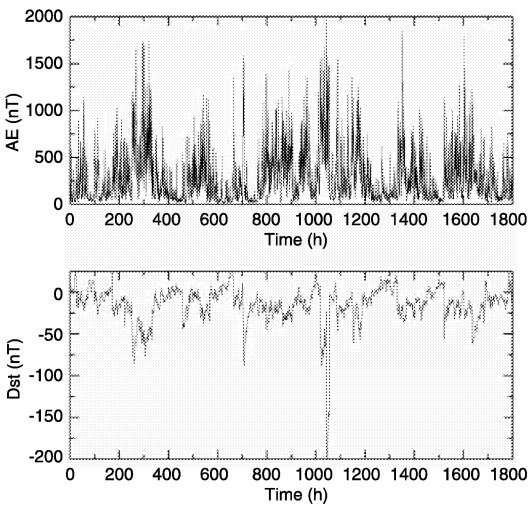

Fig. 2. Occurrences of *AE* and *Dst* during the period of study. The data consist of 40 intervals covering 1800 h selected from the period Nov. 1973 to Dec. 1974

4.1 Predictions from the solar wind alone

The basic properties of *AE* predictions with ANNs that are fed with solar-wind data have been demonstrated by Gleisner and Lundstedt (1997). Figure 3 shows an example of predictions with a network that is fed with 100 min of solar-wind n, V, B_y, and B_z. The predicted *AE* disturbances resemble a smoothed version of the observed disturbances. The accuracy of the predictions depend on the physics encoded into the network: the temporal length of the input sequence and the set of solar-wind variables being fed to the network.

The predictions improve with increasing temporal length of the input sequence for all sets of input variables that are used in this study. Predictions continue to improve up to an input sequence length $T \approx 100$ min. For much longer input sequences, the predictions starts to deteriorate as an increased number of weights in the networks makes the overfitting problem worse.

It is evident from Figs. 4 and 5 that there are significant differences between the various combinations of solar-wind parameters. The differences most likely reflect their different abilities to account for the actual mechanisms of energy transfer from the solar wind. The results in Fig. 4 show that it is essential to use all four variables n, V, B_y, and B_z. Although V and B_z are the most important quantities, the exclusion of n or B_y impairs the network performance.

Due to the risk of overfitting, the number of input variables should be as small as possible in order to minimize the number of weights. While the nets must be fed with all relevant information, this information should be given in the form of as few input parameters

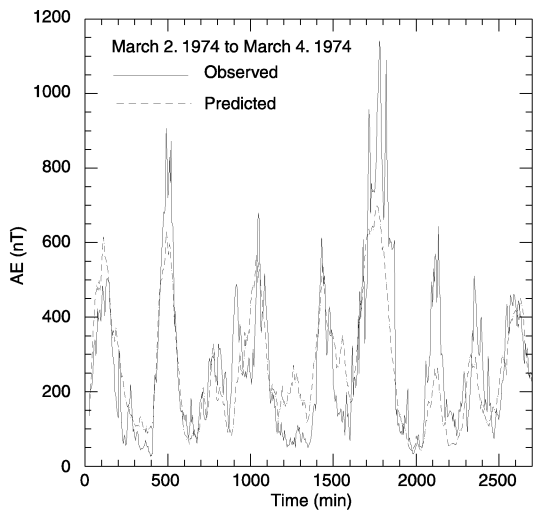

Fig. 3. Observed and predicted *AE* based on 100 min of solar-wind parameters n, V, B_y, and B_z. This particular interval started on the morning of 2 March 1974 (08.10 UT), and ended two days later on 4 March 1974 (07.15 UT)

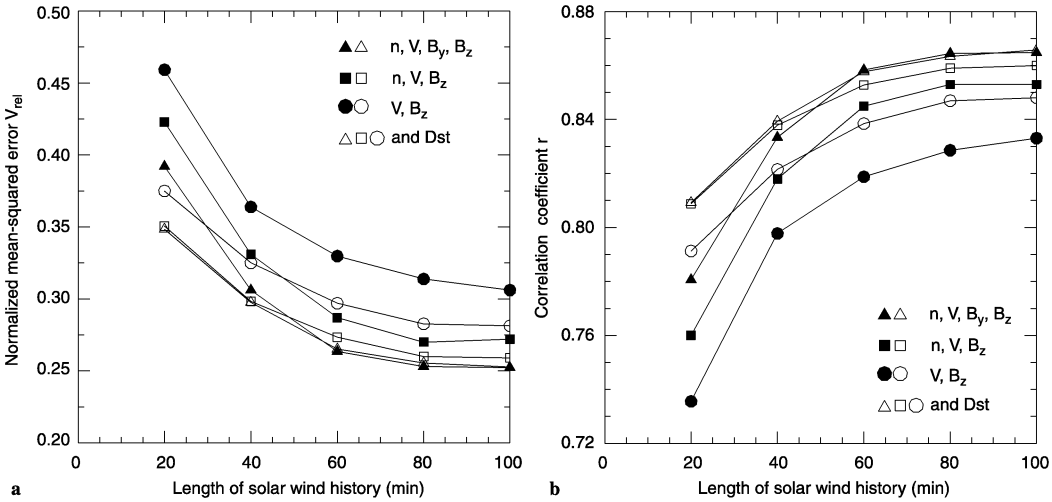

Fig. 4. **a** Normalized mean-squared error and **b** correlation between observed and predicted *AE*, for different sets of solar-wind variables and temporal lengths of the input sequence. The *solid symbols* mark predictions from the solar wind alone, while the *open symbols* mark predictions from *Dst* and the solar wind. The *AE* predictions are only improved by *Dst* if less than the optimum solar-wind data are available to the net

as possible. The temporal length of the input sequence is determined by the physics of the solar wind-auroral electrojet relation and cannot be reduced in any simple manner. One way of reducing the number of inputs is to combine individual solar-wind variables into coupling functions. This must be done with care, as all coupling functions are not equally relevant as a measure of the coupling between the solar wind and the magnetosphere.

A comparison of networks fed with different coupling functions demonstrate their different abilities to account for the observed *AE* activity. The function VB_s can be interpreted as the rectified dawn-to-dusk component of the solar-wind electric field. It is generally believed to be one of the most important quantities determining the rate of energy transfer from the solar wind to the magnetosphere. The results in Fig. 5 show, however, that the predictions are improved if VB_s is properly scaled with velocity V and density n. The parameter V^2B_s scaled with the square root of the solar-wind dynamic pressure, $p \sim nV^2$, was the coupling function that gave the most accurate predictions.

In general, the use of individual solar-wind variables is superior to the use of coupling functions. We have still not found a single function of solar-wind parameters that can summarize all relevant information contained in the individual variables. Relevant information on the solar wind is clearly being lost when measured variables are combined into coupling functions.

4.2 Predictions from the solar wind and Dst

The predictions described in Sect. 4.1 are based solely on solar-wind data. To each network we now add an additional input node which is fed with the hourly *Dst* index. The other input nodes are still fed with a sequence of solar-wind data. All networks are trained with data from the same training set as before, and tested on the same test data.

The predictions in this part of the study are thus based on both *Dst* and the solar wind. *Dst* contain information on the magnetospheric state, particularly the ring current, but also indirectly on previous solar-wind conditions. Improvements of the *AE* predictions can be the result of the magnetosphere's *Dst* state modulating the solar wind-*AE* relation. If the modulation is significant, we can expect a marked improvement of the *AE* predictions even when the optimum solar-wind data are available. However, any improvements of the *AE* predictions can also be the result of *Dst* indirectly providing information on the past solar-wind conditions. In this case we expect to find no improvements when the optimum solar-wind data are available, but some improvements when less than the optimum solar-wind data are available.

The results of the network runs are shown in Figs. 4 and 5, where the cases with and without the *Dst* input can be compared (the open and solid symbols, respectively). The influence that *Dst* has on the *AE* predictions obviously depends on the amount of solar-wind information that is fed to the network along with the *Dst* index. The *AE* predictions are not markedly influenced by *Dst* when the network has access to at least 100 min of interplanetary parameters n, V, B_y, and B_z. If less solar-wind data are available (Fig. 4), or if coupling functions are used (Fig. 5), the *Dst* input improves the *AE* predictions. The more information on the solar wind that is lacking, either due to a too short input sequence

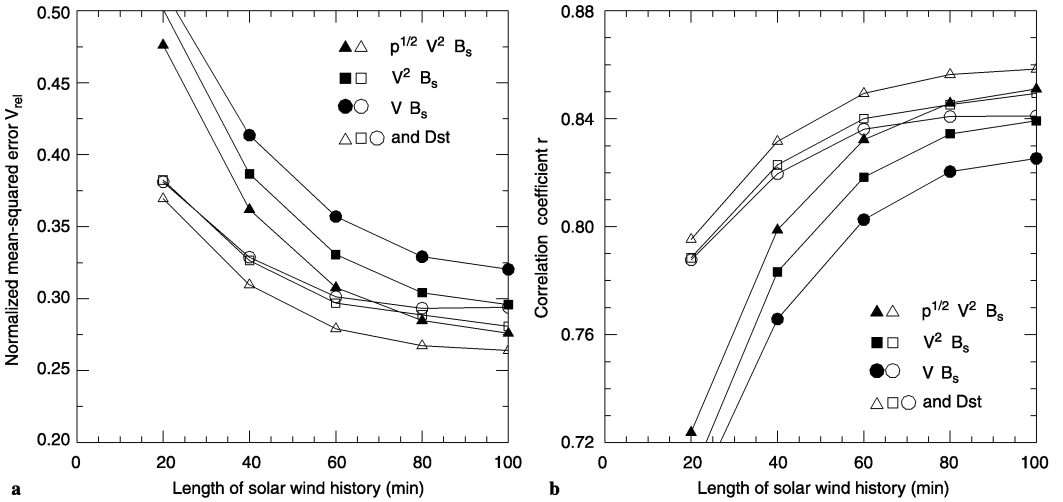

Fig. 5. **a** Normalized mean-squared error and **b** correlation between observed and predicted *AE*, for different coupling functions and temporal lengths of the input sequence. The *solid symbols* mark predictions from the solar wind alone, while the *open symbols* mark predictions from *Dst* and the solar wind. The *AE* predictions are improved by *Dst* for all the coupling functions tested here. The largest improvements occur when much information on the solar-wind conditions are missing. Some information is always lost when measured solar-wind variables are combined into coupling functions

or due to relevant solar-wind quantities being left out, the larger the difference is between the cases with and without *Dst*.

As stated in Sect. 3.2, the linear correlation between *AE* and *Dst* is 0.58 over the 21600 training and test samples. When an optimized ANN use the solar wind alone to predict *AE*, the correlation between prediction and observation is 0.86 for data not used during training. The prediction residual is nearly uncorrelated to *Dst*: $r = 0.05$. A large part of the *AE* variations can thus be explained by a nonlinear mapping from solar-wind data without the need to invoke an explicit $Dst - AE$ relation.

We conclude that the ANNs in the present study do not detect a significant *Dst* influence on the *AE* predictions when an optimal set of solar-wind data are available. Two identical 100-min sequences of solar-wind data would give nearly the same predicted *AE* irrespective of the *Dst* level. This means that the solar wind-*AE* relation, as described by an optimized ANN, is not significantly modified by the magnetosphere's *Dst* state.

A second conclusion is that the *AE* predictions are improved by *Dst* if less than, or other than, the optimum solar-wind data are available. It appears that missing information on the solar wind is partly compensated by the *Dst* index. This finding is of potential importance in operational forecasting where gaps in the stream of real-time solar-wind data are a common occurrence.

5 Conclusions and discussion

We conclude that the auroral electrojet index *AE* can be predicted from solar-wind data alone. However, even though a nonlinear mapping of solar-wind data can explain a large part of the *AE* variations, there could still be an independent influence from the ring current acting to modulate the *AE* index.

From the present study, in which we have compared *AE* predictions from *Dst* and the solar wind with predictions from the solar wind alone, two conclusions are reached: (1) with an optimal set of solar-wind data available, the *AE* predictions are not markedly improved by the *Dst* input, but (2) the *AE* predictions are improved by *Dst* if less than, or other than, the optimum solar-wind data are available to the net. It appears that the solar wind-*AE* relation described by an optimized neural net is not significantly modified by the magnetosphere's *Dst* state, but that missing information on the solar wind can be partly compensated by the *Dst* index. When the solar wind alone is used to predict *AE*, the correlation between predicted and observed *AE* is 0.86, while the correlation between prediction residual and *Dst* is very small, $r = 0.05$. Thus, we have not been able to detect a significant *Dst* modulation of the solar wind-*AE* relation.

