PREDICTABILITY OF GEOMAGNETICALLY INDUCED CURRENTS USING NEURAL NETWORKS

A thesis submitted in partial fulfilment of

the requirements for the degree of

MASTER OF SCIENCE

---

of

RHODES UNIVERSITY

---

by

Stefan Lotz

August 2008

Abstract

It is a well documented fact that Geomagnetically Induced Currents (GIC's) poses a significant threat to ground-based electric conductor networks like oil pipelines, railways and powerline networks. A study is undertaken to determine the feasibility of using artificial neural network models to predict GIC occurrence in the Southern African power grid. The magnitude of an induced current at a specific location on the Earth's surface is directly related to the temporal derivative of the geomagnetic field (specifically its horizontal components) at that point. Hence, the focus of the problem is on the prediction of the temporal variations in the horizontal geomagnetic field ($\partial B_x/\partial t$ and $\partial B_y/\partial t$). Artificial neural networks are used to predict $\partial B_x/\partial t$ and $\partial B_y/\partial t$ measured at Hermanus, South Africa ($34.27^\circ\, S$, $19.12^\circ\, E$) with a 30 minute prediction lead time. As input parameters to the neural networks, in-situ solar wind measurements made by the Advanced Composition Explorer (ACE) satellite are used. The results presented here compare well with similar models developed at high-latitude locations (e.g. Sweden, Finland, Canada) where extensive GIC research has been undertaken. It is concluded that it would indeed be feasible to use a neural network model to predict GIC occurrence in the Southern African power grid, provided that GIC measurements, powerline configuration and network parameters are made available.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of this thesis is to study the predictability of geomagnetically induced currents (GIC's) in Southern Africa, utilising a neural network (NN) model.

GIC's are observed when disturbances in the ground level magnetic field around ground-based conductor networks (e.g. power transmission grids, oil pipelines) induce geoelectric fields, which in turn, produce an electric current in the conductor system. The induced current causes spot heating due to half-cycle saturation of the transformers in the system, permanently damaging the transformers and even causing widespread blackouts, e.g. failure of the Hydro-Québec power system in March 1989 (Bolduc, 2002).

Although GIC's are more commonly observed at high-latitudes, due to the larger magnitude and higher frequency of geoeffective magnetic disturbances, they *do* occur at mid-latitudes (e.g. Kappenman, 2005; Koen, 2002). Moreover, rapid variations in the ground magnetic field, i.e. large $\partial B/\partial t$, are often observed at lower latitudes (Pirjola et al., 2005).

Regarding the *prediction* of geomagnetic parameters, one method is to use solar wind data from a spacecraft (e.g. Advanced Composition Explorer (ACE), Solar and Heliospheric Observatory (SOHO), Wind) to infer its geomagnetic effect, utilising the time delay between the spacecraft and the Earth's surface (e.g. Weigel et al., 2003; Lundstedt, 1997).

The processes leading to the rapid variations in the ground magnetic field are complex and cover wide spatial and temporal scales: from the initial activity on the solar surface which varies over a 22-year solar cycle, the 27-day solar rotation and the 24-hour rotation of the Earth around the Sun, to the complex magnetosphere-ionosphere coupling that eventually induces geomagnetic and geoelectric field variations on a 1-second timescale. The relation between solar wind parameters (e.g. density, temperature, velocity) and ground magnetic field measurements is complex and highly non-linear. It is due to the non-linearity and

complexity of the processes driving GIC's that neural networks are considered for the development of a prediction model for GIC occurrence.

Artificial neural networks are in essence a human attempt at reproducing and understanding the way a brain performs various tasks, e.g. pattern recognition. Haykin (1994) defines a neural network as *"a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use"*.

Neural networks have been successfully used to solve a number of prediction problems (McClary et al., 2008) and classification problems (De Silva et al., 2008) in different fields of research.

Given the success of using neural networks in various applications, the objective of this study is to determine whether using neural networks would yield a successful model for predicting GIC occurrence in the Southern African power grid. It must be emphasised that this is a *feasibility* study and the aim is not to predict GIC's but to determine the feasibility of using neural networks to eventually build a working prediction model for GIC's in the Southern African power grid.

Chapter 2 discusses the theoretical background of space weather and geomagnetically induced currents. It also provides an example of the calculation of GIC values for a specific conductor network configuration.

Chapter 3 treats the theoretical aspects of neural networks.

Chapter 4 describes the sources of the data sets used in the development of the neural network model.

The fifth chapter deals with the development of the prediction model, including the choices made regarding neural network configuration and parameters.

Chapter 6 discusses results, their implications in terms of project objectives, and possible future work.

Chapter 2

Space Weather and GIC's

This thesis considers one of the technological effects of space weather, namely *geomagnetically induced currents*. These are electrical currents induced into ground-based conductor systems as a result of rapid changes in the geomagnetic field caused by space weather activity.

The chain of events causing GIC's on the Earth's surface has its origins on the surface of the Sun. Activity on the solar surface is translated to the magnetosphere (and the Earth's surface) through the solar wind by a number of physical processes all contributing to the space weather phenomenon. Hence, in order to describe and understand the formation of GIC's, one must consider space weather and the physical processes driving it.

Space weather is defined by the American National Space Weather Program (NSWP, 1995) as *"... conditions on the Sun and in the space environment that can influence the performance and reliability of space-borne and ground-based technological systems, and can endanger human life or health"*.

Space weather has affected various technological systems since the 1800's, most notably electrical power systems and communication networks (wireless and landline). For example, in 1847 "anomalous currents" were reported in telegraph wires (Lanzerotti, 2001). During the last two centuries the development of large-scale ground-based power, communication and pipeline networks has meant that the adverse effects of space weather have been felt increasingly. In order to mitigate the harmful effects of space weather on technological systems, the relevant physical processes driving these phenomena should be well understood in order to predict them.

The rest of this section describes the aspects of space weather involved in the formation of GIC's.

2.1 The Sun

The processes giving rise to space weather and GIC's are driven by the Sun. The Sun is a medium-sized star (radius $\approx 7\times10^8$ km, mass $\approx 2\times10^{30}$ kg) consisting mostly of Hydrogen (91%) and Helium (8%). It serves as the *driver* of space weather, in the sense that the solar wind which pervades interplanetary space is just the outflow of plasma from the solar atmosphere (corona). The solar wind parameters are governed by solar activity. Apart from the normal outflow of plasma from the Sun, activity on the solar surface and within the corona can cause rapid variations in the various solar wind parameters, e.g. particle densities and velocities, interplanetary magnetic field (IMF) magnitude and orientation.

Violent expulsions of matter from the corona, known as *coronal mass ejections* (CME's), can hurl massive amounts of solar material (10^{12} – 10^{13}kg) in the form of ionised gas away from the Sun into interplanetary space at speeds ranging from under 50 km/s to over 1000 km/s (Gosling, 1997) at an average of around 500 km/s (Prölss, 2004). Various models for CME formation exist (Prölss, 2004). One model described by Prölss (2004) considers a magnetic flux rope forming due to a magnetic field configuration comprised of a poloidal (B_p) and toroidal (B_t) component (Fig. 2.1). This causes a broken loop of magnetic field and frozen-in plasma (from the "Frozen-in flux" condition of magneto-hydrodynamics) to form a so-called magnetic cloud extending from the solar corona outwards as the poloidal magnetic field component increases in strength.

Given a rapid and large enough increase in B_p, the flux rope can extend through the solar wind, eventually reaching the earth. In this case a shockwave can form, accelerating downstream solar wind plasma to move rapidly towards the magnetosphere before interacting with it.

Given the inclination of the magnetic cloud, the B_z component of the interplanetary magnetic field may exhibit a strong southward (negative) orientation, eventually causing an increase of energetic particle influx into the magnetosphere through magnetic reconnection (described in Section 2.3).

Coronal holes and *coronal streamers* are regions of the solar corona known for fast and slow solar wind outflows, respectively. As the solar wind flows radially outward from the rotating Sun, the difference in solar wind speed resulting from coronal holes and streamers cause a compression of the solar wind plasma and the frozen-in magnetic field (see Figure 2.2). Given $B_z < 0$, this compression amplifies the negative orientation of the IMF interacting with the magnetosphere, causing geomagnetic storms through reconnection. Due to the Sun's 27-day rotation around its axis, these corotating regions can periodically

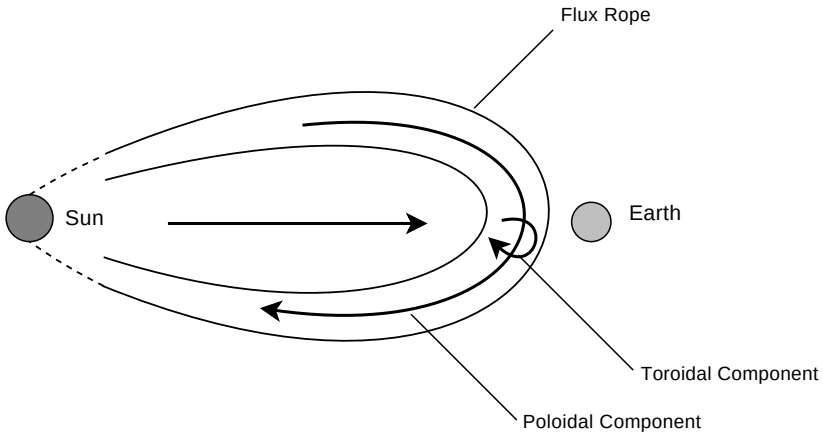

Figure 2.1: Flux rope model of a magnetic cloud. Adapted from Prölss (2004).

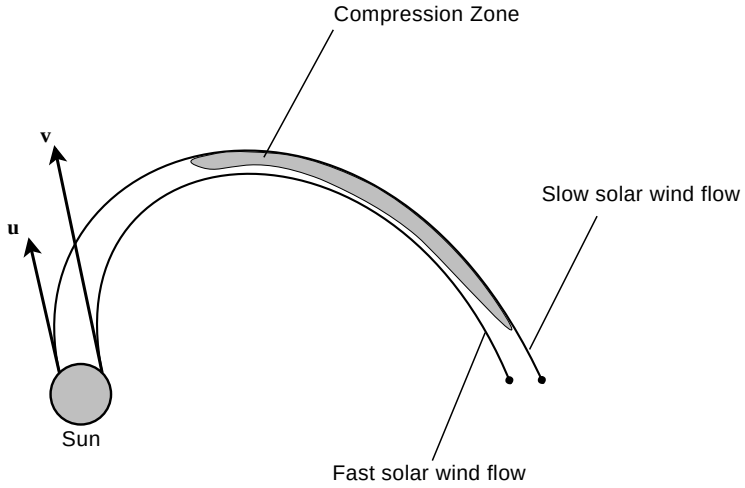

Figure 2.2: Fast ($\mathbf{v}$) and slow ($\mathbf{u}$) outflow regions in the solar corona causes compression of solar wind plasma. Adapted from Prölss (2004).

cause geomagnetic storms.

Geomagnetic storms are disturbances in the Earth's geomagnetic field characterised by the formation of a ring current around the Earth caused by space weather phenomena. The rest of this section highlights the physical processes driving the *interaction* and *reconnection* processes which cause geomagnetic storms.

2.2 Solar Wind

Space weather is closely linked to the state of the solar wind. The solar wind is defined by Hundhausen (1995) as "*a flow of ionised solar plasma and a remnant of the solar mag-*

netic field that pervades interplanetary space". It is driven from the surface of the Sun by the pressure difference between the solar atmosphere (corona) and interplanetary space, transporting charged particles (plasma) and the *frozen-in* magnetic field into interplanetary space.

This plasma serves as the *medium* through which perturbations on the surface of the Sun are transmitted through interplanetary space, moving over and interacting with the Earth's magnetosphere and ionosphere.

2.3 Solar Wind – Magnetosphere Interaction

Eventually, the plasma (and frozen-in IMF) emanating from the solar atmosphere reaches the magnetopause, where it interacts with the magnetosphere. The dipole nature of the geomagnetic field is distorted by the solar wind, compressing it on the day-side of the Earth, and stretching it into what is known as the magnetotail on the night-side.

The dominating process by which energy is transferred from the solar wind to the magnetosphere is through *magnetic reconnection*, whereby solar wind plasma particles are injected into the magnetosphere. The process of magnetic reconnection will be discussed very briefly. A more complete treatment of the subject is given by Hughes (1995).

The frozen-in flux condition states that any plasma particle is attached to a specific magnetic field line and that it cannot move to another field line; or, plasma can only mix *along* flux tubes, not *across* them. When two plasmas originating from different sources interact, they are separated by a current sheet. An example of such a current sheet separating distinct plasmas is of course the magnetopause separating the magnetospheric and interplanetary plasmas.