The *AE* index is, by its very definition, not sensitive to the *Dst* disturbance field generated by the ring current (Mayaud, 1980). Due to the limited latitude coverage of the geomagnetic observatories that are used to estimate *AE*, the *AE* index depends not only on the strength of the electrojet currents, but also on their location (Kamide and Akasofu, 1983; Akasofu *et al.*, 1983; Baumjohann, 1986). Several studies have shown that the location of the maximum electrojet currents can be observationally related to the strength of the ring current as measured by *Dst* (Feldstein, 1992; Feldstein *et al.*, 1994, 1997; Popov and Feldstein, 1996; Sumaruk *et al.*, 1989), and also to

the general level of geomagnetic activity as measured by *Kp* (Grafe *et al.*, 1983). During time-periods between substorm expansions, a relation between *Dst* and the latitudes of maximum eastward and westward electrojets is found, whereas the latitude of the maximum westward electrojet during a substorm expansion is not similarly related to *Dst* (Feldstein *et al.*, 1997). To the extent that the observed *Dst*-auroral electrojet relations are independent of the solar wind, we would expect the magnetosphere's *Dst* state to modulate the solar wind-*AE* relation. It appears, however, that the neural nets do not detect such a modulation. A conceivable reason could be that the dependence on the magnetospheric *Dst* state is relatively weak compared to the solar-wind dependence. It must also be noted that the maximum westward electrojet during substorm expansions, which is not clearly related to the *Dst* index, is a major contributor to the *AE* index.

With the availability of real-time solar-wind data from the Sun-Earth libration point *L1*, short-term forecasting of geomagnetic activity has now become possible. To produce forecasts that are as accurate and reliable as possible, it is important to make use of all data that contain information on the recent solar-wind conditions, but also all data that contain relevant information on the dynamical state of the magnetosphere. In the present work, we have studied the impact of an hourly ring-current index (*Dst*) that can be accurately predicted, on a high-time resolution auroral-electrojet index (*AE*). The conclusions show that the *Dst* index would not markedly improve the forecasts when an optimal set of solar-wind data are available. However, in an operational setting, occasional data gaps are a common occurrence. This is now handled by temporarily using a network with a shorter solar-wind input sequence than the optimum. There are also intervals of time when not all the solar-wind parameters are available. In such less-than-optimal circumstances, the additional information contained in the hourly *Dst* index can improve the forecasts.

Acknowledgements. National Space Science Data Center and World Data Center C1 for Solar-Terrestrial Physics are gratefully acknowledged for making solar-wind and geomagnetic data available.

Topical Editor K.-H. Glassmeier thanks A. Grafe and A. Klimas for their help in evaluating this paper.

References

Akasofu, S.-I., Relationships between the *AE* and *Dst* indices during geomagnetic storms, *J. Geophys. Res.*, **86,** 4820–4822, 1981.

Akasofu, S.-I., B.-H. Ahn, Y. Kamide, and J. H. Allen, A note on the accuracy of the auroral electrojet indices, *J. Geophys. Res.*, **88,** 5769–5772, 1983.

Arnoldy, R. L., Signature in the interplanetary medium for substorms, *J. Geophys. Res.*, **76,** 5189–5201, 1971.

Bargatze, L. F., D. N. Baker, R. L. McPherron, and E. W. Hones Jr., Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res.*, **90,** 6387–6394, 1985.

Baumjohann, W., Merits and limitations of the use of geomagnetic indices in solar wind-magnetosphere coupling studies, in *Solar Wind-Magnetosphere Coupling*, eds. Y. Kamide and J. A. Slavin, Terra Sci., Tokyo, 3–15, 1986.

Cade III, W. B., J. J. Sojka, and L. Zhu, A correlative comparison of the ring current and auroral electrojets using geomagnetic indices, *J. Geophys. Res.*, **100,** 97–105, 1995.

Clauer, C. R., The technique of linear prediction filters applied to studies of solar wind-magnetosphere coupling, in *Solar Wind-Magnetosphere Coupling*, eds. Y. Kamide and J. A. Slavin, Terra Sci., Tokyo, 39–57, 1986.

Davis, T. N., and R. Parthasarathy, The relationship between polar magnetic activity *Dp* and growth of the geomagnetic ring current, *J. Geophys. Res.*, **72,** 5825, 1967.

Feldstein, Y. I., Modelling of the magnetic field of magnetospheric ring current as a function of interplanetary medium, *Space Sci. Rev.*, **59,** 83–165, 1992.

Feldstein, Y. I., A. E. Levitin, S. A. Golyshev, L. A. Dremuhina, U. B. Veshchezerova, T. E. Valchuk, and A. Grafe, Ring current and auroral electrojets in connection with interplanetary medium parameters during magnetic storms, *Ann. Geophys.*, **12,** 602–611, 1994.

Feldstein, Y. I., A. Grafe, L. I. Gromova, and V. A. Popov, Auroral electrojets during magnetic storms, *J. Geophys. Res.*, **102,** 14223–14235, 1997.

Freeman, J., A. Nagai, P. Reiff, W. Denig, S. Gussenhoven-Shea, M. Heinemann, F. Rich, and M. Hairston, The use of neural networks to predict magnetospheric parameters for input to a magnetospheric forecast model, in *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, eds. J. Joselyn, H. Lundstedt and J. Trolinger, Lund, Sweden, 167–181, 1993.

Gleisner, H., H. Lundstedt, and P. Wintoft, Predicting geomagnetic storms from solar-wind data using time-delay neural networks, *Ann. Geophys.*, **14,** 679–686, 1996.

Gleisner, H. and H. Lundstedt, Response of the auroral electrojets to the solar wind modeled with neural network, *J. Geophys. Res.*, **102,** 14269–14278, 1997.

Gonzalez, W. D., J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, and V. M. Vasyliunas, What is a geomagnetic storm?, *J. Geophys. Res.*, **99,** 5771–5792, 1994.

Grafe, A., Electrojet boundaries and electron injection boundaries, *J. Geomagn. Geoelectr.*, **35,** 1–15, 1983.

Hernandez, J. V., T. Tajima, and W. Horton, Neural net forecasting for geomagnetic activity, *Geophys. Res. Lett.*, **98,** 7673, 1993.

Hertz, J., Generalization in neural networks: theory and practice, in *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, eds. J. Joselyn, H. Lundstedt and J. Trolinger, Lund, Sweden, 55–64, 1993.

Hertz, J., A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, Mass., 1991.

Kamide, Y., and S.-I. Akasofu, Notes on the auroral electrojet indices, *Rev. Geophys.*, **21,** 1647–1656, 1983.

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

Mayaud, P. N., *Derivation, Meaning, and Use of Geomagnetic Indices*, AGU, Washington, USA, 1980.

McPherron, R. L., The role of substorms in the generation of magnetic storms, in *Magnetic Storms*, AGU, Washington, USA, 1997.

McPherron, R. L., D. N. Baker, L. F. Bargatze, C. R. Clauer, and R. E. Holzer, IMF control of geomagnetic activity, *Adv. Space Res.*, **8,** 71, 1988.

Murayama, T., Coupling functions between the solar wind and the D_{st} index, in *Solar Wind-Magnetosphere Coupling*, eds. Y. Kamide and J. A. Slavin, Terra Sci., Tokyo, 119–126, 1986.

Popov, V. A., and Y. I. Feldstein, About a new interpretation of "Harang discontinuity", *Geomagn. Aeron.*, **36,** 43–51, 1996.

Rostoker, G., S.-I. Akasofu, J. Foster, R. A. Greenwald, Y. Kamide, K. Kawasaki, A. T. Y. Lui, R. L. McPherron, and C. T. Russell, Magnetospheric substorms: definition and signatures, *J. Geophys. Res.*, **85,** 1663, 1980.

Rumelhart, D. E., G. Hinton, and R. Williams, Learning representations by back-propagating errors, *Nature*, **323,** 533, 1986.

Sumaruk, P. V., Feldstein, Y. I., and B. A. Belov, The dynamics of magnetospheric activity during an intense magnetic storm, *Geomagn. Aeron.*, **29,** 110–115, 1989.

Tsurutani, B. T., W. D. Gonzalez, A. L. C. Gonzalez, F. Tang, J. K. Arballo, and M. Okada, Interplanetary origin of geomagnetic activity in the declining phase of the solar cycle, *J. Geophys. Res.*, **100,** 21717–21733, 1995.

Vassiliadis, D., The input-state space approach to the prediction of auroral geomagnetic activity from solar wind variables, in *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, eds. J. Joselyn, H. Lundstedt and J. Trolinger, Lund, Sweden, 145–151, 1993.

Vassiliadis, D., A. J. Klimas, D. N. Baker, and D. A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res.*, **100,** 3495–3512, 1995.

Wu, J. -G., and H. Lundstedt, Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks, *Geophys. Res. Lett.*, **23,** 319, 1996.

Wu, J. -G., and H. Lundstedt, Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks, *J. Geophys. Res.*, **102,** 14255–14268, 1997.

IV

To appear in the *Journal of Geophysical Research*

A neural network-based local model for prediction of geomagnetic disturbances

H. Gleisner[1] and H. Lundstedt[2]

[1]Lund Observatory, Box 43, SE-22100 Lund, Sweden
[2]Swedish Institute of Space Physics, Solar-Terrestrial Physics Division, Scheelevägen 17, SE-22370 Lund, Sweden

Abstract. This study shows how locally observed geomagnetic disturbances can be predicted from solar-wind data with artificial neural network (ANN) techniques. After subtraction of a secularly varying base level, the horizontal components X_{Sq} and Y_{Sq} of the quiet-time, daily variations are modeled with radial-basis function networks taking into account seasonal and solar-activity modulations. The remaining horizontal disturbance components ΔX and ΔY are modelled with gated, time-delay networks taking local time and solar-wind data as input. The observed geomagnetic field is not used as input to the networks, which thus constitute explicit nonlinear mappings from the solar wind to the locally observed geomagnetic disturbances.

The ANNs are applied to data from Sodankylä Geomagnetic Observatory located near the peak of the auroral zone. It is shown that 73% of the ΔX variance, but only 34% of the ΔY variance, is predicted from a sequence of solar-wind data. The corresponding results for prediction of all transient variations $X_{Sq} + \Delta X$ and $Y_{Sq} + \Delta Y$ are 74% and 51%, respectively. The local-time modulations of the prediction accuracies are shown, and the qualitative agreement between observed and predicted values are discussed. If driven by real-time data measured upstream in the solar wind, the ANNs here developed can be used for short-term forecasting of the locally observed geomagnetic activity.

1 Introduction

The peak geomagnetic activity occurring in the northern auroral zone can be related to prior solar-wind conditions, as demonstrated by many studies of the AE, AU, and AL indices. The relations can be highly complicated, particularly during substorms when the magnetospheric response to the solar wind contain a dominating nonlinear component [*Bargatze et al.*, 1985; *McPherron et al.*, 1988]. Nonlinear models have been developed to predict the peak activity from solar-wind data [*Vassiliadis*, 1993; *Vassiliadis et al.*, 1995; *Gleisner and Lundstedt*, 1997; *Weigel et al.*, 1999]. These models have provided an improved understanding of the magnetospheric response to the solar wind, and have given us a set of tools for making operational forecasts.

Geomagnetic AE indices are, however, limited to the *time domain*. They do not give any information about the location of the peak activity, or about the simultaneous activity at other locations. For some purposes, this neglect of the *spatial domain* may limit the usefulness of the prediction models, particularly at high latitudes where geomagnetic variations have a considerable degree of spatial structure. The most valuable predictions will be those that describe the geomagnetic variations at particular locations, or the spatial distribution of geomagnetic disturbances. Such predictions would have decisive advantages over predictions of geomagnetic indices.