It is in the vicinity of such a current sheet where the frozen-in flux condition can break down, allowing plasma to mix across field lines. Consider a predominantly southward interplanetary magnetic field interacting with the magnetosphere at the magnetopause. At this location the magnetospheric field direction is predominantly northward, due to Earth's dipole-like magnetic field. Having this configuration of two (approximately) antiparallel magnetic fields separated by a current sheet results in a flow of magnetic flux from opposite sides directed perpendicular to the current sheet. Enforcing the steady state condition ($\partial/\partial t = 0$) upon this process results in a very thin (compared to interplanetary scale lengths) current sheet of scale length $l = 1/\mu_0\sigma u$, separating the two fields. Here, μ_0 is the permeability of free space, σ is the conductivity, and u is the rate of magnetic flux flow.

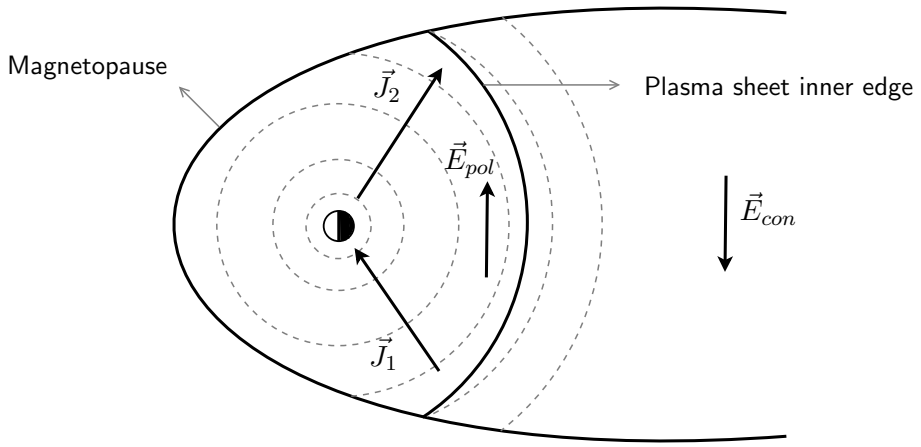

Figure 2.3: Polarisation and convection electric fields, and field-aligned currents near the inner edge of the plasma sheet. Stippled curves denote contours of constant equatorial magnetic field. Adapted from Wolf (1995).

In order for this process to be physically sound, an outflow of plasma must also be allowed. This is achieved by allowing magnetic field lines from the IMF to break and reconnect with field lines from the magnetosphere at the magnetopause. These open-ended field lines are swept back by the solar wind flow toward the night side of the Earth, reconnecting in the magnetotail, returning some of the magnetic field and plasma to the magnetosphere.

This influx of energy from the solar wind into the magnetosphere causes dynamic variations in magnetospheric currents through the *solar wind–magnetosphere* coupling (Pulkkinen, 2003).

2.4 Magnetosphere – Ionosphere Interaction

Changes in magnetospheric dynamics caused by the influx of energy from the solar wind during solar active times are coupled to ionospheric dynamics mostly through the field-aligned currents between the two regions (Wolf, 1995). Deformation of the plasma sheet inner boundary generate polarisation ($\vec{E}_{pol}$) and convection ($\vec{E}_{con}$) electric fields either side of the plasma sheet inner boundary; and field-aligned currents flowing into ($\vec{J}_1$) and out from ($\vec{J}_2$) the ionosphere (Wolf, 1995). Figure 2.3 illustrates the electric fields and currents generated.

2.5 Ground Magnetic Field Effects

Finally, these events have a measurable effect on terrestrial systems (e.g. power lines, magnetometers). Ionospheric currents ($\mathbf{J}$) instantaneously disturb the geomagnetic field

(**B**) due to the *Biot-Savart* law

$$d\mathbf{B} = \frac{\mu_0}{4\pi}\frac{(\mathbf{J}dV)\times\mathbf{r}}{|\mathbf{r}|^3}\ , \tag{2.1}$$

where **r** is the displacement vector from the current element to the point where the magnetic field is being computed, dV the volume element and μ_0 the permeability of free space.

Temporal changes in the geomagnetic field alter the spatial derivatives of the geoelectric field due to Faraday's law

$$\nabla\times\mathbf{E} = -\frac{\partial B}{\partial t}\ . \tag{2.2}$$

This induced electric field generates a sub-surface current in the Earth.

2.6 GIC Calculation

Computing the exact electric field from the magnetic field time-derivative is time consuming and not suitable for a prediction model (Pirjola, 2002). The horizontal components of the geoelectric field generated due to temporal variations in B are rather approximated. One method used by Pirjola (1989), among others, assumes that the electric field generated by ionospheric currents are plane waves propagating vertically downwards towards a uniform Earth.

The relation between the time-derivative of horizontal B and **E** is approximated (Pirjola, 1989) by

$$E_x(t) = \frac{1}{\sqrt{\pi\mu_0\sigma}}\int_0^\infty \frac{g_y(t-u)}{\sqrt{u}}du\ , \quad \text{and} \tag{2.3}$$

$$E_y(t) = -\frac{1}{\sqrt{\pi\mu_0\sigma}}\int_0^\infty \frac{g_x(t-u)}{\sqrt{u}}du\ . \tag{2.4}$$

Here, μ_0 is the permeability of free space; σ is the ground conductivity of the Earth; $g_{x,y} = \partial B_{x,y}/\partial t$ are the temporal derivatives of the horizontal components (x or y) of the ground magnetic field; t and u are times; x and y indicate the north-south and east-west directions, respectively. It is important to note that the electric field depends on all previous values of $\partial B/\partial t$. Of course, in practise measurements up to some finite time in the past are considered, e.g. 29 hours (Pirjola, 1989) or 12 hours (Viljanen and Pirjola, 1989).

Given the time-dependent values of the horizontal ground electric field components, the current induced into a ground-based network of conductors is computed (Koen, 2002) using

$$GIC(t) = aE_x(t) + bE_y(t)\ . \tag{2.5}$$

Here a and b are coefficients, measured in $A\,km/V$, characteristic for each transformer and power line configuration (Viljanen and Pirjola, 1994). Electric field components $E_{x,y}$ are measured in V/km. The configuration of the transformers and power line refers to the line length, the orientation of the network (e.g. directed east-west or north-south), resistance of the line, resistance between neutral and ground, and the number of transformers.

Thus, a and b are characteristic coefficients which depend on the geometry and resistance of the system, and changes whenever the geometry and resistances are changed. This is important to note since it implies that these network coefficients should be measured regularly in order to compute GIC's accurately. Alternatively, coinciding $GIC(t)$ and $\partial B/\partial t$ measurements can be used to compute the network coefficients (Pulkkinen et al., 2007a). Either way, measurements *within* the power network are required to compute GIC's.

This is an important point since measured GIC and power grid configuration data from power companies are critical to the study of GIC's.

2.7 Conductor Network Simulation

In this section the magnitude of induced current in a conductor network is computed using horizontal geomagnetic field (B_x and B_y) values, measured at Hermanus, South Africa ($34.27^\circ\,S$, $19.12^\circ\,E$).

Firstly, equations 2.3 and 2.4 may be rewritten as

$$E_x(t) = \frac{1}{\sqrt{\pi\mu_0\sigma}} \int_{t-M}^{t} \frac{g_y(w)}{\sqrt{t-w}}\, dw\,, \tag{2.6}$$

$$E_y(t) = -\frac{1}{\sqrt{\pi\mu_0\sigma}} \int_{t-M}^{t} \frac{g_x(w)}{\sqrt{t-w}}\, dw \tag{2.7}$$

by substituting the argument to $g_{x,y}$ with w, i.e. $t-u \rightarrow w$, and replacing the infinite interval $\{0,\,\infty\}$ with $\{0,\,M\}$, where M is finite. The notation $g_{x,y}$ refers to the shorthand form of g_x and g_y.

Since the functions $g_{x,y}(w)$ and $\sqrt{t-w}$ are practically realised as discrete sequences of values, the integrals given above are replaced by summations from $t-M$ to t. Furthermore, the substitutions

$$g_x(w) \rightarrow \Delta B_x(w)\,, \tag{2.8}$$

$$g_y(w) \rightarrow \Delta B_y(w) \tag{2.9}$$

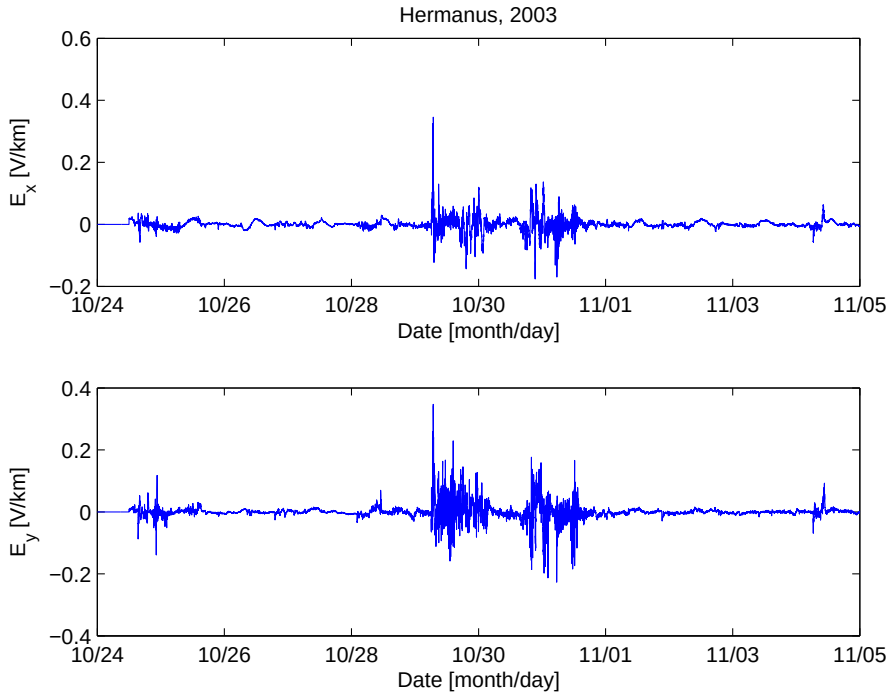

Figure 2.4: E_x and E_y calculated by means of equations 2.11 and 2.12, with B_x and B_y measurements from the same location. The interval depicted here includes the famous "Halloween storm" from 29 October to 1 November 2003.

are made. The temporal variations in B_x and B_y are approximated by applying the difference operator to a N-valued sequence of $B_{x,y}$ measurements:

$$\partial B_{x,y}/\partial t \approx \Delta B_{x,y} = B_{x,y}(i+1) - B_{x,y}(i)\,, \quad \forall\, i \in \{1, \ldots, N-1\}\,. \tag{2.10}$$

Summation by parts, the discrete analogue to integration by parts, is used to simplify the integrals (2.6, 2.7), yielding

$$E_x(t) = \frac{2}{\sqrt{\pi \mu_0 \sigma}} \left[R_t - R_{t-1} - \sqrt{M}\, \Delta B_y(t-M) \right], \text{ with} \tag{2.11}$$

$$R_t = \sum_{n=t-M+1}^{t} \Delta B_y(n) \sqrt{t-n+1} \tag{2.12}$$

for the north-south electric field component at a time t (Pirjola, 1989; Pulkkinen, 2003). Exchanging x and y subscripts above yields the relation for E_y. The horizontal components of the induced electric field are computed from equations 2.11 and 2.12, with ground conductivity set to $\sigma = 10^{-3}\,\Omega^{-1} m^{-1}$ (Koen, 2002) and the summation interval extending 12 hours in the past (i.e. $M = 12$ hours), as suggested by Viljanen and Pirjola (1989).

The time-dependent magnitude of the induced current can now be computed using equa-

tion 2.5, if the network coefficients a and b are known. Network constants for the Grassridge substation, near Port-Elizabeth, South Africa, are used. Their values are $a = -80\,A\,km/V$ and $b = 15\,A\,km/V$, as obtained from Koen (2002). Now equation 2.5 yields a sequence of GIC magnitudes as induced by the temporal variations in the horizontal geomagnetic field, ΔB_x and ΔB_y. Figure 2.4 depicts the E_x and E_y electric field components at Hermanus due to the variations in the horizontal geomagnetic field. The corresponding GIC magnitudes are computed from the $E_{x,y}$ measurements (Fig. 2.4), by means of equation 2.5.