Several models have been developed that describe the ionospheric convection patterns and geomagnetic activity in the polar caps as a linear, static function of solar-wind parameters [e.g., *Feldstein and Levitin*, 1986; *Papitashvili et al.*, 1994; *Weimer*, 1996]. These models provide spatial information that go beyond geomagnetic indices. However, while valuable for the knowledge they provide, the models are mainly valid in an average sense and/or for intervals dominated by quasi-steady convection effects. Another approach has been to model the geomagnetic disturbance patterns without any reference to the solar-wind driver, thus providing a "nowcast" of the disturbance field [*Rostoker and Nashi*, 1997].

A different type of model for the nonlinear relations between the solar wind and the geomagnetic activity at auroral-zone latitudes, has been reported by *Vassiliadis et al.* [1998] and *Valdivia et al.* [1999]. They presented a nonlinear, dynamic model based on a generalization of the nonlinear filters previously used by *Vassiliadis et al.* [1995] for prediction of the AL index. The model describes how the spatially resolved magnetic disturbances, as measured by a latitudinal chain of magnetometers, develop under the influence of an external solar-wind input.

In the present study, we have developed artificial neural networks (ANNs) for prediction of the geomagnetic disturbances at specific locations. The methods employed are similar to those used by *Gleisner and Lund-*

stedt [1997] for prediction of the AE index. The most important difference is the use of gating networks to synthesize predictions produced by specialized networks into a single predicted value.

The ANNs are here applied to geomagnetic data from Sodankylä Geomagnetic Observatory (SOD) located at 64° geomagnetic latitude. To extract the disturbance field from the observed geomagnetic records, we first define a quiet-time reference field using radial-basis function networks. These networks describe the quiet-time daily variations, X_{Sq} and Y_{Sq}, and their seasonal and solar-activity modulations. The horizontal disturbance components, ΔX and ΔY, are obtained by subtraction of the quiet-time reference field from the observed geomagnetic data. The local disturbance field is then modeled with gated, time-delay networks taking local time (LT) and a sequence of solar-wind data as input (Fig. 1) The observed geomagnetic field is not used as input to the networks, which thus constitute explicit nonlinear mappings from the solar-wind input to a local geomagnetic-activity output.

The data sets are described in section 2. A brief description of ANNs is given in section 3, with more specific details of the present models in section 4. In section 5, some of the results are presented, which are then further discussed in section 6.

2 Data Sets

The solar-wind parameters considered in this study are the wind speed, V, the proton number density, n, and the southward component, B_z, of the interplanetary magnetic field as measured in the GSM coordinate system. These data were obtained from the records of the earth-orbiting spacecraft IMP-8. No attempts were made to correct the data for the time it takes the solar wind to propagate from the IMP-8 position to the magnetopause. The years 1978 to 1985 were scanned for time intervals with long, nearly unbroken sequences of solar-wind data. A total number of 165 intervals was found that had less than 10% missing data, had no data gaps longer than 15 minutes, and that were at least 8 hours long. The magnetic records from SOD contain a few short data gaps that were linearly interpolated. The original geomagnetic data were converted from 1-min to 5-min averages to be consistent with the solar-wind data.

The 165 intervals were divided into a training set (110 intervals; 15550 samples) and a test set (55 intervals; 7550 samples). The test data are not used during training; their only role is to evaluate the network performance. The method used to select the test data was simply to use every third interval for testing. As the original 165 intervals were selected based on the availability of solar-wind data (which is largely governed by the IMP-8 orbit and thus uncorrelated to the geomagnetic activity), the selection is wholly random in relation to the geomagnetic disturbance conditions. For this reason, the test data should constitute a reasonably representative sample. The standard deviations for the horizontal disturbance components are 95 nT and 29 nT, respectively, as calculated over the test set. The corresponding figures based on all data from the years 1978 to 1985 are very similar: 99 nT and 33 nT, respectively.

The data that are used for the daily variation model consist of geomagnetic observations from UT days that are both globally and locally quiet. First, all the truly quiet days from the 5 international quietest days for the months were selected. The criteria for a day to be "truly quiet" are based on the Ap and Kp indices [*Mayaud*, 1980]. Then, days that are globally quiet, but locally disturbed based on the local K index, were removed from the selection. To ensure that all levels of solar activity are represented, two complete solar cycles (1976-1998) were scanned for quiet days, which provided data for more than a thousand days. Any irregular, small-scale variations that still appear in the selected data are not interesting for the regular variations. Hence, we used one hour averages centered on half hours to fit and evaluate the daily variation model. Half of these data were used to train the networks, and half of the data were used to evaluate, or test, the networks.

3 Artificial Neural Networks

In the present study, two types of feed-forward networks are used: the time-delay network (TDN), which is a feed-forward network with the input organized as a time-delay line, and the radial-basis function network (RBN), which is a feed-forward network with gaussian transfer functions at the hidden nodes [*Haykin*, 1999]. TDNs are used to predict the geomagnetic disturbances ΔX and ΔY from local time and solar-wind data, whereas RBNs are used to predict the quiet-time daily variations X_{Sq} and Y_{Sq} from local time, season, and the solar-activity index $F_{10.7}$.

The TDN can be viewed as a layered structure of interconnected nodes, where the strength of the connection between any two nodes is determined by a modifiable weight. The output of a TDN is given by

$$O^{\mu} = \sum_j W_j \tanh(\sum_k w_{jk} I_k^{\mu} + \theta_j) + \theta \qquad (1)$$

where $\{I_k^{\mu}; k = 1, 2, \ldots, m\}$ is an input data vector with m components. The specific input vectors to the TDNs, predicting the geomagnetic disturbances ΔX and ΔY, are more thoroughly described in section 4.3. Superscript μ labels the input-output samples. Index j refers to a hidden node and index k refers to an input node. An additional input node, the bias node, is set to a fixed value and connect to all hidden and output nodes through the weights θ_j and θ.

The RBN can similarly be viewed as a layered structure of interconnected nodes, but instead of the sigmoid-shaped function which is the common choice for TDNs, a gaussian transfer function is used at the hidden nodes. The output of a RBN is given by

$$O^{\mu} = \sum_j W_j \exp(-\sum_k \frac{(I_k^{\mu} - w_{jk})^2}{\sigma^2}) \qquad (2)$$

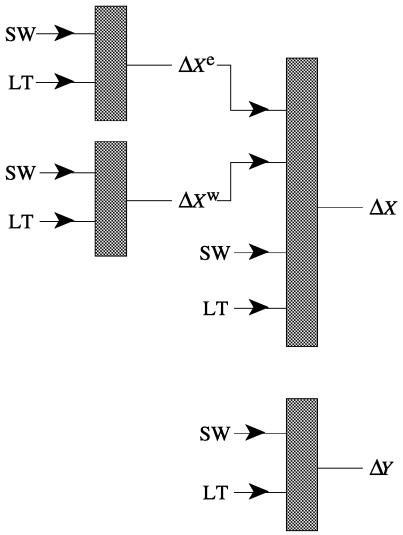

Fig. 1. Block scheme of the ANN model for the geomagnetic disturbance components ΔX and ΔY. Each box indicates a time-delay network that takes local time (LT) and a sequence of solar-wind data (SW) as input. The two leftmost networks predict ΔX^e and ΔX^w, i.e. the positive and negative northward disturbances. The final prediction of ΔX is produced by a gating network that synthesize the predicted ΔX^e and ΔX^w into a single value.

where w_{jk} defines the center of the gaussian function for the j-th hidden node, and σ defines the width of the gaussian shape. The specific input vectors to the RBNs, predicting the regular, daily variations X_{Sq} and Y_{Sq}, are described in section 4.2.

The response properties of an ANN is determined by the set of weights. The network is optimized, or trained, by adjusting the weights until the network produce a response similar to the input-output samples in the training set. The network's ability to produce a "correct" output is monitored by the cost function

$$C(\mathbf{w}) \equiv \frac{1}{2Q_{\text{trn}}} \sum_{\mu=1}^{Q_{\text{trn}}} (O^\mu - T^\mu)^2 \qquad (3)$$

where $\mathbf{w}$ is the set of weights, O^μ is the actual output of the network, T^μ is the "correct" output (or target), and Q_{trn} is the number of samples in the training set.

The TDNs in the present study are optimized using the Levenberg-Marquardt algorithm [*Press et al.*, 1992; *Hagan and Menhaj*, 1994]. At the cost of an increased complexity and memory requirement, this method is much faster than the more commonly used error back-propagation algorithm. The RBNs are optimized with the orthogonal least squares algorithm [*Chen et al.*, 1991; *Wintoft and Lundstedt*, 1999], which is fast but also requires a large amount of computer memory.

Much of the practical use of neural networks relies on their ability to make sensible generalizations. This ability can be defined as the average performance on a randomly chosen data sample. The generalization ability is usually estimated by the network performance on the test set. However, the cost function $C(\mathbf{w})$ measures a network's ability to memorize the training data. In order to achieve an ability to generalize to new data, rather than a perfect fit to the training data, the training procedure need to be constrained. This is done simply by excluding a small part of the training set from the actual training, and use these data, the validation set, to determine when to stop the iteration. In this way the problem of overfitting is avoided, or at least lessened.

4 The Model

4.1 Geomagnetic field variations

For the purposes of this study, the northward and eastward horizontal components of the observed geomagnetic field are described by a base level, $\{X_0, Y_0\}$, on which regular, daily variations, $\{X_{Sq}, Y_{Sq}\}$, and magnetic disturbances, $\{\Delta X, \Delta Y\}$, are superposed.

$$X(t) = X_0(t) + X_{Sq}(t) + \Delta X(t)$$
$$Y(t) = Y_0(t) + Y_{Sq}(t) + \Delta Y(t)$$

The secularly varying base level and the regular, daily variations together describe a quiet-time reference level, and the disturbance field is *defined* as the departure from this reference. It is assumed that the quiet-time reference field provides an adequate description of the geomagnetic variations during periods of weak solar wind-magnetosphere interactions. This assumption of a well defined quiet-time ground state makes it possible to separate contributions to ΔX from eastward and westward ionospheric currents, respectively.

4.2 The quiet-time reference field

The base level $\{X_0, Y_0\}$ is defined by quiet-time annual means, where the means only involve the selection of quiet days described in section 2. The secular variation of the base level is modeled by a piecewise linear fit to these annual means. During the period 1978 to 1985, the average change per year of the base level was -20 nT and +17 nT for X_0 and Y_0, respectively.

The traditional way of modeling the regular, daily variation is by harmonic analysis. The data used as input to the analysis are normally the departures of hourly averages from a daily reference level, often taken to be the midnight value. To account for seasonal and solar-cycle modulations of the daily variation, separate sets of harmonic coefficients are required for each month of the year and for different solar-activity levels.

It was recently pointed out by *Sutcliffe* [1999] that ANN methods have certain advantages over harmonic analysis. We have chosen to model the regular variations with radial-basis function networks taking into account seasonal and solar-activity modulations. Unlike the traditional harmonic analysis, we use the baseline $\{X_0, Y_0\}$

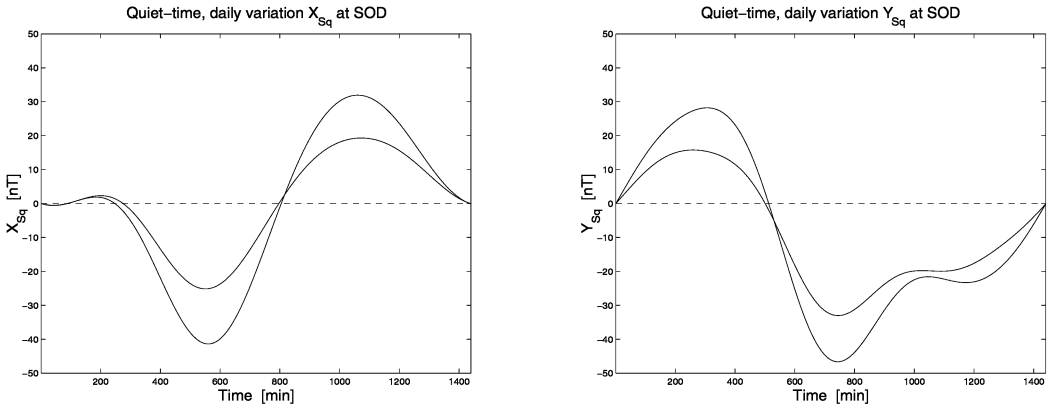

Fig. 2. Computed quiet-time, daily variations X_{Sq} and Y_{Sq} at SOD during summertime, for both low ($F_{10.7} = 75$) and high ($F_{10.7} = 200$) solar-activity levels. The regular variations were computed with radial-basis function networks taking local time, day number, and $F_{10.7}$ as input.

as a reference in the computation of the regular variations. The radial-basis networks were designed to have one hidden layer with 36 nodes and a single, linear output node. Each net take LT, day number, and solar activity as input and produce a component of the daily variation as output. Solar activity is quantified by $F_{10.7}$, the daily values of the solar 10.7 cm radio flux. Day number and LT are each split into two inputs using sine and cosine functions. Each net thus have a total number of five inputs, and the input data vector I^{μ} in Eq. 2 has five elements. Examples of computed quite-time, daily variations $\{X_{Sq}, Y_{Sq}\}$ are shown in Fig. 2.

4.3 The geomagnetic disturbance field

The disturbance field, $\{\Delta X, \Delta Y\}$, is obtained by subtracting the quiet-time reference field from the observed data. At a given location, the disturbance field is assumed to be a function of LT and a finite-length sequence of solar-wind data. This function is highly complicated, being nonlinear and possibly nonstationary. A part of the nonlinearity is a result of the fact that several current systems contribute to the disturbance field. The eastward and westward electrojets, generating positive and negative ΔX disturbances, are known to behave differently and to have partly different physical causes. Another complication is the implementation of a proper LT modulation of the geomagnetic response to the solar wind.

A single ANN constitute a function that is both locally and globally nonlinear. We can, however, expect the overall solar wind-geomagnetic activity relations to be more accurately described by several ANNs that are specialized on separate regimes of the input-output space. The outputs of the specialized networks can be synthesized into a single value by a gating function, which itself can be implemented by a neural network [Ramamurti and Ghosh, 1998; Weigel et al., 1999]. The use of specialized networks and gating functions increases the dynamical range of the neural networks, and should be considered when the input-output space consists of separate regimes with widely different characteristics.

We choose to train separate TDNs on positive and negative ΔX disturbances, and then use a gating network to synthesize the two predictions into a single value. The gating network is an ordinary TDN taking the two specialized ΔX predictions as input, along with LT and solar-wind data (see Fig. 1). The ΔY disturbances are predicted with a single TDN. The basic reason for separating positive and negative disturbances for ΔX, but not for ΔY, is that positive and negative ΔX disturbances for clearly represent two different regimes, namely the effects of the eastward and westward electrojets, respectively. It is not obvious that positive and negative ΔY disturbances represent separate regimes.

One way to deal with the LT modulation of the geomagnetic response would be to develop separate networks for different intervals of LT. We choose instead to use LT as an additional input to each network and to let the network find its own LT modulation as an integrated part of the overall mapping from the solar wind to the geomagnetic response. The overall network configuration is shown in Fig. 1.

The input to the networks consist of an 80-min sequence of the solar-wind dynamic pressure, $\sqrt{nV^2}$, and the coupling function, $V^2 B_s$. All solar-wind parameters have a time resolution of 5 minutes. Each net also use LT, split into two inputs using sine and cosine functions. Each net thus have 16 inputs with $\sqrt{nV^2}$, 16 inputs with $V^2 B_s$, and 2 inputs that determine the local time. The input data vector I^{μ} in Eq. 1 thus has a total number of 34 elements.

5 Results

The observed test data and the corresponding predictions are shown in Fig. 3. The quality of the predictions

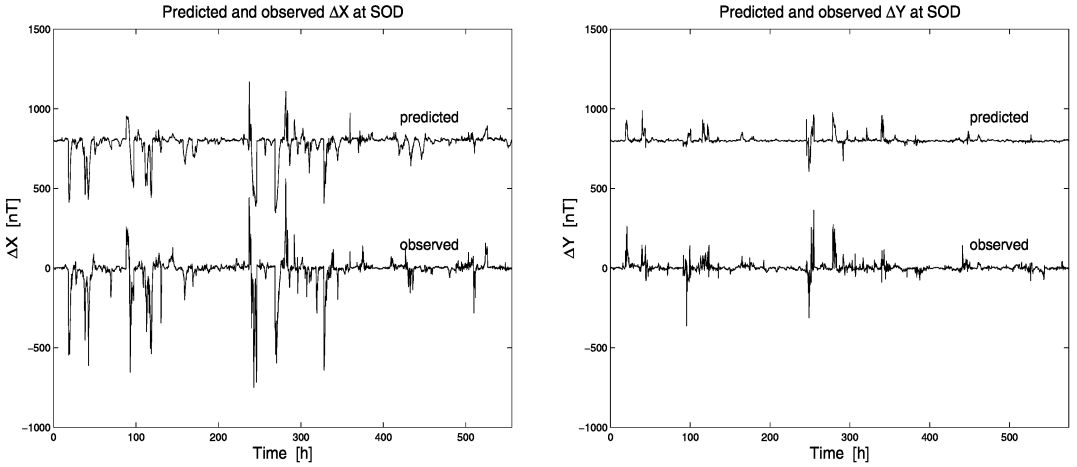

Fig. 3. Predicted and observed horizontal disturbance components ΔX and ΔY at SOD. The predictions were made with gated time-delay neural networks taking local time and a sequence of solar-wind data as input. All test data are shown; 55 separate intervals have been concatenated into a single sequence, and the predicted values have been vertically shifted.

are measured by three statistics: the RMS errors, the average relative variances, ARV, and the linear correlations, ρ, between observed and predicted values. In Table 1, results from predictions of ΔX and ΔY are shown, together with the corresponding results from predictions of all transient variations, $X_{Sq} + \Delta X$ and $Y_{Sq} + \Delta Y$.

During training, the ANNs are optimized with respect to the the sum square difference between network output and observed values. After network optimization, only very small improvements (in the least-squares sense) are possible by a simple linear transformation of the network outputs. We can then interpret ρ^2 as the fraction of the observed variance that is accounted for by the model [*Reiff*, 1983]. The quantity 1-ARV has a similar interpretation: it is a measure of the accuracy of the predictions relative to the accuracy of just using a simple average.

5.1 Accuracy of disturbance predictions

The transient variations of the geomagnetic field consists of a regular part, $\{X_{Sq}, Y_{Sq}\}$, and an irregular part, $\{\Delta X, \Delta Y\}$. These two parts can not be separated unambiguously. Rather, as emphasized in section 4.1, the irregular variations are *defined* as the departure from the quiet-time field model. It is only the total transient variations, $\{X_{Sq}+\Delta X, Y_{Sq}+\Delta Y\}$, that can be directly compared to observations. The irregular variations, or disturbances, can only be compared to a combination of observation and model. However, if the average regular variations are small compared to the irregular variations, a comparison between observed and predicted $\{\Delta X, \Delta Y\}$ still provides relevant information. The same would be true if we had a perfect model for the regular variations. The better our quiet-time field model is, the less strict is the requirement that the regular variations should be small compared to the irregular variations.

The observed and predicted $\{\Delta X, \Delta Y\}$ are compared under the assumption that the regular variations are small and/or that our quiet-time reference field can adequately account for the regular variations. It is found that the RMS prediction error for ΔY is smaller than the error for ΔX by a factor of two: 24 nT and 49 nT, respectively. This is largely a result of the observed ΔY variance being smaller than the ΔX variance. Taking the observed variances into account, predictions of ΔX are considerably more accurate than predictions of ΔY: only 34% of the ΔY variance is predicted from the solar-wind data, while 73% of the ΔX variance is predicted.

The variation of the RMS prediction errors with local time is shown in Fig. 4 (solid lines), together with the variation of the observed standard deviations (dashed lines). For both ΔX and ΔY the RMS errors are smallest around local noon. However, the variance of the observed data is also smallest around noon. In relation to the observed variances, the predictions are in fact most accurate around midnight.

The predicted variations $\{X_{Sq} + \Delta X, Y_{Sq} + \Delta Y\}$ are also compared to observations. From Table 1, we see that the prediction of the eastward component, $Y_{Sq} + \Delta Y$,

Table 1. Quality of geomagnetic-activity predictions at SOD. The upper panel shows the results from predictions of the disturbance field. The lower panel shows the corresponding results from predictions of all transient variations, including both the regular and irregular parts.

	RMSE	ρ	ρ^2	ARV
ΔX	49 nT	0.85	0.73	0.27
ΔY	24 nT	0.58	0.34	0.66
$X_{Sq} + \Delta X$	49 nT	0.86	0.74	0.26
$Y_{Sq} + \Delta Y$	24 nT	0.71	0.51	0.49

can explain 51% of the observed variance. This should be compared to the predictions of ΔY alone that only explain 34% of the observed variance. For the northward component, the corresponding figures are 74% and 73%, respectively, i.e. predictions of $X_{Sq} + \Delta X$ are only marginally better than predictions of ΔX alone. These differences between the northward and the eastward components is related to the fact that at SOD we can expect the average intensity of the regular variations to be small compared to the irregular variations for X, but not for Y. The regular variations, Y_{Sq}, contribute significantly to the total variance of Y, whereas X_{Sq} does not similarly contribute to the total variance of X.

5.2 Qualitative results

All 55 intervals included in the test set are shown in Fig. 3. The dominating features of the disturbance record are reproduced by the predictions, although with underestimated amplitudes and a tendency to be broadened.

Four of these 55 intervals are shown in more detail in Fig. 5. Dotted lines show observed values, and the solid lines show the predictions. The relations between observation and prediction show the same qualitative characteristics as predictions of the *AE* index: the predictions are effectively smoothed versions of the observations. Much of the small-scale, high-frequency variations are not being reproduced, especially in the ΔY records. Some intensifications of ΔX are also missed. This can be seen in intervals #13 and #30: both intervals start with a -200 nT ΔX disturbance that is not predicted. A rather odd feature can be seen in the ΔY record of interval #13. The disturbance at the end of the interval is being predicted, but with the wrong sign.

One source of error is when predicted disturbances have a relatively correct appearance, but are shifted in time compared to the observations. An example of this behaviour can be seen in the ΔX record of interval #35. The steep 150 nT rise at approximately 3.5 hours after the beginning of the interval is being predicted, but the prediction lies ahead of the observation. The fact that no allowance is made for the travel time between IMP-8 and the subsolar magnetopause is a contributing factor to these kind of errors.

Apart from the above limitations, the main features of the disturbance records are actually being predicted. The two substorm-like features in the ΔX record of interval #5 are accurately predicted, although the magnitude of the disturbances are underestimated. The simultaneous disturbances in ΔY are also reproduced. The prediction of the large ΔX depression in interval #13 show similar characteristics. Interval #30 ends with two positive ΔX intensifications: both are predicted, but the largest have only half the observed amplitude.

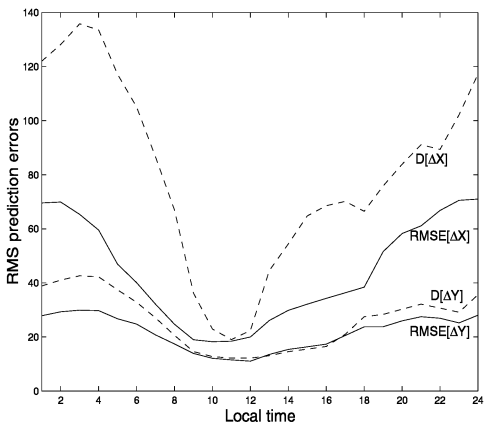

Fig. 4. The RMS prediction errors (solid lines) and the observed standard deviations (dashed lines) for the disturbance field, as a function of local time. The results are based solely on the test data.

5.3 Accuracy of the daily variation model

Although the main topic of this paper is prediction of geomagnetic disturbances, the accuracy of the daily variation model is also of some interest. As described in section 2, the test data consist of hourly averages from more than 500 quiet days. The RMS errors are around 9 nT for X_{Sq} and 6 nT for Y_{Sq}, and the correlations between observed and modeled values are 0.80 for X_{Sq} and 0.90 for Y_{Sq}.

It should be emphasized that the solar-activity index $F_{10.7}$ does not provide the neural networks with enough information to accurately account for the day-to-day variability of the daily variations. A part of the day-to-day variability may have internal ionospheric causes [*Greener and Schlapp*, 1979], and at high latitudes much of the variability depends on the solar-wind conditions and on magnetospheric processes. The role of $F_{10.7}$ is mainly to account for variations of the solar ionizing radiation on a solar-rotational or solar-cycle time scale [*Lean*, 1987], and not to account for the day-to-day variability.