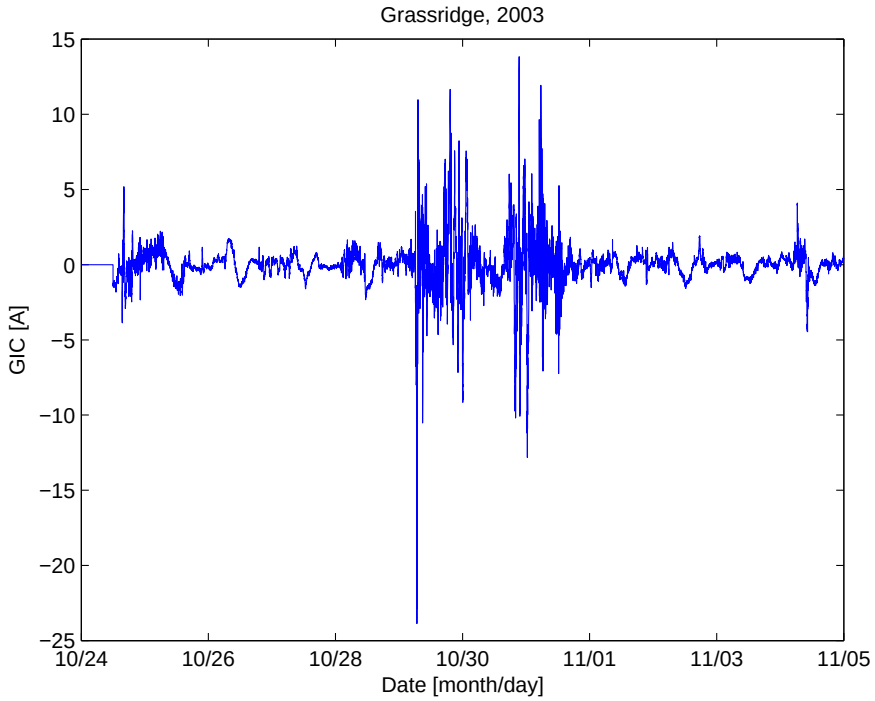

Figure 2.5: GIC magnitude as computed from equation 2.5, utilising the electric field components calculated by means of equations 2.11, 2.12 and depicted in figure 2.4. The network constants a and b used in the GIC equation (2.5) were determined by Koen (2002) for the Grassridge substation.

This section described the procedure for calculating GIC magnitudes from geomagnetic field measurements. This is a critical step in modelling GIC-related behaviour since it provides the link between the theory, i.e. the temporal geomagnetic fluctuations that drive GIC's, and the actual GIC magnitudes which are needed in practise.

In the next chapter neural networks are discussed. Enough information is provided such that the development of a neural network based prediction model (Chapter 5) might be followed.

Chapter 3

Neural Networks

In 1911 *Santiago Ramón y Cajál* introduced the notion of *neurons* being the elemental constituents of the brain (Haykin, 1994, pg. 1) which send and receive information to and from other neurons through synaptic connections.

Using this understanding of its structure, the brain can be viewed as a vast network of interconnected neurons, hence the term *neural network*. It is by means of these neurons and their connecting synapses that the brain has the capability to perform certain tasks or calculations, e.g. pattern recognition (Haykin, 1994, pg. 2).

Artificial neural networks model the way in which the brain learns and executes tasks. Since they are usually simulated through computer software, neural networks are utilised in various fields of research and commerce to solve complicated problems.

Suppose a set of parameters $\mathbf{Y} = \{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_N\}$ are determined by another set of parameters $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_M\}$, through some unknown function $\mathcal{F}$:

$$\mathbf{Y} = \mathcal{F}(\mathbf{X})\,. \tag{3.1}$$

In order to find $\vec{y}_{\{1,\ldots,N\}}$, given a set of values for $\vec{x}_{\{1,\ldots,M\}}$, the relation between $\mathbf{X}$ and $\mathbf{Y}$ should be known, or at least approximated. Presenting a series of values for $\mathbf{X}$ and $\mathbf{Y}$ to a neural network causes it to *learn* the relation between $\mathbf{X}$ and $\mathbf{Y}$, through a process known as neural network training. In neural network terminology, the dependent variables ($\mathbf{Y}$) are known as *output parameters*, whereas the independent variables ($\mathbf{X}$) are known as *input parameters*. Note that the number of input and output parameters need not be equal, i.e. $N \neq M$ is valid. Now, given previously unseen values of the input parameters, an algorithm based on the *trained* network will estimate the output parameters utilising the knowledge gained during training. The aim of neural network modelling is to set up a sound training process and network configuration such that the network will be able to

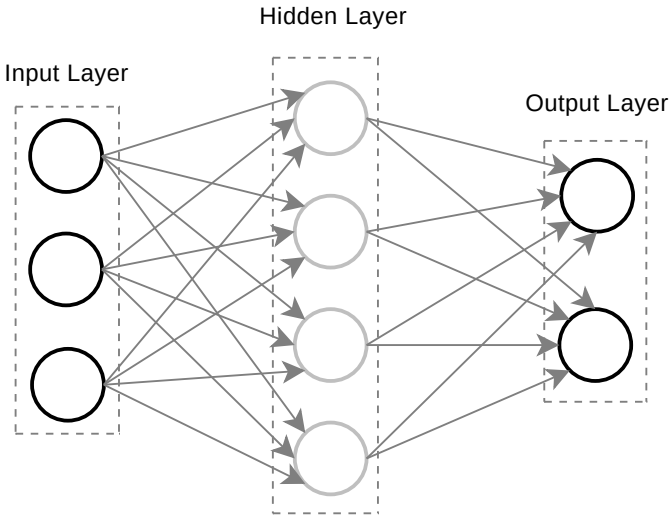

Figure 3.1: A 3-layer feedforward neural network with 3 input nodes, 4 hidden nodes and 2 output nodes.

generalise well, making accurate approximations of the output parameter.

The rest of this chapter is dedicated to the description of neural network structure and training.

3.1 Structure of a Neural Network

Neural networks are structured so as to create a "flow" from the input nodes to the output nodes. In the case of feedforward networks, the flow is strictly unidirectional. Input nodes are the "neurons" of a network corresponding to the input parameters ($\mathbf{X}$ in the example above). The series of input parameter values are represented by an input node; one input node for each input parameter. Correspondingly, the series of known output parameter values are represented by an output node; again, one output node for each output parameter. Multilayer networks include intermediary *hidden nodes*, constituting hidden layer(s) situated between input and output layers (see Fig. 3.1). These layers are hidden in the sense that the interconnections passing information into and out of the hidden nodes are contained within the network with no input from, or output to, external elements. The presence of hidden layers in the neural network enables the network to extract higher-order statistics from the input data. This configuration is especially useful when a large number of input nodes are present (Haykin, 1994, pg. 19).

Recurrent neural networks allow a feedback loop of information, i.e. the flow of information is not unidirectional. Relaxing this restriction allows for an implicit representation of temporal timeshifts in input data – a very useful characteristic under some circumstances.

Elman networks (a type of recurrent network) are discussed further in Section 3.3.

3.1.1 Structure of a Node

The information passing into and out of a node is processed by a number of components which ultimately constitute a node (see Fig. 3.2). The various inputs into a node forms the input signal. Weights are applied to the inputs before adding all the weighted inputs. The sum of the weighted inputs forms the argument to the *activation function*, usually defined by a sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}} \ . \tag{3.2}$$

The activation function imposes a limitation on the amplitude of the output value (Haykin, 1994, pg. 8). The result of the activation function forms the output of the node itself. Other commonly used activation functions are *threshold* and *piecewise linear* functions. A more detailed description of node structure and activation functions can be found in Haykin (1994), Ch. 1.4.

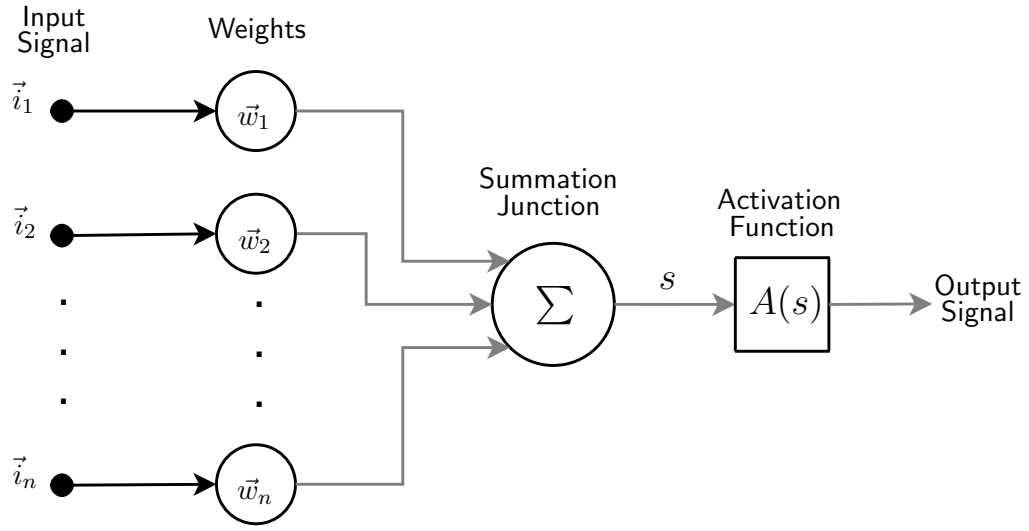

Figure 3.2: Structure of a node. Adapted from Haykin (1994), pg. 10.

3.2 Training and Testing

The process of training a network is critical to the success of the model as it determines the state of the trained network. Training is an iterative process whereby a network is repeatedly

- trained,
- its performance evaluated,

- modifications made to the network and training setup.

The procedure is repeated until satisfactory performance is achieved. In order to realise this procedure, the entire data set (of input and output parameter values) is divided into three subsets. The training algorithm is applied to the *training set* whereby weights are adapted to decrease the error. The class of training process described here is known as *supervised* training. Other training processes include unsupervised and reinforcement training. In a supervised training environment the *validation data set* is used in parallel to the training algorithm to compute the output from the network and to calculate the error between the response and targeted output values. Given the response – normally gauged by the squared-difference between the target and response – weights are adapted accordingly. Finally the *testing set* is used to judge the performance of the trained network. Here the judge is an *external* observer, as opposed to the case of the validation set where the observer is set within the training process itself. Upon completion of the training algorithm, the trained network is processed and its performance evaluated by presenting it with unseen input parameter values and comparing the response to the known output parameter values contained in the testing set. In the case of a prediction model the testing set would be used to compare the predicted and measured values of the output parameter(s). If the observer is satisfied with the performance, the training process concludes – otherwise changes to the network configuration and parameters, training algorithm, or the data sets are made and the process (illustrated in Fig. 3.3) is repeated.

3.2.1 Error-Correction Training

Different types of training *algorithms* or *rules* exist, some aimed at specific training processes (e.g. supervised, unsupervised). The algorithm described here is of the *error-correction* class.

Error-correction training makes use of the validation set described above (Section 3.2) in order to gauge the performance of the network *within* the training algorithm and to adapt the free network parameters (i.e. connection weights) accordingly. Suppose the response of an output node, n, is given by y_n and the target response of that node is t_n. Then the error in the response of node n is

$$e_n = t_n - r_n \,. \tag{3.3}$$

The goal of the error-correction training algorithm is to minimise the error e_n over all the output nodes $n = \{1, \ldots, N\}$, where N denotes the total number of output nodes. Practically, this is achieved by minimising the sum of the squared error

$$\mathcal{E}(i) = \frac{1}{2} \sum_n e_n^2(i) \tag{3.4}$$

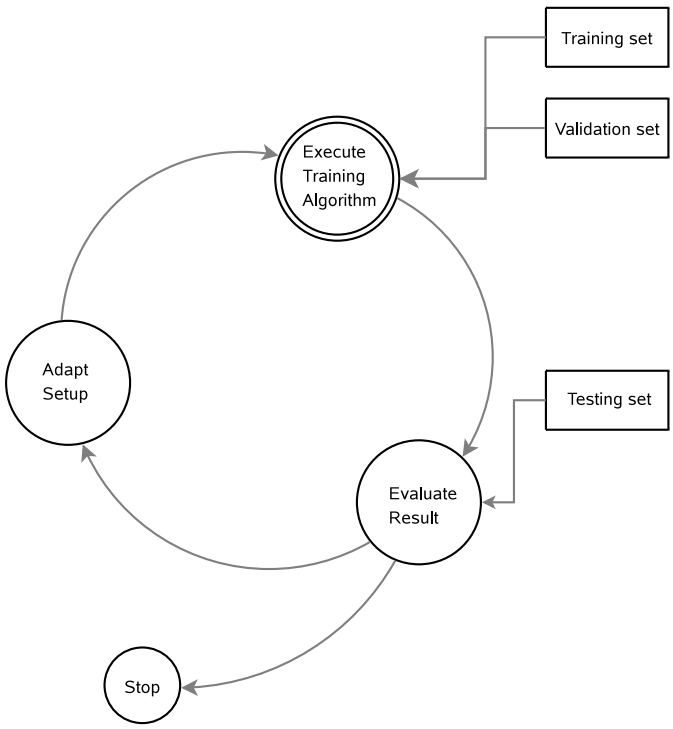

Figure 3.3: Overview of the neural network training procedure. A network is trained, using the *training data set* and validated using the *validation set*, by the training algorithm. The performance of the trained network is evaluated using the *testing set*. Depending on its performance, the process stops or the network and/or training setup is adapted and the process is repeated.

with respect to the connection weights. Here $e_n(i)$ denotes the error of node n at iteration i.