6 Discussion

This study shows how the locally observed geomagnetic variations can be predicted from solar-wind data with ANN techniques. It also shows what accuracies to expect using relatively simple neural networks. Several ways to improve on the networks are not explored in this paper to keep the networks as simple as possible.

We found a large difference between the prediction accuracies of ΔX and ΔY: 73% of the observed ΔX variance is predicted, whereas only 34% of the observed ΔY variance is predicted. The predictability from solar-wind data appears to be higher for ΔX than for ΔY. Several effects may contribute to this. One potentially important factor is the different spatial scales of the source currents. Large-scale ionospheric electrojet currents have a predominant east-west flow direction. These are the currents that generate ΔX. The spatial scales of currents

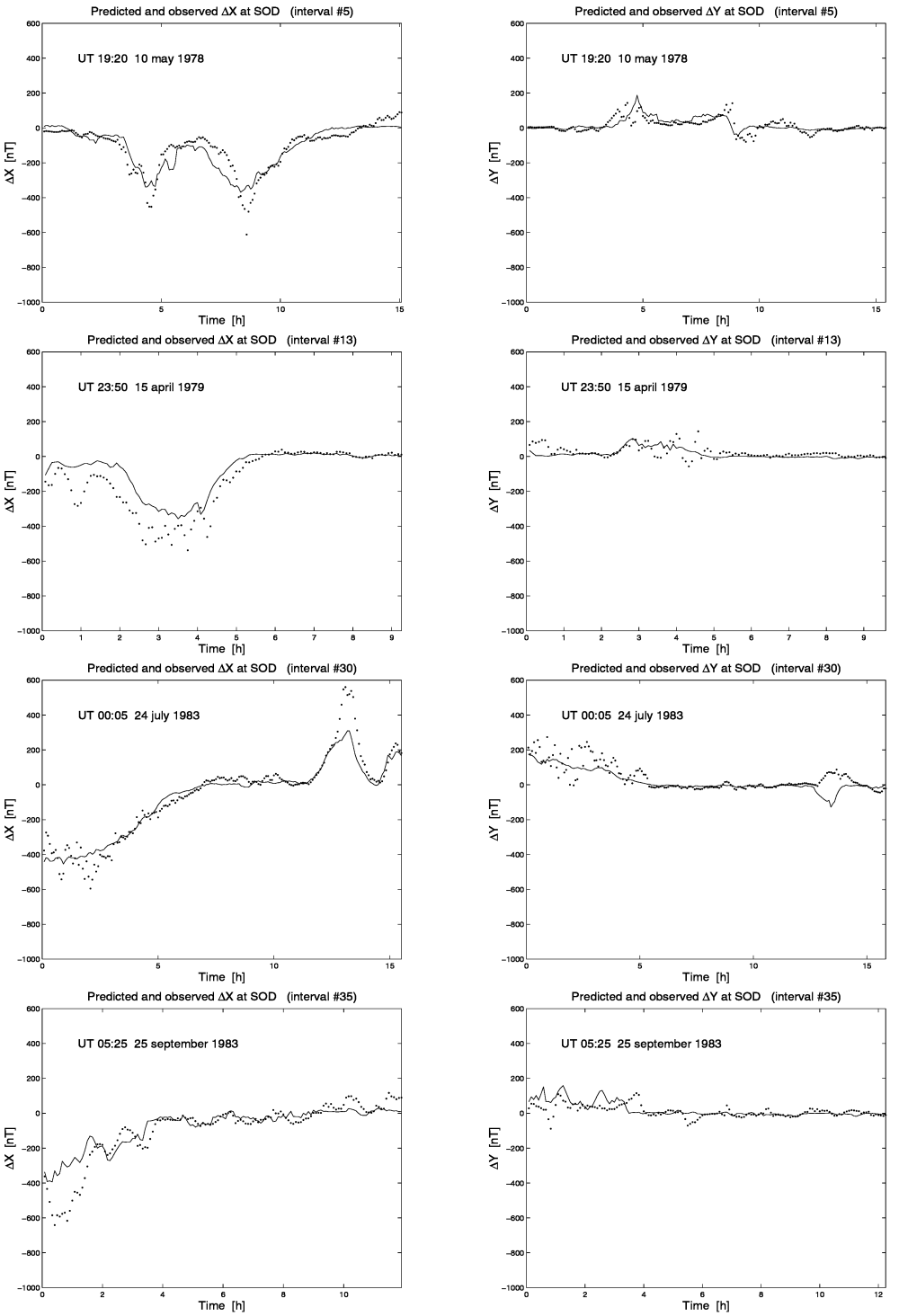

Fig. 5. Predicted and observed disturbance components ΔX and ΔY for four of the 55 intervals in the test set. The dotted lines show the observed values, and the solid lines show the predictions.

flowing in the north-south direction, which are the currents that generate ΔY, tend be smaller, often being associated with current vortices or the Harang discontinuity [*Untiedt and Baumjohann*, 1993].

The predictions presented here are based on a finite-length sequence of solar-wind data, and the neural networks are purely feed-forward, i.e. they have no feedback connections. This has two consequences. Firstly, disturbances caused by current system with an internal time scale longer than the input sequence are not accurately reproduced. For example, the slow variations caused by the ring current decay are not accounted for. While this effect is relatively small for an individual station near the peak of the auroral zone, it can become important for stations at sub-auroral latitudes. Secondly, during prolonged periods of constant solar-wind input the predicted disturbances settle onto a predicted output that vary only with local time. Under these circumstances only the directly driven geomagnetic response can be predicted.

Nevertheless, the prediction accuracies obtained show that ANN techniques can be used to make useful predictions of the geomagnetic disturbance field at particular locations. Further, the characteristics of ANN make them suitable for real-time operation. Using real-time solar-wind data from the sun-earth libration point L_1, forecasts up to an hour ahead of the locally observed geomagnetic disturbances should be possible within the limitations that are described in this paper.

Acknowledgements. The authors would like to thank the staff at Sodankylä Geomagnetic Observatory for making the geomagnetic data available.

References

1. Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones Jr., Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res., 90*, 6387, 1985.
2. Chen, S., C.F.N Cowan, and P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, *IEEE Transactions on Neural Networks, 2*, 302, 1991.
3. Feldstein, Y.I. and A.E. Levitin, Solar wind control of electric fields and currents in the ionosphere, *J. Geomagn. Geoelectr., 38*, 1143, 1986.
4. Gleisner, H. and H. Lundstedt, Response of the auroral electrojets to the solar wind modeled with neural networks, *J. Geophys. Res., 102*, 14269, 1997.
5. Greener, J.G. and D.M. Schlapp, A study of day-to-day variability of S_q over Europe, *J. Atmos. Terr. Phys., 41*, 217, 1979.
6. Hagan, M.T. and M. Menhaj, Training feedforward networks with the Marquardt method, *IEEE Transactions on Neural Networks, 5*, 989, 1994.
7. Haykin, S., *Neural networks - a comprehensive foundation*, 2nd ed., Prentice Hall, Upper Saddle River, New Jersey, 1999.
8. Lean, J., Solar ultraviolet irradiance variations: a review, *J. Geophys. Res., 92*, 839, 1987.
9. Mayaud, P.N., *Derivation, Meaning, and Use of Geomagnetic Indices*, AGU Geophysical Monograph 22, AGU, Washington D.C., USA, 1980.
10. McPherron, R.L., D.N. Baker, L.F. Bargatze, C.R. Clauer, and R.E Holzer, IMF control of geomagnetic activity, *Adv. Space Res., 8*, 71, 1988.
11. Papitashvili, V.O., B.A. Belov, D.S. Faermark, Ya.I. Feldstein, S.A. Golyshev, L.I. Gromova, and A.E. Levitin, Electric potential patterns in the northern and southern polar regions parameterized by the interplanetary magnetic field, *J. Geophys. Res., 99*, 13251, 1994.
12. Press W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in FORTRAN: the art of scientific computing*, 2nd ed., Cambridge University Press, New York, 1994.
13. Ramamurti, V. and J. Ghosh, On the use of localized gating in mixture of expert networks, in proceedings of the *SPIE Conference on Applications of Science and Computational Intelligence*, SPIE Proc. Vol., Orlando, 1998.
14. Reiff, P.H., The use and misuse of statistical analyses, in *Solar-Terrestrial Physics*, eds., R.L. Carrovillano and J.M. Forbes, Reidel Publishing Company, 1983.
15. Rostoker, G. and H.A. Nashi, Technique for space weather nowcasting, in *Proceedings of the Space Weather Workshop*, Saskatchewan, Canada, 1997.
16. Sutcliffe, P.R., The development of a regional geomagnetic daily variation model using neural networks, *Ann. Geophys., 18*, 120, 2000.
17. Untiedt, J., and W. Baumjohann, Studies of polar current systems using the IMS scandinavian magnetometer array, *Space Science Rev., 63*, 215, 1993.
18. Valdivia, J.A., D. Vassiliadis, A.J. Klimas, and A.S. Sharma, Modelling the spatial structure of the high latitude perturbations and the related current systems, *Physics of Plasmas, 6*, 4185, 1999.
19. Vassiliadis, D., The input-state space approach to the prediction of auroral geomagnetic activity from solar wind variables, in *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, eds., J. Joselyn, H. Lundstedt, and J. Trolinger, Lund, Sweden, 1993.
20. Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res., 100*, 3495, 1995.
21. Vassiliadis, D., A.J. Klimas, and J.A. Valdivia, Substorm expansion as seen from the ground: models of the geomagnetic signature, in *Proceedings of the Fourth International Conference on Substorms*, eds., Y. Kamide and S. Kokubun, Tokyo, Terra Scientific, 1998.
22. Weigel, R.S., W. Horton, T. Tajima, and T. Detman, Forecasting auroral electrojet activity from solar wind input with neural networks, *Geophys. Res. Lett., 26*, 1353, 1999.
23. Weimer, D.R., A flexible, IMF dependent model of high-latitude electric potentials having "space weather" applications, *Geophys. Res. Lett., 23*, 2549, 1996.
24. Wintoft, P. and H. Lundstedt, A neural network study of the mapping from solar magnetic fields to the daily average solar wind velocity, *J. Geophys. Res., 104*, 6729, 1999.

V

To appear in the *Journal of Geophysical Research*

Auroral electrojet predictions with dynamic neural networks

H. Gleisner[1] and H. Lundstedt[2]

[1]Lund Observatory, Box 43, SE-22100 Lund, Sweden
[2]Swedish Institute of Space Physics, Solar-Terrestrial Physics Division, Scheelevägen 17, SE-22370 Lund, Sweden

Abstract. Neural networks with internal feedback from the hidden nodes to the input [*Elman*, 1990] are developed for prediction of the auroral electrojet index AE from solar-wind data. Unlike linear and nonlinear ARMA models, such networks are free to develop their own internal representation of the recurrent state variables. Further, they do not incorporate an explicit memory for past states; the memory is implicitly given by the feedback structure of the networks. It is shown that an Elman recurrent network can predict around 70% of the observed AE variance, using only a single sample of solar-wind n, V, and B_z as input. A neural network with identical solar-wind input, but without a feedback mechanism, only predicts around 45% of the AE variance. It is also shown that 4 recurrent state variables are sufficient: the use of more than 4 hidden nodes does not improve the predictions, and with less than 4 hidden nodes the network performance drops. The Elman recurrent networks are compared to time-delay networks taking a sequence of time-lagged solar-wind data as input. To reach comparable prediction accuracies as the recurrent network, a time-delay network needs up to 100 minutes of solar-wind input data.

1 Introduction

Auroral electrojet activity, as measured by the geomagnetic indices AU, AL, and AE, depends on both the external solar-wind forcing and the internal dynamics of the magnetosphere. The important roles played by the solar-wind density, velocity, and magnetic field have been extensively studied during the last 30 years, first using statistical, correlative methods [e.g., *Arnoldy*, 1971; *Murayama and Hakamada*, 1975], and later using linear filters [e.g., *Iyemori et al.*, 1979; *Clauer et al*, 1981]. The studies based on linear filters have demonstrated the fundamental nonlinearity of the geomagnetic response to the solar wind [*Bargatze et al.*, 1985], and have contributed to a better understanding of the time scales involved.