Since the error function is dependent on many weights, it forms a multi-dimensional *error surface*. This surface may contain multiple local minima. The minimisation method attempts to find the *global minimum* of the error-surface by adjusting the connection weights of the network at every iteration of the training algorithm. Minimisation is performed by means of the *gradient descent* method. The specific technique used for applying gradient descent to neural networks is known as *error back-propagation*. The back-propagation method works by efficiently determining the partial derivatives of the approximation function, $\widetilde{\mathcal{F}}(\vec{w} \,|\, \mathbf{X})$, given by the network with respect to the weight vector, given the input parameters $\mathbf{X}$ (Haykin, 1994, pg. 185). The details of this method will not be covered in this study. It is discussed in detail by Haykin (1994) and others.

3.3 Elman Neural Networks

Recurrent networks differ from feedforward networks in that they contain at least one feedback loop (Haykin, 1994, pg. 20). The type of recurrent network considered in this study

are Elman neural networks. Other types of recurrent neural networks include Jordan and Hopfield networks. Elman networks were developed in order to represent time implicitly within the realm of neural networks (Elman, 1990). These networks are structured such that each of the h hidden layers are duplicated, forming h *context* layers. Information is projected from each hidden layer to its corresponding context layer, and back to the hidden layer at every iteration of the training algorithm (see Fig. 3.4).

The feedback loop between a hidden layer and its context layer results in the context layer representing the state of the hidden layer one iteration in the past. This feedback loop is known as a recurrent connection. Since the hidden layers receive input signals not only from the input layer, but also from the context layer (which is just the previous instance of the hidden layer), the training algorithm is guided by the temporal structure of the input and output parameter vectors; and reacts to it through the subsequent modification of connection weights. Hence, Elman networks are said to have "memory".

Refer to Elman (1990) for a detailed description of the development of Elman networks.

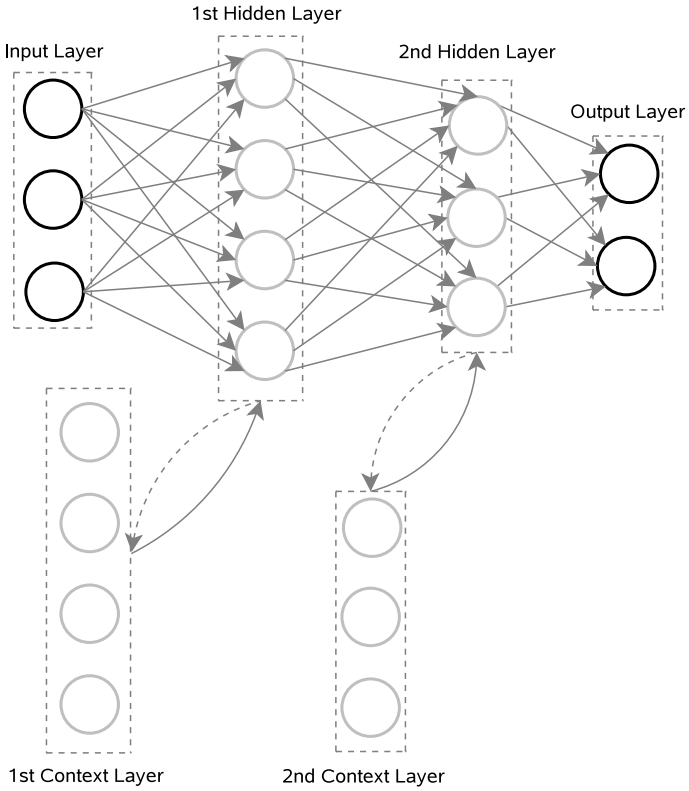

Figure 3.4: An Elman neural network with 3 input nodes, 7 hidden nodes in two hidden layers and 2 output nodes. The hidden layers are duplicated to form *context layers*.

Table 3.1: Representation of the training file format containing 3 descriptive parameters, p inputs and q output parameters. Every parameter comprises one column.

1	r_1	t_1	$input_1^1$	$\cdots$	$input_1^p$	$output_1^1$	$\cdots$	$output_1^q$
2	r_2	t_2	$input_2^1$	$\cdots$	$input_2^p$	$output_2^1$	$\cdots$	$output_2^q$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
N	r_N	t_N	$input_N^1$	$\cdots$	$input_N^p$	$output_N^1$	$\cdots$	$output_N^q$

3.4 Training and Testing Procedure

The training data set is written to electronic file format in the form of a $N \times (p + q + 3)$ matrix (see Table 3.1). The length of the training set is N; p input parameters, q output parameters and 3 *descriptive* parameters are included as columns. The three descriptive parameters represent the index of each row, i.e. $i = 1, 2, \ldots, N$; a set of random numbers $R = r_1, r_2, \ldots, r_N$; and the time at which input and output parameters are measured, $T = t_1, t_2, \ldots, t_N$.

The training set is shuffled by sorting all columns with respect to the random numbers in the second column. All input and output parameter values are normalised to fall within the $(-1, 1)$ interval. A 30% fraction of the training data set is randomly selected and identified as the validation set. Thus 70% and 30% of the original training data set is written to two files, the *training* and *validation* files, respectively.

The testing data set is written to file in the same configuration as the training and validation files, except that the output parameters are not included. After training, the trained network is processed into an executable file. The "testfile" serves as input to the executable, which predicts the output parameters.

The software package used to simulate neural networks is the Stuttgart Neural Network Simulator (SNNS), version 4.2, developed at the University of Stuttgart and the University of Tübingen. The reader is referred to the user manual (Zell et al., 1998) for information on the simulator.

3.5 Summary

The short introduction to neural networks given in this chapter should provide enough clarity in order to follow the rest of this thesis. Many useful texts examining neural networks and their applications are available (Haykin, 1994; Bishop, 1995; Fausett, 1994). Chapter 5 describes the testing and configuration procedure which yields the neural network-based model presented in this study. The sources of input and output data are outlined in Chapter 4. Note that the norm is to not shuffle the training data set when working with

recurrent neural networks. However, we've applied both methods (i.e. using shuffled and ordered training sets) and found that shuffling yields slightly better results.

Chapter 4

Description of Data

As mentioned in the previous chapter, neural networks require a training data set in order to observe and *learn* the complex relations between the input and output parameters of the modelled system. Since the training data set is the only source of problem-specific information available to the network, it is the most critical component in the setup of a neural network model. Regardless of how well the network is optimised for learning, without a data set reflecting the nature of the process to be modelled (and the relation between the input and output parameters), training will not be successful. The goal of the training and testing data sets is to provide a balanced representation of the system under consideration, ideally including every possible scenario, so as not to create a bias towards certain conditions.

Creating a data set which will enable the network training regime to perform well requires careful consideration as to the choice of input and output parameters, length of the time interval spanned by the set and the time resolution of the measurements made. These choices are subject to constraints set by the type of network used, source of the measurements, the goal of the model and the nature of the modelled system.

This chapter describes the measured physical quantities and their sources, which were considered for the development of input and output spaces for the neural network model presented in this study.

4.1 Input Parameter Data Set

The Sun, through the solar wind as medium, is the source of energy influx into the magnetosphere which is responsible for the perturbative geomagnetic effects relating to space weather. Therefore solar wind parameters (e.g. particle densities and velocities measured within the solar wind) are considered as possible inputs to a neural network model tasked

with predicting the geomagnetic effects driving GIC's.

Measurements of solar wind parameters are made by several spacecraft (e.g. SOHO, ACE, Wind) located in near-Earth space. For this study the Advanced Composition Explorer satellite (ACE) is used as the source of solar wind data due to the availability of historic measured data, its (approximately constant) position relative to the Earth and the Sun, and the fact that its measurements has been used by other authors who have conducted similar studies (Wintoft, 2005; Weigel et al., 2003, for example).

ACE and SOHO both orbit the first Lagrangian point in the Earth-Sun system (see section 4.1.1), while Wind follows a complex orbit pattern, moving from one type of orbit to another (Wind, 1999). The orbit radius of ACE is significantly smaller than that of SOHO. This implies that there is less variation in the position of ACE, relative to the Earth, compared to the location of SOHO (see next subsection and Figure 4.2).

4.1.1 Advanced Composition Explorer

According to Stone et al. (1998) ACE is tasked with determining the composition of the solar corona, interplanetary and local interstellar medium (solar wind) and galactic matter through a series of measurements on elemental and isotopic level; and transmitting measurements in real time to NOAA (National Oceanic and Atmospheric Administration) ground stations. The spacecraft, developed by the Johns Hopkins University (Applied Physics Laboratory), was launched on 25 August 1997 from Cape Canaveral Air Station aboard a Delta II rocket.

After a successful launch, ACE was delivered into a highly elliptical transfer orbit, eventually moving to a halo orbit around the first Sun-Earth *libration* point (L1), approximately 1.5×10^6 km from the Earth on the Sun-Earth line. Libration or *Lagrangian* points refer to positions within a three-body system given by the *equilibrium solutions*. It is assumed that the two larger bodies (the Sun and the Earth in this case) move in circular orbits around their centre of mass and that the mass of the third body (ACE) is negligible with respect to the other two. The equilibrium solutions are obtained by setting the velocity and acceleration of the third body to zero, and fixing its position to the ecliptic plane (Vallado, 1997, pg. 126). Thus, placing the third body of a three-body system at one of these Lagrangian points causes it to remain stationary with respect to the other two (larger) bodies. The positions of the five Lagrangian points in the Sun-Earth-satellite system are depicted in Figure 4.1.

This effect yields significant advantages for observational satellite missions like ACE. The advantages of ACE orbiting L1 are three-fold:

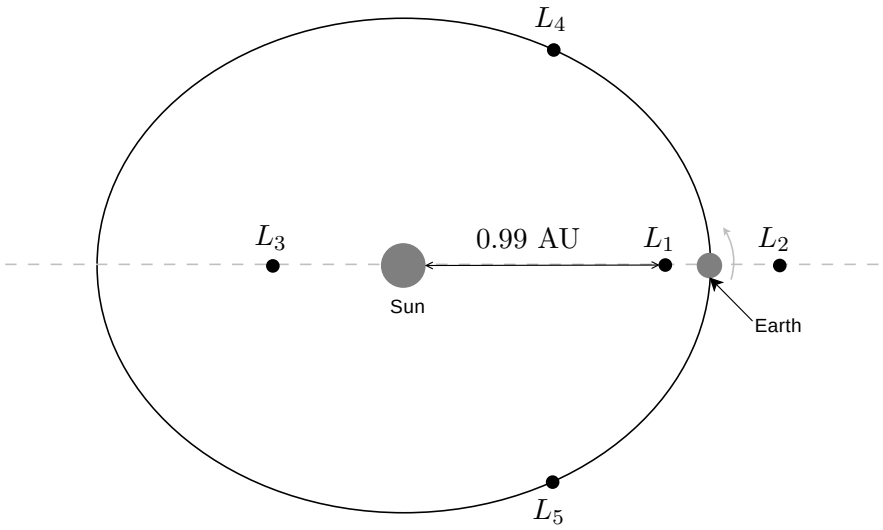

Figure 4.1: The five Lagrangian points in the Sun-Earth-satellite three body system

- It ensures that solar wind measurements are made in the *upstream* solar wind, i.e. *before* it reaches Earth;
- the position of the observer is kept (approximately) constant, avoiding the difficulties related to accounting for a moving observer;
- and fixes its position outside the hostile magnetosphere, where strong magnetic field variations could damage instruments (Stone et al., 1998).