A consequence of the findings by Bargatze et al., is that accurate predictions of geomagnetic activity from solar-wind data require nonlinear methods. *Vassiliadis* [1993], *Price et al.* [1994], and *Vassiliadis et al.* [1995] used a locally linear, but globally nonlinear, filter technique to predict the AE and AL indices. These nonlinear filters were developed from ideas put forward by *Casdagli* [1992] and *Hunter* [1992]. The range of applications for the nonlinear filters have later been extended to the ring current index Dst [*Klimas et al.*, 1998], and to spatial patterns of geomagnetic disturbances at high latitudes [*Valdivia et al.*, 1999].

Neural networks is a related technique that relatively recently has found a wide range of applications. Different type of neural networks have been developed for prediction of geomagnetic indices. Time-delay networks (TDNs), taking a sequence of time-lagged solar-wind data as input, were used by *Lundstedt and Wintoft* [1994] and *Gleisner et al.* [1996] to predict Dst one hour ahead. Later studies by *Wu and Lundstedt* [1996,1997] have shown that Elman recurrent networks (ERNs) can predict the Dst index from a single sample of solar-wind data, i.e. with no explicit reference to time-lagged solar-wind inputs.

In 1993, *Hernandez et al.* published a study on AL predictions using two types of neural networks: a nonlinear ARMA filter and a nonlinear MA filter, the latter being identical to a TDN. This work was extended by *Weigel et al.* [1999] to address certain problems due to clipping of high amplitude variations. In 1997, *Gleisner and Lundstedt* presented a study on AE predictions using TDNs. It was shown that up to 100 minutes of solar-wind data is required, and that the use of raw solar-wind parameters gives more accurate AE predictions than any of the most common coupling functions (i.e. a function that combines several solar-wind parameters into a single quantity). *Takalo and Timonen* [1997] also studied AE predictions using nonlinear ARMA filters fed with a time-lagged sequence of solar-wind data and AE itself.

The two types of recurrent neural networks that have been used to predict geomagnetic indices are thus nonlinear ARMA filters and Elman recurrent networks. Unlike nonlinear ARMA filters, ERNs are free to develop their own representation of the states that are fed back to the input. Further, they do not incorporate an explicit mem-

ory for past states, other than the most recent. The memory is implicitly given by the feedback structure of the network. If fed with a single sample of solar-wind data, any dynamic behaviour of an ERN must be the result solely of the network's feedback structure. The studies by *Wu and Lundstedt* [1996, 1997] show that ERNs have the ability to approximate at least a part of the underlying dynamics of the solar wind-Dst relation. The aim of the present study is to evaluate the abilities of ERNs to model the dynamics of the solar wind-AE relation.

2 Neural Networks

2.1 Nonlinear, dynamic mappings

At the most general level, the geomagnetic activity, O, can be described as a function of a vector, $\mathbf{I}$, of time-lagged solar-wind inputs

$$O_t = F(\mathbf{I}_{t-1}) \tag{1}$$

where

$$\mathbf{I}_{t-1} \equiv \{I_{t-1}, I_{t-2}, \ldots, I_{t-T_I}\} \tag{2}$$

and T_I is the temporal length of $\mathbf{I}$. A more powerful assumption is that the geomagnetic activity depends on both the solar-wind input and prior geomagnetic activity

$$O_t = F(\mathbf{I}_{t-1}, \mathbf{O}_{t-1}) \tag{3}$$

where

$$\mathbf{O}_{t-1} \equiv \{O_{t-1}, O_{t-2}, \ldots, O_{t-T_O}\} \tag{4}$$

and T_O is the temporal length of $\mathbf{O}$. Alternatively, it is a set of internal states, $\mathbf{Y}$, that is fed back to the input

$$O_t = F(\mathbf{I}_{t-1}, \mathbf{Y}_{t-1}) \tag{5}$$

where $\mathbf{Y}_t$ represents the internal state at time t. An interesting difference between the feedback mechanisms of Eqs. 3 and 5, is that in the former equation the time dimension is explicitly represented by a memory for past outputs. In Eq. 5, there is no explicit memory for past states (other than the most recent), and time is only represented implicitly by its effects on processing.

Eqs. 1 to 5 could represent low-dimensional magnetospheric dynamics at the most general level. The simplest possible implementations are the linear moving-average (MA) filter and the auto-regressive moving-average (ARMA) filter. For a magnetosphere with linear response properties, these filters would be perfectly sufficient. However, as the study by *Bargatze et al.* [1985] pointed out, the response properties vary with the level of geomagnetic activity. The magnetospheric response is in fact nonlinear.

One way of introducing nonlinearities is to approximate F locally by linear MA or ARMA filters whose response properties vary between different regions of the input-state space. Such locally linear, but globally nonlinear, filters have been used by *Vassiliadis et al.* [1993, 1995] to study the auroral electrojet response to the solar wind, and the technique has also been described by *Klimas et al.* [1996] in a broader context of nonlinear, magnetospheric dynamics.

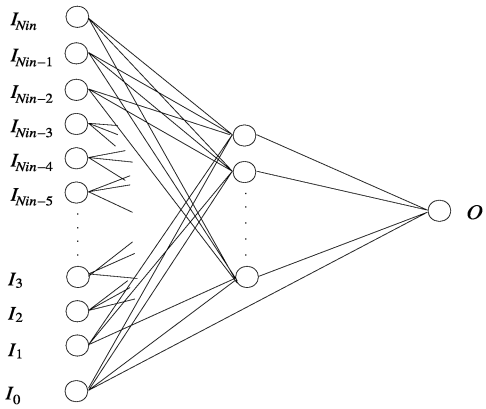

Fig. 1. The time-delay network (TDN) with a sequence of solar-wind data to the left and a single output node to the right. The processing performed at the hidden nodes is nonlinear, while the output node performs a linear, weighted summation.

Neural networks have much in common with the nonlinear filters, but they also differ from them in important respects. For a neural network, the function F is both locally and globally nonlinear. The time-delay network (Figure 1, and further described in section 2.2), with a sequence of time-lagged solar-wind data as input, is essentially a realization of Eq. 1. An Elman recurrent network (Figure 2, and further described in section 2.3), with feedback of internal network activations to a set of context nodes at the input, is a realization of Eq. 5.

2.2 Time-delay networks

Standard feed-forward neural networks are described thoroughly in many text books [e.g., *Haykin*, 1999]. Time-delay networks are simply feed-forward networks that are fed with a temporal sequence of time-lagged external inputs [*Lapedes and Farber*, 1987; *Waibel*, 1989]. It is the organization of the input data that gives the TDN a dynamic behaviour; the mapping itself is static. The processing performed on the input data is given by

$$O^{\mu} = \mathrm{g}_o\Big(\sum_{\mathrm{j}=1}^{\mathrm{N_{hid}}} W_{\mathrm{j}} \mathrm{g}_h\Big(\sum_{\mathrm{k}=0}^{\mathrm{N_{in}}} w_{\mathrm{jk}} I_{\mathrm{k}}^{\mu}\Big) + BI_0\Big) \tag{6}$$

where the input-output samples $\{I_{\mathrm{k}}^{\mu}, O^{\mu}\}$ are labeled by superscript μ. Index j refers to a hidden node and index k refers to an input node. The bias input, I_0, is assigned a fixed value and is connected to all hidden and output nodes in the network through a set of bias weights. In the present study, the activation functions for nodes in the hidden and output layers are

$$g_h(x) = \tanh(x) \; ; \quad g_o(x) = x \tag{7}$$

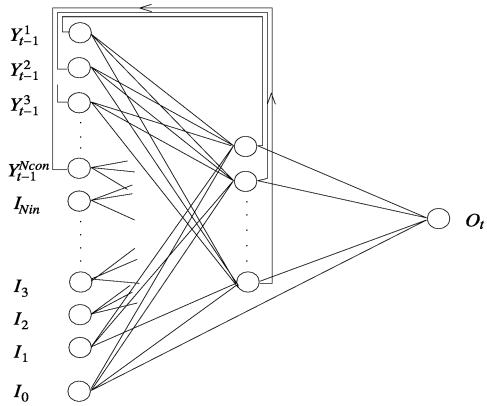

Fig. 2. The Elman recurrent network (ERN) with feed-back connections from the hidden nodes to a set of context nodes at the input. The processing performed at the hidden nodes is nonlinear, while the output node performs a linear, weighted summation.

The incoming signals at a hidden node are thus processed by a nonlinear, differentiable, saturating function, and the processing performed at the output node is simply a linear, weighted summation.

The number of input and output nodes is determined by the set of data to which the network is applied, while the number of hidden nodes, N_{hid}, essentially is a free parameter. The factors that should be considered when determining N_{hid} are briefly discussed in section 4.1.

2.3 Elman recurrent networks

Partially recurrent networks often incorporate the essential features of time-delay networks, but they also include a limited set of fixed feedback connections [*Jordan*, 1989; *Cleeremans et al.*, 1989; *Elman*, 1990]. The recurrent networks devised by *Elman* [1990] feed back the hidden-node activations to a set of context nodes at the input. Unlike the TDN, the dynamical properties of an ERN is a result of the mapping itself being dynamic. The processing performed by an ERN is given by

$$O^\mu_t = \mathrm{g}_o\Big(\sum_{\mathrm{j}=1}^{\mathrm{N_{hid}}} W_{\mathrm{j}} \mathrm{g}_h\Big(\sum_{\mathrm{k}=0}^{\mathrm{N_{in}}} w_{\mathrm{jk}} I^\mu_{\mathrm{k}} + \sum_{\mathrm{c}=1}^{\mathrm{N_{con}}} w_{\mathrm{jc}} Y^{\mathrm{c}}_{t-1}\Big) + BI_0\Big) \quad (8)$$

where index c denotes the context units and t is a time variable. Similar to a TDN, the external input data to the ERN are organized as a temporal sequence of time-lagged data. This sequence is normally relatively short. It often consists of only the current input data, i.e. the ERN is not fed with any time-lagged input data at all. The activation functions are the same as for the TDN described above, nonlinear at the hidden nodes and linear at the output node.

The actual feedback of an ERN is performed by copying the hidden node activations onto the context nodes. No weights are associated with the feedback connections, which are kept fixed during the training process. This is an important property of a recurrent network. As the modifiable connections are purely feed-forward, we can use the same training algorithm as for a TDN.

2.4 Network training

A network's ability to produce a "correct" output is quantified by

$$C(\mathbf{w}) \equiv \frac{1}{2} \sum_{\mu=1}^{\mathrm{Q_{trn}}} (O^\mu - T^\mu)^2 \quad (9)$$

where $\mathbf{w}$ is the set of weights, O^μ is the actual output of the network, T^μ is the "correct" output (or target), and Q_{trn} is the number of input-output samples in a set of training data. Network training is the process of optimizing the cost function, $C(\mathbf{w})$, under certain restrictions. In the present study we have used the error back-propagation algorithm [*Rumelhart et al.*, 1986]. The weights are iteratively updated according to the rule

$$\Delta w_i \leftarrow -\eta \left(\frac{\partial C}{\partial w}\right)_i + \alpha \Delta w_{i-1} \quad (10)$$

$$w_{i+1} \leftarrow w_i + \Delta w_i \quad (11)$$

where w is a single weight and subscript i denote the iteration. Normally, it is only a subset of the Q_{trn} training samples that is used in each iteration, and the actual update is in an approximate gradient direction. The size, Q_{bat}, of this subset is a parameter that, along with η and α, controls the training process. For TDNs, the Q_{bat} samples in the training batch are selected wholly randomly from the training set. For ERNs, intervals of data, rather than individual samples, are randomly selected.

In the present study, the parameters that control the training process have been assigned the values

$$Q_{bat} = \begin{cases} 800 & \text{for TDNs} \\ 3000 - 10000 & \text{for ERNs} \end{cases} \quad (12)$$

$$\eta = \frac{0.012}{Q_{bat}} \quad (13)$$

$$\alpha = 0.90 \quad (14)$$

The varying batch size, Q_{bat}, for the ERNs is a consequence of the fact that the batch consists of a fixed number of data intervals, rather than a fixed number of samples. Further, the training process is more unstable for ERNs than for TDNs, requiring either a smaller learning rate, η, or a larger batch size, Q_{bat}. We have chosen the latter.