Figure 4.2 depicts the locations of ACE and SOHO orbiting L1, and the magnetosphere bow-shock nose, as measured by ACE during 2001 (GSFC, 2008). Positions are given in Earth radii (R_E) in the GSE x-y plane, with the Sun located beyond ACE and SOHO on the $y = 0$ line. As stated in the third "advantage" mentioned above, the spacecraft orbits L1 at about $230\,R_E$ from Earth, well beyond the magnetosphere. As is evident from Figure 4.2, the extent of the Earth's magnetosphere varies with time. Violent solar coronal activity cause increased particle densities and velocities within the solar wind. The resultant increase in pressure exerted by the solar wind plasma on the magnetosphere causes the magnetosphere to be compressed. Hence the variation in the position of the magnetospheric bow-shock over the one year period illustrated in Figure 4.2. The position of ACE, combined with its (near) real-time transmission ability, provides a natural lead time on solar wind effects. The transmission time from ACE to Earth is negligible (Weigel et al., 2002), while the solar wind takes an average of approximately 55 minutes to travel from L1 to Earth. This is computed by dividing the approximate distance from L1 to Earth ($1.5 \times 10^6\,km$) by the average solar wind speed of $450\,km/s$ (Hundhausen, 1995). During magnetic storms, however, solar wind speeds can reach up to $1000\,km/s$ – resulting in a propagation time of only 25 minutes.

4.1.2 ACE Measurement Instruments

Nine instruments constitute the measurement capacity of ACE. They measure mass, energy, ionic and nuclear charge and magnetic field quantities in the solar wind plasma. Measurements from two of these instruments (SWEPAM and MAG) are utilised in the development of the neural network model presented in this study.

SWEPAM (Solar Wind Electron Proton Alpha Monitor) measures the proton (H, He) and electron energy and charge distributions of the incoming solar wind plasma. From these measurements, quantities like proton number density (N_p) and solar wind velocity are derived. These measurements are critical in the determination of the current state of the solar wind. The MAG instrument consists of two triaxial fluxgate magnetometers measuring the x, y and z geocentric solar magnetospheric (GSM) components of the interplanetary magnetic field.

4.1.3 Sources of ACE Data

All ACE measurement data sets utilised in this study were downloaded in electronic (`.cdf`) format from the Goddard Space Flight Centre (Space Physics Data Facility) web-

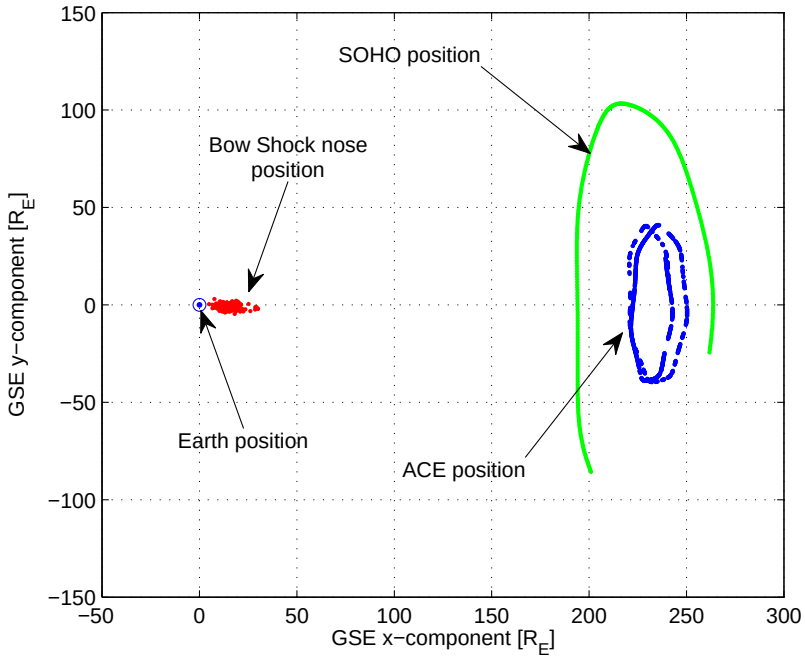

Figure 4.2: Positions of the ACE and SOHO satellites, together with the *bow shock nose* position in the ecliptic plane, as measured by the spacecraft during 2001. Coordinates are GSE (geocentric solar ecliptic), measured in Earth radii (R_E), with Earth at the origin.

site (GSFC, 2008). Near real-time data is available at a 1-minute sampling rate via a website maintained by the NOAA Space Weather Prediction Centre (SWPC, 2003).

4.1.4 Solar Wind Parameters

The actual parameters measured by these two instruments are listed in Table 4.1. These are included in the testing and development of the presented model. The density, temperature and velocity measurements are made by the SWEPAM instrument at 64-second intervals, while the IMF parameters are measured by the magnetometers (MAG) every 16 seconds.

Table 4.1: Input parameters measured by ACE. Density, velocity and temperature measurements are made by the SWEPAM instrument on board ACE at 64 second intervals, while the IMF measurements are made by the MAG instruments at a cadence of 16 seconds.

Symbol	**Description**	**Units**	**Δt**
N_p	Proton number density	$n.cm^{-3}$	$64s$
V_p	Bulk solar wind speed	$km.s^{-1}$	$64s$
T_{pr}	Radial component of proton temperature	$^{\circ}K$	$64s$
N_α	Alpha to proton number density ratio	-	$64s$
$V_{x,\,y,\,z}$	Solar wind velocity components (GSM)	$km.s^{-1}$	$64s$
B^{imf}	Interplanetary magnetic field magnitude	nT	$16s$
$B^{imf}_{x,y,z}$	IMF components (GSM)	nT	$16s$
B^{imf}_{rms}	Root-mean-square of IMF magnitude	nT	$16s$

Since the output parameter data set has a 1-minute resolution (see next section), the MAG and SWEPAM sets are interpolated to 60-second intervals, so that the time resolution of the respective measurements of input and output parameters coincide, as is required by the neural network modelling software.

4.1.5 Input Parameters

A set of 30 quantities is derived from local time and the solar wind parameters listed in Table 4.1. This forms the set of "candidate" input parameters, from which an optimal subset is selected in the next chapter.

Table 4.1 lists the solar wind parameters measured by ACE. These are considered as input parameters to the prediction model. As is the case for the output parameters, functions of the measured solar wind quantities are used as inputs to the model. The *running mean* and *running standard deviations* of the 12 parameters listed in Table 4.1 are considered as possible input parameters.

The running mean (rm_n), sometimes referred to as a "boxcar average", of a sequence of measurements made N times, $\vec{x} = \{x_1, \ldots, x_N\}$, is computed by determining the arithmetic mean of a n-valued subset of these values. The number n is known as the "window-width" and denotes the size of the subset. The running mean of a N-valued sequence $\vec{x}$, written as $rm_n(\vec{x})$, is calculated by applying the definition given below

$$rm_n(\vec{x}) = \frac{x_{t-(n-1)} + x_{t-(n-2)} + \cdots + x_t}{n} \quad \forall\, t \in \{n, \ldots, N\}\,, \tag{4.1}$$

with the window-width n determined independently. Thus, the running mean of some sequence $\vec{x}$ at time t is just the arithmetic mean of the last n values of $\vec{x}$, including x_t.

Analogous to the running mean, the n-minute running standard deviation is defined as

$$rstd_n(x_t) = std\left(x_{t-(n-1)},\, x_{t-(n-2)},\, \ldots,\, x_t\right)\,, \tag{4.2}$$

where $std(\cdot)$ is the standard deviation. In both cases the assignment $n = 10\,min$ is made, so that the 10-minute running mean (rm_{10}) and 10-minute running standard deviation ($rstd_{10}$) of all 12 solar wind parameters are calculated and used as input parameters.

Solar wind electric field is defined by the velocity–magnetic field cross product $\vec{E} = -\vec{V}\times\vec{B}$. The cross-product of $rm_{10}(\vec{V}^{sw})$ and $rm_{10}(\vec{B}^{imf})$ is computed and the magnitude of the result is taken as an input parameter:

$$\mathcal{E}_{sw} = rm_{10}(\vec{V}_{sw}) \times rm_{10}(\vec{B}^{imf})\,. \tag{4.3}$$

According to Pulkkinen et al. (2007b) and references therein, the solar wind dynamic pressure

$$p_{sw}^{dyn} = mNV^2\,, \tag{4.4}$$

with m mass, N density of particles and V denoting solar wind speed, plays an important role in the configuration and dynamics of the magnetosphere. Since the temporal derivative of the geomagnetic field is of interest and noting that the pressure is proportional to NV^2, the temporal derivative yields (Wintoft, 2005)

$$\frac{dp_{sw}^{dyn}}{dt} \propto V^2\frac{dN}{dt} + 2NV\frac{dV}{dt}\,. \tag{4.5}$$

The 10-minute running means of the solar wind bulk velocity V_p and particle density N_p are used to compute the parameter

$$\mathcal{D} = \nu^2\frac{d\eta}{dt} + 2\eta\nu\frac{d\nu}{dt}\,, \tag{4.6}$$

with ν denoting $rm_{10}(V_p)$ and η denoting $rm_{10}(N_p)$. The parameter $\mathcal{D}$ is included in the

Table 4.2: List of all input parameters considered.

$rm_{10}(N_p)$	$rstd_{10}(N_p)$	LTS	$\mathcal{E}_{sw}$
$rm_{10}(V_p)$	$rstd_{10}(V_p)$	LTC	$\mathcal{D}$
$rm_{10}(T_{pr})$	$rstd_{10}(T_{pr})$	DNS	
$rm_{10}(N_\alpha)$	$rstd_{10}(N_\alpha)$	DNC	
$rm_{10}(V_x)$	$rstd_{10}(V_x)$		
$rm_{10}(V_y)$	$rstd_{10}(V_y)$		
$rm_{10}(V_z)$	$rstd_{10}(V_z)$		
$rm_{10}(B^{imf})$	$rstd_{10}(B^{imf})$		
$rm_{10}(B_x^{imf})$	$rstd_{10}(B_x^{imf})$		
$rm_{10}(B_y^{imf})$	$rstd_{10}(B_y^{imf})$		
$rm_{10}(B_z^{imf})$	$rstd_{10}(B_z^{imf})$		
$rm_{10}(B_{rms}^{imf})$	$rstd_{10}(B_{rms}^{imf})$		

set of inputs.

In addition to the resulting 26 parameters, four time-related variables are considered as possible input parameters. Since all the measured quantities mentioned above are ultimately driven by solar processes, it is important to take the diurnal and seasonal effects into account. These are realised as the sine and cosine of local time (LTS, LTC) and day number of the year (DNS, DNC). The time-derived parameters are defined below, with $mins$ denoting the minute of the day, and dn the day number of the year.

$$LTS = \sin\left(\frac{2\pi \cdot mins}{1440}\right) \tag{4.7}$$

$$LTC = \cos\left(\frac{2\pi \cdot mins}{1440}\right) \tag{4.8}$$

$$DNS = \sin\left(\frac{2\pi \cdot dn}{365.25}\right) \tag{4.9}$$

$$DNC = \cos\left(\frac{2\pi \cdot dn}{365.25}\right) \tag{4.10}$$

4.2 Output Parameter Data Set

Chapter 2 described the various processes in the solar-terrestrial environment leading to GIC's. From equations 2.2 – 2.5 it is clear that the temporal derivatives of the horizontal ground magnetic field components, $\partial B_x/\partial t$ and $\partial B_y/\partial t$, drive GIC's. In order to approximate their temporal derivatives, the difference operator $\Delta(\cdot)$ is applied to both B_x and

B_y:

$$\Delta B_x \approx \partial B_x/\partial t\,, \tag{4.11}$$
$$\Delta B_y \approx \partial B_y/\partial t\,. \tag{4.12}$$

The difference operator is defined by

$$\Delta\alpha(i) = \alpha(i+1) - \alpha(i)\,, \quad i = \{1,\,\ldots\,,\,N-1\} \tag{4.13}$$

for some N-valued sequence, α. Therefore, ΔB_x and ΔB_y are defined by

$$\Delta B_x = B_x(i+1) - B_x(i)\,, \tag{4.14}$$
$$\Delta B_y = B_y(i+1) - B_y(i) \quad \forall\, i = \{1,\,\ldots\,,\,N-1\} \tag{4.15}$$

for N-valued B_x and B_y vectors.

4.2.1 Output Parameters

It was established in Chapter 2 and Section 4.2 that the temporal variation in horizontal geomagnetic field magnitudes drive GIC's. Therefore, ΔB_x and ΔB_y (defined by equations 4.14 and 4.15) are considered as output parameters of the GIC occurrence prediction model.

Weigel et al. (2002) studied the solar wind – magnetosphere coupling and predicted the 30-minute mean of $|\Delta B_x|$. Wintoft (2005) found that taking a temporal average of the time-difference B_x component results in a loss of information and that the root-mean-square (RMS) of $\Delta B_{x,y}$ conserves more of the signal than temporal averaging.