Much of the practical use of neural networks relies on their ability to make sensible generalizations. This ability can be defined as the average performance on a randomly chosen data sample. However, the cost function $C(\mathbf{w})$ measures a network's ability to memorize the training data, rather than the ability to generalize to new data. In order to optimize the generalization ability, the training procedure need to be constrained. This is done

by excluding a small part of the training set from the actual training, and use these data to determine when to stop the iteration. In this way the problem of overfitting is avoided, or at least lessened.

3 Data Set

The data were obtained from the Bargatze data set, originally compiled for linear MA filter studies [*Bargatze et al.*, 1985]. These data, or subsets of them, have been used in several studies of magnetospheric dynamics [e.g., *Hernandez et al.*, 1993; *Vassiliadis et al.*, 1995; *Takalo and Timonen*, 1997; *Weigel et al.*, 1999]. A similar data set, largely overlapping in time but with a lower time resolution (5 min instead of 2.5 min), was used in the study by *Gleisner and Lundstedt* [1997].

The data set contains solar-wind data from the Earth-orbiting spacecraft IMP-8, observed between November 1973 and December 1974: velocity, V, proton number density, n, and the interplanetary magnetic field components B_x, B_y, and B_z, given in Geocentric Solar Magnetospheric (GSM) coordinates. The data set also includes the geomagnetic activity index AE at a time resolution of 2.5 minutes. The AE data have been time shifted to account for the solar-wind travel time from the IMP-8 position to the magnetopause.

The data are divided into 34 intervals covering 42216 samples. Each interval contains isolated auroral activity preceded and followed by relatively quiet periods. The intervals are ordered from low to high activity, such that interval 1 is very quiet and interval 34 very disturbed. Every third interval is used as test data to ensure that the networks are tested on a variety of activity levels. The test data are not used during training; their only role is to evaluate the network performance.

4 Studies

4.1 Network setup: input parameters and network size

Neural networks are completely data based in the sense that the model parameters, i.e. the network weights, are determined solely from the data available during the training process. Physics enter the neural-network model through the choice of physical quantities to use as input and output data. The input data should contain a maximum amount of information on the solar wind, but should also exclude all data that are of no relevance for the geomagnetic activity.

Another consideration is the risk for overfitting, which was briefly addressed by *Gleisner and Lundstedt* [1997]. The overfitting problem increases with the number of free parameters in the network, i.e. both with the number of input data and with the number of hidden nodes. The negative effects of overfitting can be reduced by a proper choice of training procedure, but they can not be completely compensated for. An excessive number of weights tends to decrease the performance of any neural network.

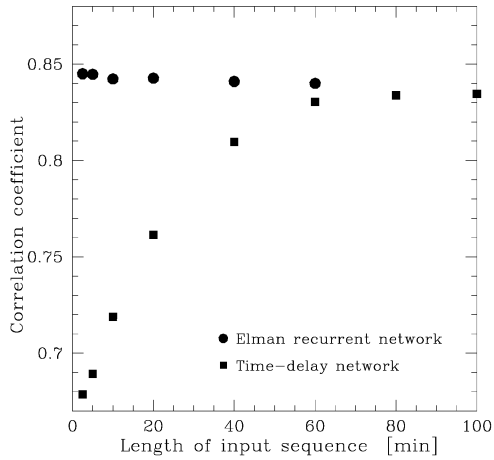

Fig. 3. Correlation between observed and predicted test data as a function of T_I, the temporal length of the solar-wind input sequence. The predictions are based on solar-wind parameters n, V, and B_z. The performance of the Elman recurrent network (●) is nearly independent of T_I, whereas the performance of the time-delay network (■) depends strongly on T_I.

The number of weights in the network, and thus the number of input data, should be as small as possible, while still be large enough to represent the full complexity of the problem. This should favour the use of solar-wind coupling functions. However, previous studies have shown that it is better to use the raw solar-wind parameters n, V, and B_z, than to use any of the most common coupling functions constructed from these parameters [*Gleisner and Lundstedt*, 1997], presumably because a loss of relevant information when combining separate solar-wind parameters into a single quantity.

In the present studies, we have used the solar-wind parameters n, V, and B_z as input to the networks. The networks have 10 hidden nodes and a single output node. For an ERN, the number of context nodes, N_{con}, is identical to the number of hidden nodes, N_{hid}. The basic ERN, which only use a single solar-wind sample as input, have 3 input nodes in addition to the 10 context nodes, one each for n, V, and B_z

$$\mathbf{I}_t = \{n_t, V_t, B_{z,t}\}$$

The basic TDN, which take a 100-min sequence of solar-wind data as input, have a total number of 120 input nodes, 40 each for n, V, and B_z

$$\mathbf{I}_t = \{n_t, \ldots, n_{t-39}, V_t, \ldots, V_{t-39}, B_{z,t}, \ldots, B_{z,t-39}\}$$

In addition to the solar-wind inputs, each network also have a bias input connected to all hidden and output nodes as shown in Figures 1 and 2.

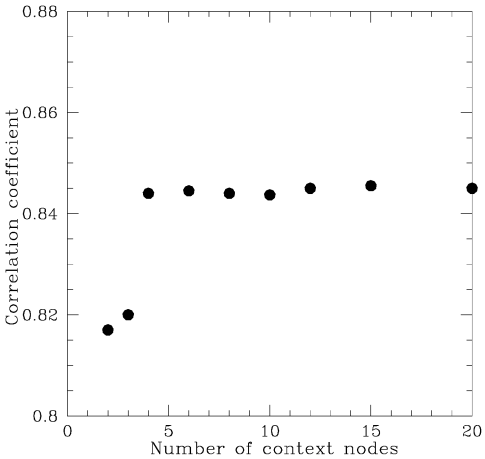

Fig. 4. Correlation between observed and predicted test data as a function of the number of context nodes for an Elman recurrent network. The predictions are based on the solar-wind parameters n, V, and B_z. Four recurrent state variables are sufficient: the prediction accuracy does not improve by adding more than 4 context nodes, and with less than 4 context nodes the network performance drops.

4.2 Length of the solar-wind input sequence

It is the organization of the input data that gives the TDN a dynamic behaviour; the mapping itself is static. Based on a time-lagged vector of past solar-wind inputs, the TDN can display such properties as delayed response and modulation of the response properties by prior inputs. For a TDN, it is crucial that the temporal size, T_I, of the time-delay line is large enough to accommodate all relevant dependences on prior solar-wind inputs. The performance of a TDN is systematically improved with a larger T_I, up to roughly 100 minutes.

Contrary to the TDN, the dynamical properties of the ERN is a result of the mapping itself being dynamic. If the network input consists of only the instantaneous solar-wind data, i.e. if no time-lagged solar-wind data are presented to the ERN, any dynamic behaviour must be the result solely of the network's feedback structure. If, in fact, important aspects of the solar wind-AE dynamics are encoded in the feedback structure, we can expect the performance of the network to be relatively independent of T_I. This property of the ERN differs markedly from the TDN, and can be used as an indicator of the role played by the feedback.

To study this aspect of recurrent networks, a sequence of ERNs were trained with different temporal lengths of the solar-wind input sequence, from 2.5 to 60 minutes. A corresponding sequence of TDNs were trained on identical input data. All networks used the solar-wind parameters n, V, and B_z as input. The results are shown in Figure 3, where the correlation between predicted and observed values over the test set is plotted as a function of T_I.

As expected from previous studies, the performance of the time-delay network is strongly dependent on T_I. A TDN with an input-sequence length, T_I, around 80 to 100 minutes, predicts 70% of the observed AE variance. With $T_I = 2.5$ min, i.e. with only the instantaneous solar-wind data presented to the network, the TDN only predicts 45% of the observed variance.

Unlike the time-delay network, the performance of the recurrent network is nearly independent of the input-sequence length T_I. Using only the instantaneous solar-wind data, i.e. with $T_I = 2.5$ min, the ERN predicts 71% of the AE variance, which is marginally better than the best performing TDN.

We thus conclude that adding time-lagged solar-wind data to the ERN input does not result in an improved prediction accuracy as measured by the correlation between observed and predicted values. The dynamical behaviour of an ERN appears to be entirely due to the feedback structure of the network, and this feedback structure can replace a sequence of time-lagged solar-wind data.

4.3 Number of recurrent state variables

For an ERN that is fed with only the instantaneous solar-wind parameters, i.e. that does not receive any time-lagged external inputs, the number of recurrent state variables becomes an important design parameter of the network. This number provides an upper limit to the dimensionality of the dynamics that can be described by the network. To clarify the role played by the number of state variables, we trained a sequence of ERNs with a varying number of context nodes, and thus a varying number of recurrent state variables. The external solar-wind input consists of a single sample of solar-wind n, V, and B_z. The results are summarized in Figure 4.

With 4 context nodes, the ERNs predict 71% of the AE variance. The prediction accuracy does not improve by adding more context nodes, but with less than 4 context nodes the network performance suddenly drops. We conclude that 4 recurrent state variables are sufficient to give the network a dynamic behaviour that accounts for 71% of the observed AE variance. A larger number of state variables does not contribute significantly to improved predictions.

The drop in network performance from 4 to 3 context nodes is rather abrupt (Figure 4). The performance for 2 or 3 context nodes is, however, relatively high compared to TDNs with short sequences of solar-wind data. Around 66% of the observed variance is still predicted with only 2 context nodes, corresponding to a TDN using around 45 minutes of solar-wind input data.

4.4 Qualitative results

All test data are shown in Figure 5: the TDN results are presented in Figure 5a, and the ERN results in Figure 5b. The 11 test intervals have been concatenated into a single sequence and observed values are vertically shifted.

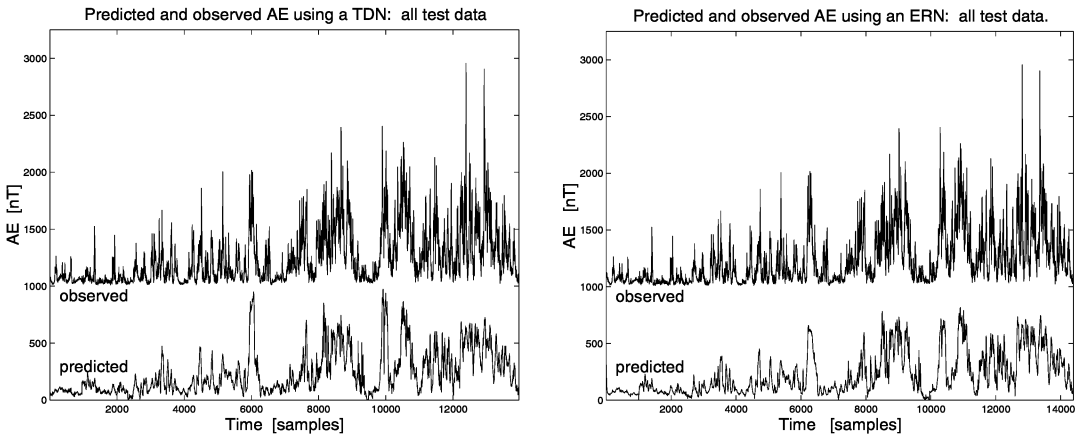

Fig. 5. Observed and predicted AE for (a) the basic time-delay network using 100 minutes of n, V, and B_z as input, and (b) the basic Elman recurrent network using a single sample of n, V, and B_z as input. All test data are shown; 11 separate intervals have been concatenated into a single sequence, and the observed values have been vertically shifted.

Observe that the number of test data are not exactly the same for TDNs and ERNs. This is due to the different temporal lengths, T_I, of the solar-wind input sequences. For the TDN, T_I corresponds to 40 samples, whereas for the ERN it corresponds to a single sample. The first 40 samples are lost from each data interval when using a TDN. When using an ERN, only a single sample is lost from each interval.

Figure 5 gives a rough overview of the correspondence between observation and prediction. The dominating features of the observed AE record are reproduced by the predictions. The amplitudes are, however, underestimated and most narrow, high-frequency features are broadened and some of them are even completely missed. An interesting observation is that even though the statistical performance of the ERN is slightly better than the performance of the TDN, this is not immediately obvious from a visual inspection of Figures 5a and 5b. In fact, it rather tends to give the opposite impression. Several of the dominating features are less underestimated and less broadened by the TDN. The ERN appears to produce smoother predictions than the TDN. This is also confirmed by the variability of the predicted AE: the ERN predictions have a standard deviation of 184 nT, compared to 194 nT for the TDN predictions. As a comparison, the standard deviation of the observed test data is 235 nT.