For this study a RMS-derived function, the *running root mean square* (RRMS) defined by equation 4.16, is used. Here, the term "*running*" is used in the same sense as with a *running mean* (Section 4.1.5, equation 4.1). It is used in this way: suppose a quantity ($\vec{x}$) is measured at 1-minute intervals over a N-minute period, yielding an array

$$\vec{x}_t = \{x_1,\,x_2,\,\ldots\,,\,x_N\}\,.$$

Now the n-minute running root-mean-square of $\vec{x}$ at time t is defined by

$$RRMS_n(x_t) = RMS\left[\,x_{t-(n-1)},\,x_{t-(n-2)},\,\ldots\,,\,x_t\,\right]\,, \tag{4.16}$$

Table 4.3: Variances (σ^2) of $\Delta B_{x,y}$, $\psi_{x,y}$ and $\omega_{x,y}$ are compared. The definitions of $\psi_{x,y}$ and $\omega_{x,y}$ are given by equations 4.18 and 4.19. The percentages α and β are defined by equations 4.20 and 4.21. They denote the fraction of variance preserved by $\psi_{x,y}$ and $\omega_{x,y}$, respectively.

i	$\sigma^2(\Delta B_i)$	$\sigma^2(\psi_i)$	$\sigma^2(\omega_i)$	$\alpha\%$	$\beta\%$
x	0.3401	0.0999	0.2195	29.36	64.55
y	0.2131	0.0850	0.1150	39.88	53.95

with RMS defined by

$$RMS(\vec{y}) = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n}}\,, \tag{4.17}$$

for a n-valued $\vec{y}$.

Setting $n = 10\,min$, this function (equation 4.16) is applied to $\Delta B_{x,y}$, yielding two variables with 1-minute sampling resolution:

$$RRMS_{10}(\Delta B_x) \text{ and } RRMS_{10}(\Delta B_y)\,.$$

The variances (σ^2) of $\Delta B_{x,y}$, $RRMS_{10}(\Delta B_{x,y})$ and the 10-minute running mean $RM_{10}(\Delta B_{x,y})$ are computed and listed in Table 4.3. As a shorthand notation, the definitions

$$\psi_{x,y} = RM_{10}(\Delta B_{x,y})\,, \text{ and} \tag{4.18}$$

$$\omega_{x,y} = RRMS_{10}(\Delta B_{x,y}) \tag{4.19}$$

are made. The columns labelled α and β contain the percentage of total variance preserved by RM_{10} and $RRMS_{10}$:

$$\alpha = \sigma^2[\psi_i]/\sigma^2[\Delta B_i] \tag{4.20}$$

$$\beta = \sigma^2[\omega_i]/\sigma^2[\Delta B_i]\,, \tag{4.21}$$

with $i = (x, y)$. From the variances listed in Table 4.3 it is clear that, as suggested by Wintoft (2005), the root mean square conserves more of the variation of the signal than temporal averaging.

Thus the logarithm of $RRMS_{10}(\Delta B_{x,y})$ is used in order to limit the range of the output parameter amplitude. Figure 4.3 depicts the "development" from geomagnetic field to the two output parameters $log(\omega_x)$ and $log(\omega_y)$, with ω_x and ω_y defined by equation 4.19 . The chosen output parameters $log(\omega_{x,y})$ are to be used to predict mean-square GIC magnitudes (Wintoft et al., 2005). In the next chapter these output parameters are predicted using the input parameters described in Section 4.1.

The horizontal ground magnetic field components, B_x and B_y, used to compute the output parameters are measured at Hermanus by the Hermanus Magnetic Observatory. The data set employed for the prediction model development (see Chapter 5) covers a seven year period, from 1999 to 2005, measured at 1-minute intervals.

4.3 Summary

Twelve solar wind-related quantities measured with satellite based instruments at 16 and 64 second intervals over a seven year period (1999 – 2005) are considered for the development of the input space, while two geomagnetic quantities are considered for the output space. The sets of input and output parameters described in this chapter represent the *total set* of measurements considered as possible inputs and outputs to the neural network model. In the following chapter these parameters (and functions thereof) are tested and adapted, along with neural network parameters and other constraints, so as to construct an optimal prediction model.

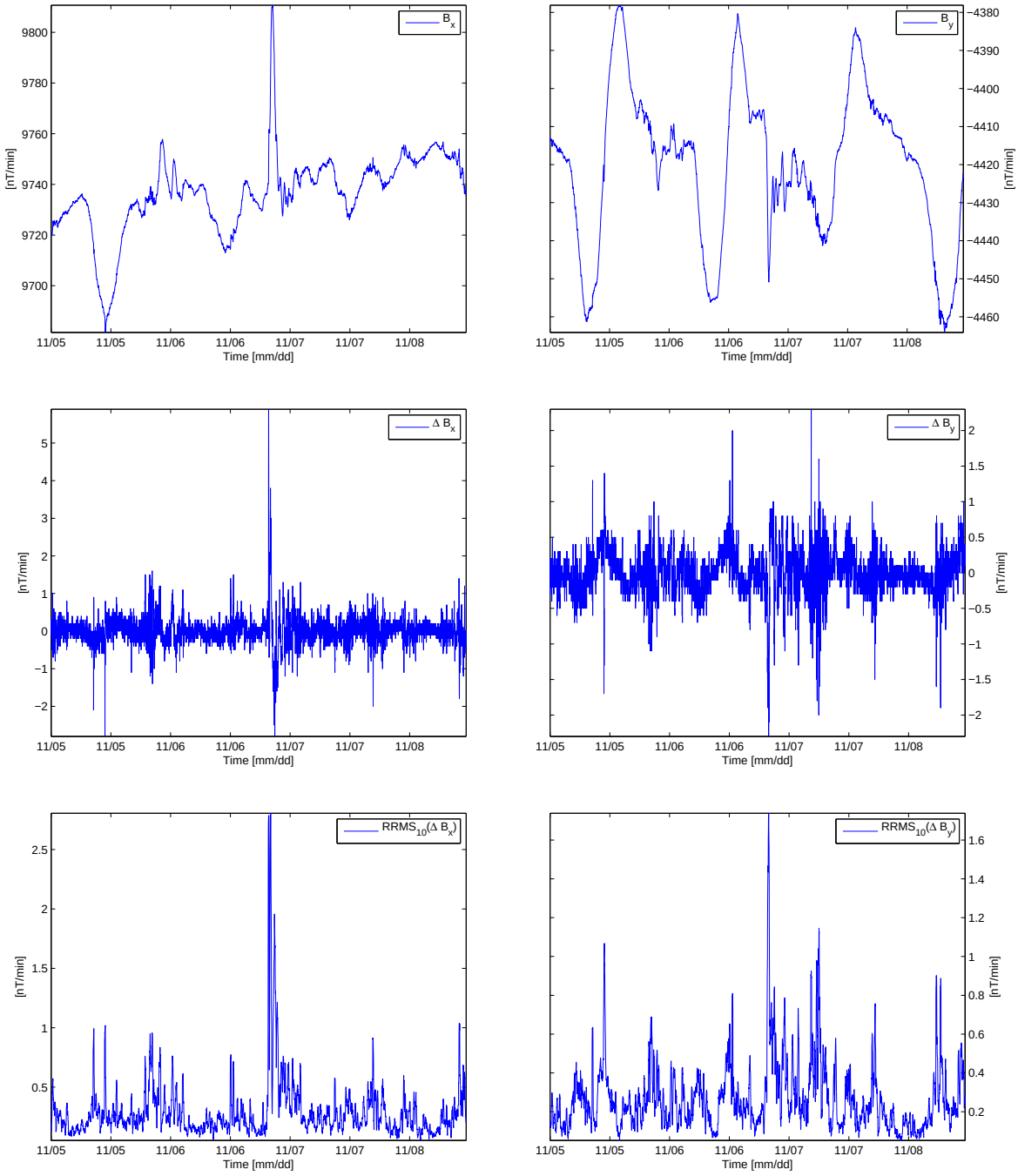

Figure 4.3: Development from measured geomagnetic field ($B_{x,y}$) to the output parameters, ω_x and ω_y.

Chapter 5

Model Development

This chapter describes the development of a neural network-based prediction model using solar wind-based input parameters (Section 4.1.4) and predicting output parameters derived from the geomagnetic quantities ΔB_x and ΔB_y (see Section 4.2). The chapter is divided into 5 sections. Each section describes an aspect of the development process, highlighting the choices made during development. In the first section below the choice of neural network type is described and motivated.

5.1 Network Type Determination

Given the cause-effect structure of a neural network training data set – the input parameter being the cause of the corresponding output – the input parameter values presented to the network should be physically responsible for the corresponding output values in order for the model to learn the underlying relation correctly. In the case of a feed-forward neural network, this condition is rigid in the sense that the solar wind parameter at index i of the input vector should correspond exactly to the *output parameter* value at the same index i. However, an inherent and significant delay between the time of solar wind measurement (by ACE at L1) and the time of its geomagnetic effect when it eventually interacts with the magnetosphere, exists on account of the time it takes to travel from L1 to Earth. If the said delay, t_d, were constant, this problem could be dealt with by shifting the output parameter vector by t_d, such that a solar wind measurement at time t corresponds to the geomagnetic parameter at time $t + t_d$, and training the network with this modified data set.

From Figure 5.1, however, it is clear that the solar wind velocity, and hence the propagation time from L1 to Earth is highly variable. Correcting for an inconsistent delay in the manual manner mentioned above would be difficult at best. A more elegant and natural way of solving this problem is to compensate for the delay by extracting the underlying structure

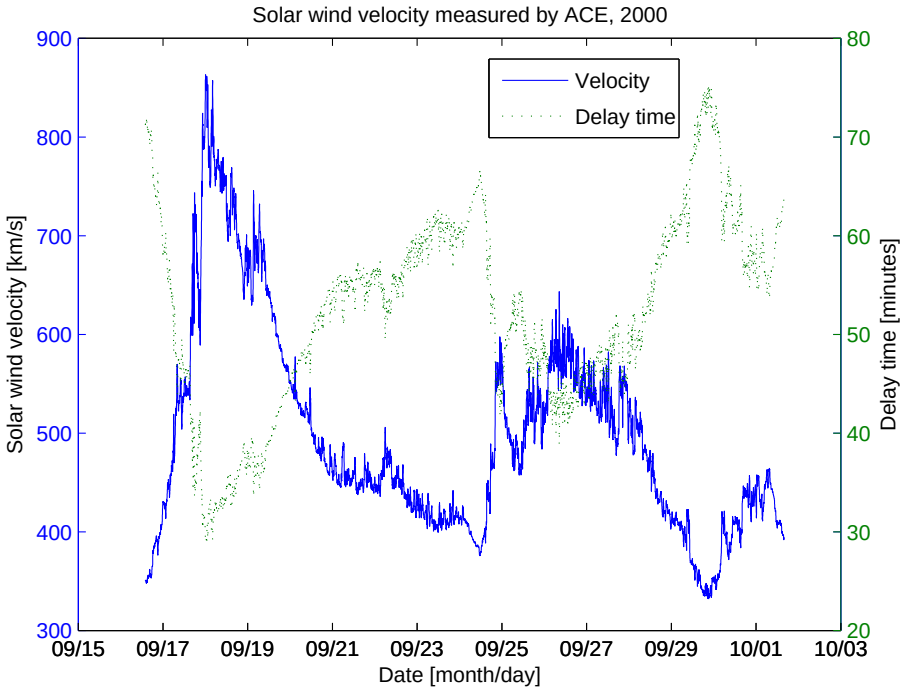

Figure 5.1: Solar wind velocity measured by ACE during 2000 with the delay time t_d from ACE to Earth. The delay time was computed by dividing the distance $1.5 \times 10^6\, km$ from L1 to Earth by the solar wind speed.

within the data set itself. This is achieved by employing a recurrent neural network, specifically of the *Elman* type, instead of a simple feed-forward network. As described in Chapter 3, Elman networks possess an internal memory of its previous states. Practically, this results in the network correcting for the delay by virtue of its internal structure, i.e. without externally applied modifications made to the data set.

5.2 Choice of Training and Testing Data Sets

As mentioned in Chapter 4 the training data set should include as much information as possible about the modelled system, the ideal being to include data representing every possible state. The ability of a neural network model to generalise – accurately responding to events not included in training – is dependent on the selection of data presented to the network training algorithm. Two Elman neural networks are trained to illustrate this statement.

The two networks are trained to form two prediction models, predicting the same geomagnetic field parameter. The output parameter for both networks is the absolute temporal difference of the horizontal geomagnetic field, $|\Delta B_H|$, measured during May 2005. The

Table 5.1: The setup of two networks used to predict the horizontal geomagnetic field variation $|\Delta B_H|$. Configurations are identical, while the training sets differ.

Model	**Network architecture**	**Training set**	**Testing set**
α	[10:20:15:1] Elman	2000.03.15 – 2000.05.15	2005.05.15 – 2005.05.17
β	[10:20:15:1] Elman	2000.01.01 – 2001.01.01	2005.05.15 – 2005.05.17

horizontal geomagnetic field is defined by

$$B_H \equiv \left| \sqrt{B_x^2 + B_y^2} \right| , \tag{5.1}$$

and $\Delta(\cdot)$ denotes the difference operator as defined by equation 4.13, to form $|\Delta B_H|$. The difference operator serves as an approximation of the temporal derivative (see equations 4.11 and 4.12). Both networks employ the same input parameter set of solar wind parameters, as defined in Table 4.1:

$$B^{imf},\ B^{imf}_{x,y,z}\,,\ V_{x,y,z}\,,\ V_p,\ N_p,\ N_\alpha\,.$$

However, the two training sets span different intervals: a two-month set and a one-year set. The configurations of the two models are listed in Table 5.1. The aim is to observe the difference in performance between employing a large (1 year) and small (2 month) training set. The two-month period stretches from 15 March to 15 May 2000, containing the geomagnetic storm of 6-7 April 2000; the one-year training set covers the year 2000, from day 1 to day 365.

A prediction based on training with two months' worth of one-minute sampled data (model α) yields the *green* curve in Figure 5.2. Training with a one-year data set (model β), predicting the same 2005 event, yields the *red* curve. The blue curve denotes the measured, or targeted, values of $|\Delta B_H|$.

The difference in prediction performance between the two models is clearly observed in Figure 5.2. Although the prediction resulting from model β is far from perfect, the improvement over the performance of model α is significant. From this result it is clear that the smaller training set is not sufficient. It is not necessarily the relatively small size of its training data set that causes model α to fail, but the inadequate representation of the modelled system in the training set.

Consider model β, where the training data set includes all measurements from the year 2000. During this period the bulk of measurements were made during solar-quiet times. Due to the statistical nature of the neural network training algorithm, rarely occurring phenomena are perceived as outliers or anomalies; and since GIC's are directly related to

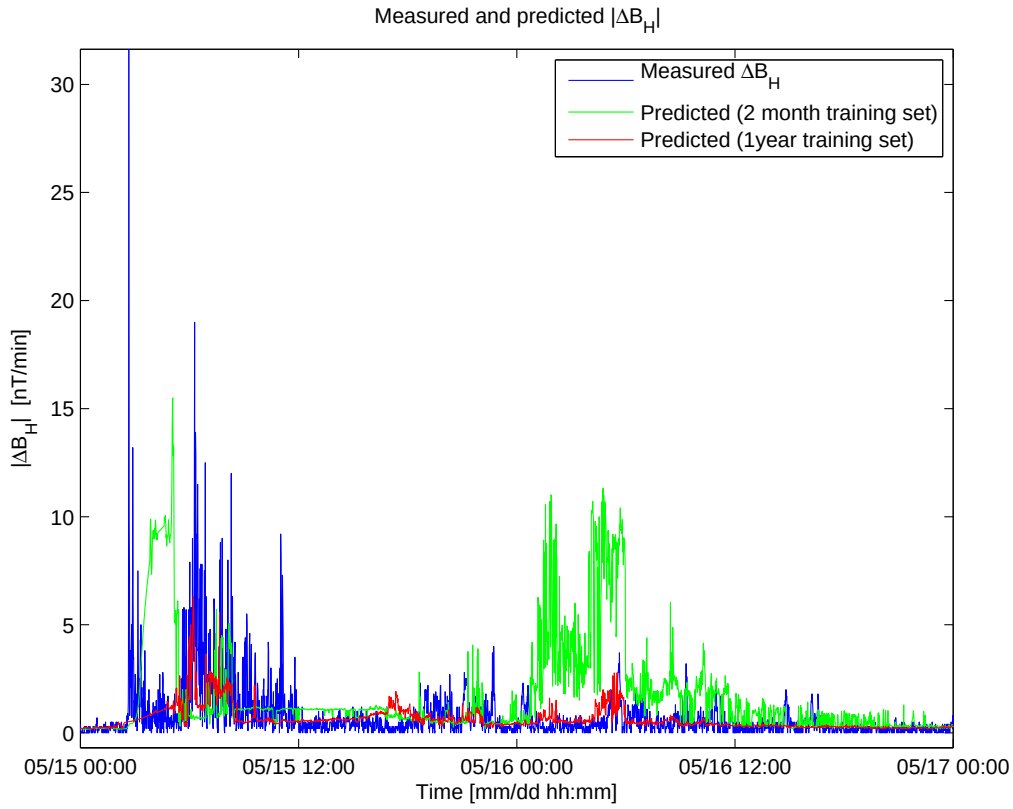

Figure 5.2: The performance of two identical Elman neural networks is compared. The green curve represents the network trained on a data set spanning two months, while the red curve depicts an identical network trained using a one-year data set. Also see Table 5.1.

solar-*active* periods, this is less than ideal. Thus, using indiscriminately selected data as the training data set, as in Figure 5.2, results in an underestimation of storm-time output response.

Solar activity, and hence the structure of the solar wind and its processes, changes over the course of a solar cycle. Hence, it is important to include in the data set measurements taken throughout the evolution of the (last) solar cycle. Solar wind and geomagnetic measurements were gathered respectively from ACE and HMO, for the seven-year period from 1999 to 2005. As mentioned in Chapter 4, the solar wind measurements from ACE are interpolated to one-minute samples so as to coincide with the geomagnetic field measurements. Utilising the entire set as training input to a neural network model would result in the underestimation of the predicted outputs during a solar-active period, due to the relative rarity of active periods compared to quiet periods. A process for selecting relevant sequences within the master data set (1999 – 2005) is derived from the method followed

by Wintoft (2005). The aim is to identify instances in the *master set* (1999 – 2005) where the magnitude of the relevant output parameter exceeds some threshold; and to extract sequences of all input and output parameters *around* the identified instances. This results in a subset of the master data set which excludes much of the solar-quiet measurements. The selection process is described below.

Consider the master data set ($\mathcal{M}$) as a collection of p input (in_i) and q output parameter (out_i) vectors, each comprising N one-minute samples stretching from 1999 to 2005:

$$\mathcal{M} = \begin{bmatrix} in_1(1) & in_2(1) & \cdots & in_p(1) & out_1(1) & \cdots & out_q(1) \\ in_1(2) & in_2(2) & \cdots & in_p(2) & out_1(2) & \cdots & out_q(2) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ in_1(N) & in_2(N) & \cdots & in_p(N) & out_1(N) & \cdots & out_q(N) \end{bmatrix} \tag{5.2}$$

The sequences are identified by means of the following algorithm.

1. Select relevant output parameter f from $\mathcal{M}$: $f = out_k$
2. Define threshold z for output parameter f, where f is a vector of length N
3. Identify all W instances j of f, such that $f(j) \geq z$
4. For each j: let $g_j \equiv \{j - 48\,hrs\,, \ldots\,, j + 48\,hrs\}$
5. Merge overlapping g_j sequences, such that $Q \leq W$ sequences of indices exist

Thus sequences g_i, with $i = \{1, \ldots\,, Q\}$, exist, each at least 96 hours in length. These sequences of indices are applied to $\mathcal{M}$ in order to find the relevant values for each input and output parameter, forming the combined training and testing data set, $\mathcal{T}$:

$$\mathcal{T} = \begin{bmatrix} in_1(g_1) & in_2(g_1) & \cdots & in_p(g_1) & out_1(g_1) & \cdots & out_q(g_1) \\ in_1(g_2) & in_2(g_2) & \cdots & in_p(g_2) & out_1(g_2) & \cdots & out_q(g_2) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ in_1(g_Q) & in_2(g_Q) & \cdots & in_p(g_Q) & out_1(g_Q) & \cdots & out_q(g_Q) \end{bmatrix} \tag{5.3}$$

Thus, $\mathcal{T}$ is a subset of $\mathcal{M}$, containing those values of each input and output parameter considered as relevant to the modelling process.

The set $\mathcal{T}$ includes all data to be used for training and testing purposes. From $\mathcal{T}$ a number of the sequences are set apart to be used as a testing data set. Subsequent development is based on the ability of a network to predict the testing data set sufficiently.

Table 5.2: Details of training and testing data sets. Six sequences are chosen from both the ω_x and ω_y training sets to be used as the testing data sets.

Output parameter	**Data points**	**Training sequences**	**Testing sequences**	**Total sequences**
ω_x	429688	49	6	55
ω_y	234182	25	6	31

The procedure above is followed to yield the training and testing data sets upon which all subsequent network training is based. Since two parameters ($\omega_{x,y}$) are predicted (Section 4.2), two separate selection procedures are followed, resulting in separate training and testing data sets for $log(\omega_x)$ and $log(\omega_y)$ predictions. For both selection procedures the threshold is set to $z = 5\,nT/min$. Details of the resulting data sets are summarised in Table 5.2.

5.3 Prediction Lead Time

The inherent prediction lead time due to the propagation time of solar wind plasma from L1 to the Earth provides an average of 55 minutes of prediction lead time. However, this natural lead is not constant as is seen in Figure 5.1, and therefore unreliable to a forecasting service. An additional lead time τ is manually inserted by shifting the input and output parameter vectors with respect to each other (Wintoft, 2005). This results in the prediction function $\mathcal{F}$, based on a trained neural network, approximating the output parameter a further τ minutes in the future:

$$out(t + \tau\, min) \approx \mathcal{F}(in(t))\,. \tag{5.4}$$

An inserted lead time of $\tau = 30$ minutes is employed throughout, resulting in a total lead time of $\tau + d$, with d the response time of the local geomagnetic field to solar wind and IMF structures.

5.4 Input Parameter Selection

Thirty variables are presented as possible input parameters to the prediction model. Many neural networks are trained to predict $log(\omega_{x,y})$ and $\omega_{x,y}$, each with a different set of input parameters. Other free parameters (type of network and learning function, for example) are kept constant. In this way, each trained network is representative of its corresponding set of input parameters. The trained networks are all used to predict the same testing data set, and their prediction performance calculated. Prediction performance is gauged

by the correlation coefficient

$$cc(\theta, \tilde{\theta}) = \frac{cov(\theta, \tilde{\theta})}{\sqrt{cov(\theta, \theta) \times cov(\tilde{\theta}, \tilde{\theta})}} \,. \tag{5.5}$$

The measured output parameter is denoted by θ, while $\tilde{\theta}$ denotes the predicted quantity and $cov(\cdot\,,\,\cdot)$ denotes the covariance of two variables. The output parameters $log(\omega_x)$ and $log(\omega_y)$ are used to judge the performance of the predictions, in the sense that the networks resulting in the highest correlation between measured and predicted $log(\omega_{x,y})$ are identified and their corresponding sets of input parameters are selected as the *optimal* sets. The parameters ω_x and ω_y are only included in the output layer to serve as a comparison between the predictions of $\omega_{x,y}$ and $log(\omega_{x,y})$.

All networks trained are $[n : n + 2 : 4]$ Elman neural networks, with

- n input nodes,
- $n + 2$ hidden nodes in one hidden layer,
- and 4 output nodes.

The networks are all trained for 1000 cycles of the training algorithm, using the *back-propagation* learning function (with learning parameter 0.05), adapted for Elman networks. All networks are trained and tested using the instances of the output and relevant input parameters, as identified by the process described in Section 5.2.

Firstly, 30 networks are trained, each with one input parameter selected from the set of inputs. The trained networks are processed and test predictions are made, yielding 30 correlation coefficients. The input parameter resulting in the largest correlation coefficient is selected. In the next round of training, 29 networks are trained (each with 2 input parameters) varying the remaining 29 inputs in addition to the parameter selected after the first round. The process continues: train 28 networks, each with 3 input parameters – the 2 selected in the previous two rounds, varying the remaining 28 parameters with each network trained, and so on. The process is repeated in this manner until the increase in correlation magnitude (as is expected, due to the increasing number of input sources) becomes negligible.

The process is generally described by the *pseudo-code* algorithm below.