Despite these obvious differences, the TDN and the ERN predictions are relatively similar, even down to small-scale details. Compare, for example, the interval from $t = 2000$ up to the large peak near $t = 6000$. There is a one-to-one correspondence between the peaks of the two predictions. The same features of the AE records are predicted by both the TDN and the ERN. Where observed features fail to be predicted, as the two narrow peaks around $t = 13000$, they are simultaneously missed by both the TDN and the ERN.

Figure 6 compares, in much more detail, the predicted and observed values for the TDN (left column) and the ERN (right column). Four of the 11 test data intervals are presented: Bargatze intervals 12, 15, 18, and 27. Most comments on the qualitative aspects of Figure 5 are confirmed by the higher resolution of Figure 6. An interesting case where the TDN and the ERN predictions differ is shown in test interval 9 (Bargatze interval 27). In the TDN plot (bottom left of Figure 6), the activity around $t = 1600$ shows three peaks reaching approximately 400 nT and a fourth peak, barely discernible just after the second peak, that reach 250 nT. The corresponding part of the ERN plot (bottom right of Figure 6) shows only two peaks. The first and the second peaks are not resolved and instead form a broad feature in the ERN plot. The third, smaller peak is not found at all in the ERN plot. It probably contributes to a broadening of the base of the first two unresolved peaks. The last of the four peaks is shown in the ERN plot, but it is much broader than in the TDN plot. However, for the most part the TDN and the ERN produce predictions that are very similar. The broad activity around $t = 1200$ in the same two plots, is an example of nearly identical predictions, and many more nearly-identical features can be found in the other plots of Figure 6.

In general, the ERN tends to smooth out high-frequency AE variations more than the TDN. The tendency of the TDN to produce predictions with narrow features that not exactly correspond to observations, is probably a part of the explanation to why the ERN is somewhat better than the TDN from a statistical point of view, while a visual inspection of the two predictions gives the opposite impression.

5 Conclusions

The aim of this study is to evaluate the abilities of ERNs to approximate the dynamics of the solar wind-AE rela-

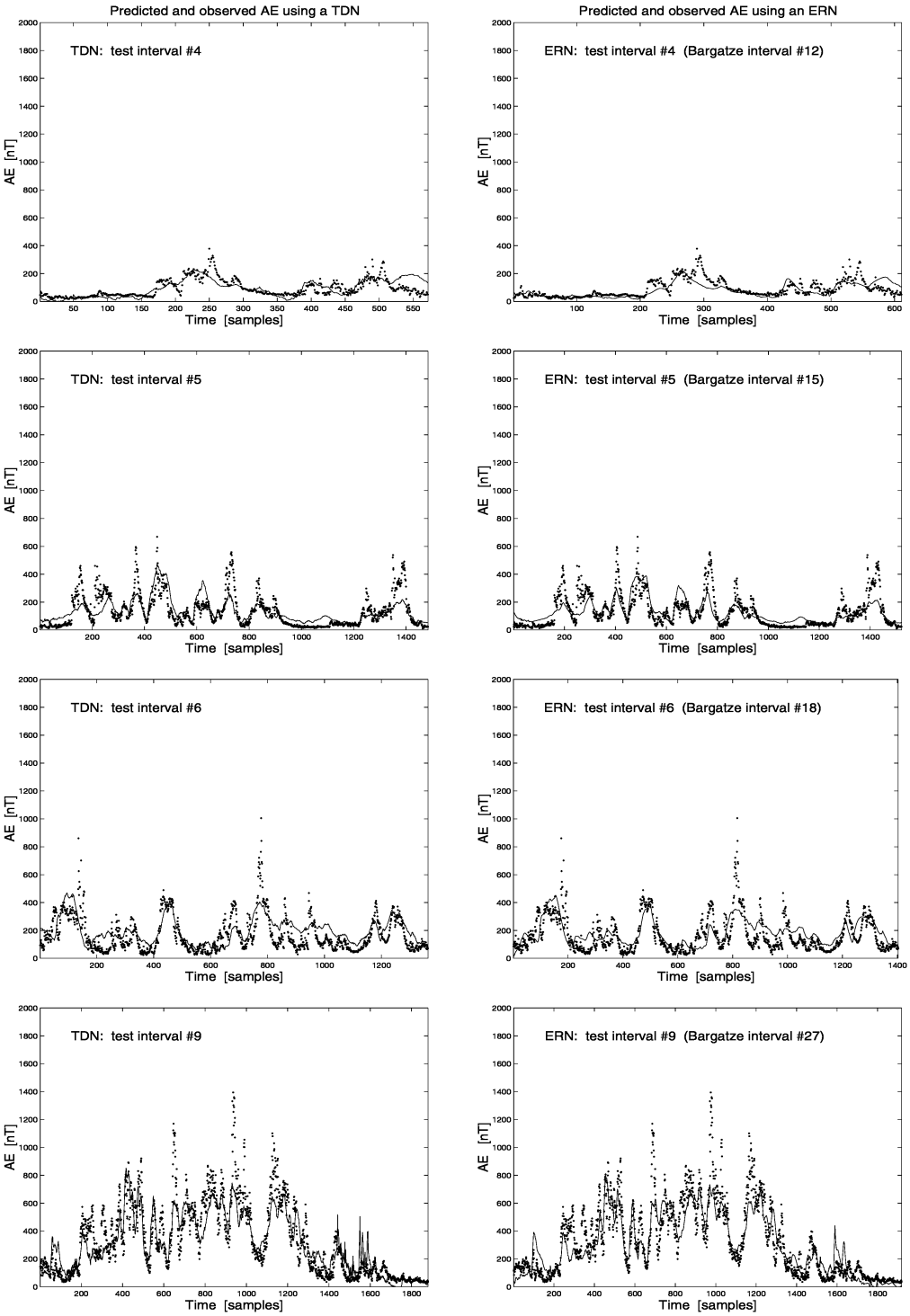

Fig.6. Predicted and observed *AE* for four of the 11 intervals in the test set. The dots show the observed values, and the solid lines show the predictions.

tion. This was done by a comparison with TDNs, whose predictive capabilities are better known. The results show that Elman recurrent networks can predict approximately the same fraction of the observed *AE* variance as the time-delay networks. The predictions are, however, qualitatively somewhat different. The TDN tends to produce predictions with a more irregular appearance, while the ERN tends to smooth out high-frequency irregularities. Observed high-amplitude disturbances tend to be more broadened and underestimated by the ERN than by the TDN. Still, the statistical performance of the ERN is somewhat better than that of the TDN.

The preferred network configurations are vastly different for the two types of network. While a TDN may require up to 100 minutes of solar-wind data, the ERN only requires a single sample of solar-wind parameters. With an equal number of hidden nodes, the number of weights is much larger for the TDN than for the ERN. The ERN can be very simple and still produce relatively accurate predictions: from 1 to 4 input nodes (depending on what solar-wind parameters are used), 4 context nodes, 4 hidden nodes, and a single output node. In fact, with only 2 hidden nodes, an thus 2 context nodes, the ERN produce predictions that are surprisingly accurate.

References

1. Arnoldy, R.L., Signature in the interplanetay medium for substorms, *J. Geophys. Res., 76*, 5189, 1971.
2. Bargatze, L.F., D.N. Baker, R.L. McPherron, and E.W. Hones, Magnetospheric impulse response for many levels of geomagnetic activity, *J. Geophys. Res., 90*, 6387, 1985.
3. Casdagli, M., A dynamical systems approach to modeling input-output systems, in *Nonlinear Modeling and Forecasting*, M. Casdagli and S. Eubank (eds.), pp. 265-281, Addison-Wesley, 1992.
4. Clauer, C.R., R.L. McPherron, C. Searls, and M.G. Kivelson, Solar-wind control of auroral zone geomagnetic activity, *Geophys. Res. Lett., 8*, 915, 1981.
5. Cleeremans, A., D. Servan-Schreiber, and J.L. McClelland, Finite state automata and simple recurrent networks, *Neural Computation, 1*, 372, 1989.
6. Elman, J.L., Finding structure in time, *Cognitive Science, 14*, 179, 1990.
7. Gleisner, H., H. Lundstedt, and P. Wintoft, Predicting geomagnetic storms from solar-wind data using time-delay neural networks, *Ann. Geophys., 14*, 679, 1996.
8. Gleisner, H., and H. Lundstedt, Response of the auroral electrojets to the solar wind modeled with neural networks, *J. Geophys. Res., 102*, 14269, 1997.
9. Haykin, S., *Neural networks - a comprehensive foundation*, 2nd ed., Prentice Hall, Upper Saddle River, New Jersey, 1999.
10. Hernandez, J.V., T. Tajima, and W. Horton, Neural net forecasting for geomagnetic activity, *Geophys. Res. Lett., 98*, 7673, 1993.
11. Hunter, N.F., Applications of nonlinear time series models to driven systems, in *Nonlinear Modeling and Forecasting*, M. Casdagli and S. Eubank (eds.), pp. 467-491, Addison-Wesley, 1992.
12. Iyemori, T., H. Maeda, and T. Kamei, Impulse response of geomagnetic indices to interplanetary magnetic fields, *J. Geomag. Geoelectr., 31*, 1, 1979.
13. Jordan, M.I., Serial order: a parallel, distributed processing approach, in *Advances in Connectionist Theory: Speech*, eds. J.L. Elman and D.E. Rumelhart, Hillsdale, Erlbaum, 1989.
14. Klimas, A.J., D. Vassiliadis, D.N. Baker, and D.A. Roberts, The organized nonlinear dynamics of the magnetosphere, *J. Geophys. Res., 101*, 13089, 1996.
15. Klimas, A.J., D. Vassiliadis, and D.N. Baker, Dst index prediction using data-derived analogues of the magnetospheric dynamics, *J. Geophys. Res., 103*, 20435, 1998.
16. Lapedes, A., and R. Farber, *Nonlinear signal processing using neural networks: prediction and system modelling*, Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, New Mexico, 1987.
17. Lundstedt, H., and P. Wintoft, Prediction of geomagnetic storms from solar-wind data with the use of a neural network, *Ann. Geophys., 12*, 19, 1994.
18. Murayama, T., and K. Hakamada, Effects of solar wind parameters on the development of magnetospheric substorms, *Planet. Space Sci., 23*, 75, 1975.
19. Price, C.P., D. Prichard, and J.E. Bischoff, Nonlinear input/output analysis of the auroral electrojet index, *J. Geophys. Res., 99*, 13227, 1994.
20. Rumelhart, D.E., G. Hinton, and R. Williams, Learning representations by back-propagating errors, *Nature, 323*, 533, 1986.
21. Takalo, J., and J. Timonen, Neural network prediction of AE data, *Geophys. Res. Lett., 24*, 2403, 1997.
22. Valdivia, J.A., D. Vassiliadis, A.J. Klimas, and A.S. Sharma, Modelling the spatial structure of the high latitude perturbations and the related current systems, *Physics of Plasmas, 6*, 4185, 1999.
23. Vassiliadis, D., The input-state space approach to the prediction of auroral geomagnetic activity from solar wind variables, in proceedings of the *International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, J. Joselyn, H. Lundstedt, and J. Trolinger (eds.), pp. 145-151, Lund, Sweden, 1993.
24. Vassiliadis, D., A.J. Klimas, D.N. Baker, and D.A. Roberts, A description of the solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res., 100*, 3495, 1995.
25. Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, Phoneme recognition using time-delay neural networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*, 328, 1989.
26. Weigel, R.S., W. Horton, T. Tajima, and T. Detman, Forecasting auroral electrojet activity from solar wind input with neural networks, *Geophys. Res. Lett., 26*, 1353, 1999.
27. Wu, J.-G., and H. Lundstedt, Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks, *Geophys. Res. Lett., 23*, 319, 1996.
28. Wu, J.-G., and H. Lundstedt, Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks, *J. Geophys. Res., 102*, 14255, 1997.

LUND UNIVERSITY

KFS AB, Lund 2000