Define: $\mathcal{A}$ is the set of all N input parameters
$\mathcal{S}$ is the set containing the selected inputs, i.e. $\mathcal{S} \subseteq \mathcal{A}$
$\mathcal{R}$ is the set of remaining inputs, i.e. $\mathcal{R} = \mathcal{A} \setminus \mathcal{S}$

Let $\mathcal{A} \leftarrow [\text{input}_1, \text{input}_2, \ldots, \text{input}_N]$
Let $\mathcal{S} \leftarrow \emptyset$ (the empty set)
Let $\mathcal{R} \leftarrow \mathcal{A}$

While increase in correlation $> \epsilon$:
for all $i \in \mathcal{R}$:
Train network with input parameters $\mathcal{I} \leftarrow \mathcal{S} \cup \mathcal{R}_i$
Process network, make test prediction, compute correlation coefficient:
$cc_i = cc(\text{measured, predicted})$
Identify input parameter corresponding to k, where $cc_k > cc_{i \neq k}$
Add input k to $\mathcal{S}$: $\mathcal{S} \leftarrow \mathcal{S} + k$
Remove input k from $\mathcal{R}$: $\mathcal{R} \leftarrow (R - k)$

Upon completion of this algorithm, the set $\mathcal{S}$ yields the optimal set of input parameters. The process is executed for the prediction of $log(\omega_x)$ and $log(\omega_y)$ separately, such that individual input parameter sets are identified for the prediction of each output. The progressive improvement in prediction performance is illustrated in Figure 5.3 and listed in Tables 5.3(a) and 5.3(b) below. The top and bottom panels depict the progression of $log(\omega_x)$ and $log(\omega_y)$ prediction, respectively.

Table 5.3: Correlations for the prediction of (a) $log(\omega_x)$ and (b) $log(\omega_y)$, as input parameters are added. Also see Figure 5.3.

(a) Correlation coefficients of measured and predicted $log(\omega_x)$.

No. of Inputs	Chosen Inputs	$cc(\log \omega_x, \widetilde{\log \omega_x})$
1	$rm_{10}(B^{imf})$	0.5453
2	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$	0.6807
3	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rm_{10}(N_p)$	0.7210
4	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rm_{10}(N_p)$, LTC	0.7359
5	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rm_{10}(N_p)$, LTC, $rstd_{10}(N_\alpha)$	0.7440

(b) Correlation coefficients of measured and predicted $log(\omega_y)$.

No. of Inputs	Chosen Inputs	$cc(\log \omega_y, \widetilde{\log \omega_y})$
1	$rm_{10}(B^{imf})$	0.5719
2	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$	0.6592
3	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rstd_{10}(N_p)$	0.6876
4	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rstd_{10}(N_p)$, LTC	0.7150
5	$rm_{10}(B^{imf})$, $rm_{10}(V_p)$, $rstd_{10}(N_p)$, LTC, LTS	0.7309

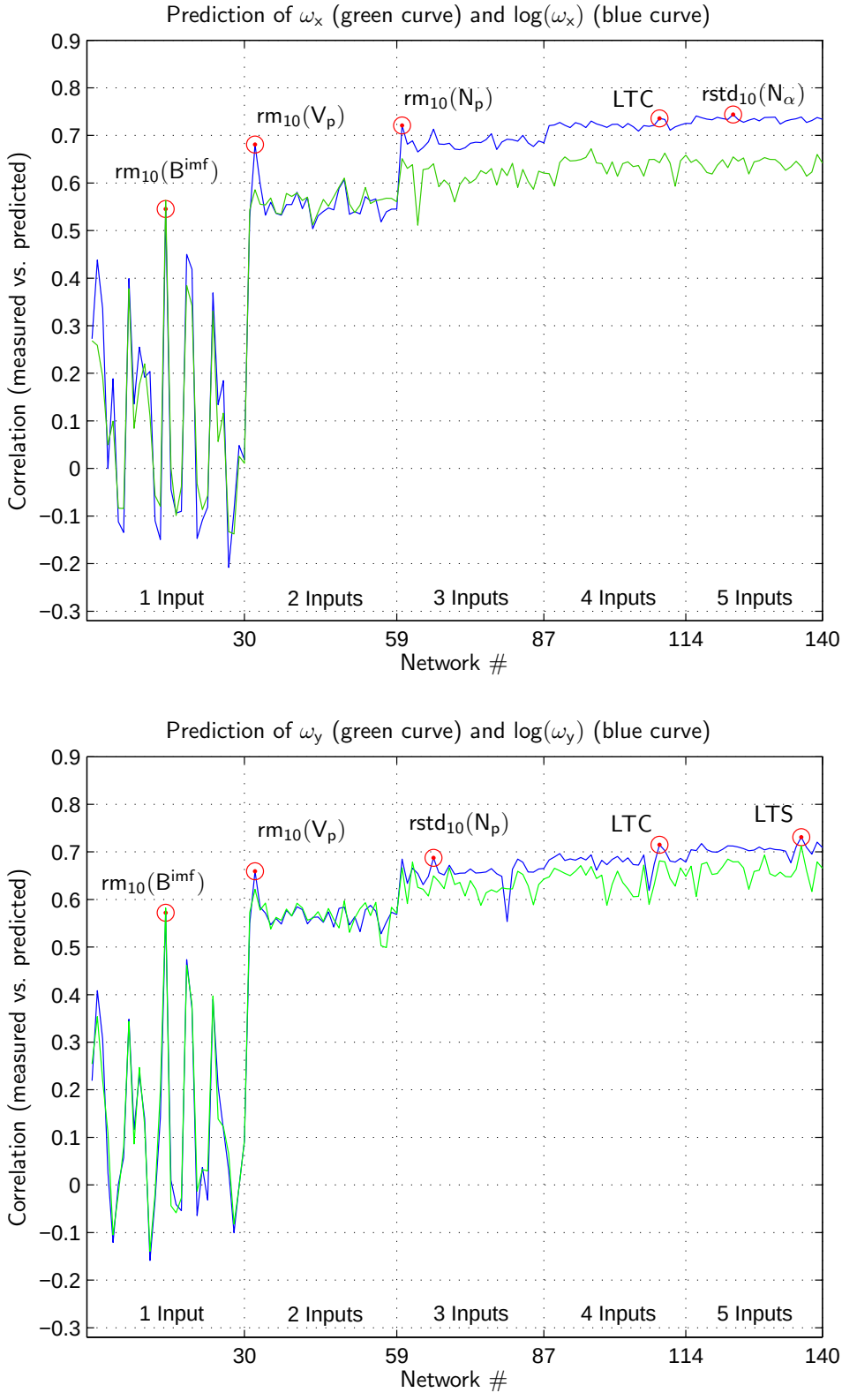

Figure 5.3: Correlation between measured and predicted values of $log(\omega_x)$ (top panel) and $log(\omega_y)$ (bottom panel), as input parameters are varied and added. The "winner" of each round is labelled and denoted by a red circle. The progressive performances in predicting ω_x and ω_y are also included (green curves).

Two sets of "optimal" input parameters are identified and listed in the last rows of Tables 5.3(a) and 5.3(b).

An inherent measure of uncertainty is present in the training of neural networks and the subsequent test predictions made. Since the selection of networks trained and tested in this section all utilise different sets of input parameters, every network is inevitably trained with different training and validation data sets. Apart from the difference in input parameters, different validation sets may also introduce a further measure of uncertainty in prediction results. Due to this fact, the correlation coefficients calculated in this section, and the differences between them, are only meaningful above a certain threshold. When the differences between corresponding correlation values are smaller than this threshold, choices based on these correlations are ambiguous.

To test this, 10 identical networks are trained using the same training set and predicting the same testing set, but the distribution of validation data is randomly chosen for each network. Input parameters

$$rm_{10}(B^{imf}),\ rm_{10}(V_p),\ rm_{10}(N_p),\ LTC,\ rstd_{10}(N_\alpha)$$

are included and $log(\omega_x)$ is predicted, and

$$rm_{10}(B^{imf}),\ rm_{10}(V_p),\ rstd_{10}(N_p),\ LTC,\ LTS$$

is used to predict $log(\omega_y)$. Each independently trained network employs the same input and output parameters. The correlations between the measured and predicted outputs for each of the 10 networks are presented in Table 5.4. From Table 5.4 the maximum and minimum correlations are identified for each output parameter. These are printed in boldface in rows 1, 8, 2 and 4 of Table 5.4. The difference between the maximum and minimum values are

$$\epsilon_x = 0.7457 - 0.7404 = 0.0053$$

for $\log(\omega_x)$ prediction and

$$\epsilon_y = 0.7334 - 0.7236 = 0.0098$$

for $\log(\omega_y)$ prediction. Thus the selection of LTC and $rstd_{10}(N_\alpha)$ for $log(\omega_x)$ prediction is ambiguous because correlations resulting from the selection of other inputs fall within ϵ_x of the maximum correlation yielded in Table 5.3(a) and Figure 5.3 (top panel).

This is illustrated in Figure 5.4 which displays the correlations yielded when using 4 and 5 input parameters, i.e. the correlations also found in Figure 5.3 (top panel) for networks 87

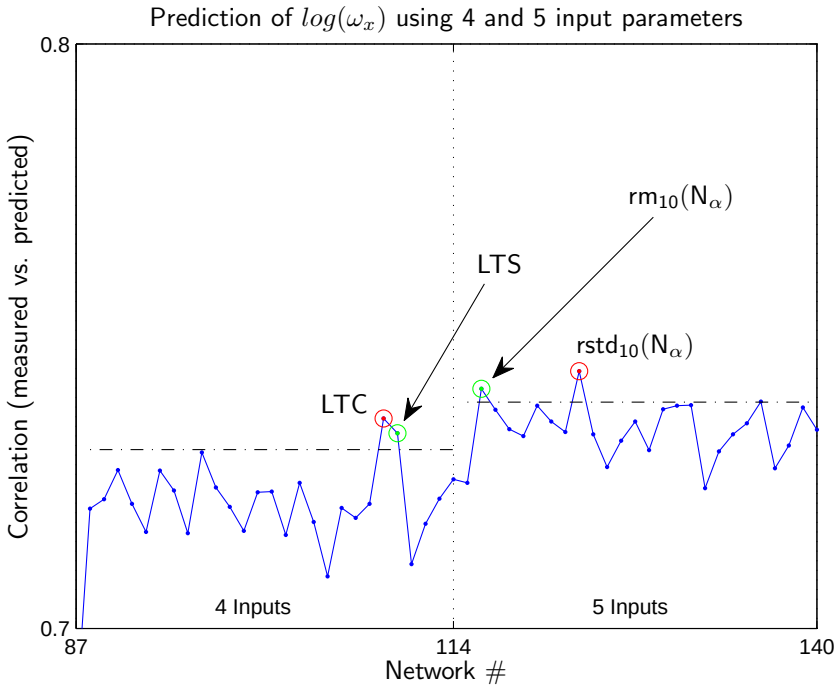

Figure 5.4: Correlations attained when predicting $log(\omega_x)$ using 4 and 5 input parameters. The deviation (ϵ_x) from each maximum is denoted with dashed lines. Green circles denote the "second placed" input parameters. Note that both the winners (red circles) and second placed inputs (green circles) are related: LTS – LTC and $rm_{10}(N_\alpha)$ – $rstd_{10}(N_\alpha)$. Also see Figure 5.3 (top panel).

to 140. Dashed lines denote deviations of magnitude ϵ_x from each maximum (red circles). Green circles mark "second placed" correlations, from which the ambiguity arises. Since both "second placed" inputs fall within ϵ_x of the two maximum correlations, it is uncertain whether LTC or LTS should be selected as the fourth input and whether $rm_{10}(N_\alpha)$ or $rstd_{10}(N_\alpha)$ should be selected as the fifth input. Note that the two ambiguous inputs in each pair are derived from the same measurements; i.e. LTC and LTS are both derived from the local time, and $rm_{10}(N_\alpha)$ and $rstd_{10}(N_\alpha)$ are both derived from the α to proton number density ratio, N_α. Thus, although the choice of fourth and fifth input parameters are somewhat ambiguous, we know that a weak dependence of ω_x on local time and N_α exists.

5.5 Neural Network Configuration

The previous section described the selection of input parameters from a larger set in order to optimally predict the output parameters. In the $log(\omega_x)$ case the neural network comprises 5 input parameters, 7 hidden nodes in one hidden layer, and 4 output parameters (since $\omega_{x,y}$ predictions are also made). Generally, and specifically in the case of Section

Table 5.4: The best performing networks from Section 5.4 are retrained 10 times each, in order to find the variation in prediction performance (correlation) due to the natural uncertainties within the network training algorithm. The minimum and maximum values are highlighted for each case.

Repetition	$cc(\log \omega_x, \widetilde{\log \omega_x})$	$cc(\log \omega_y, \widetilde{\log \omega_y})$
1	**0.7404**	0.7241
2	0.7451	**0.7334**
3	0.7439	0.7320
4	0.7430	**0.7236**
5	0.7448	0.7260
6	0.7443	0.7319
7	0.7428	0.7319
8	**0.7457**	0.7307
9	0.7432	0.7290
10	0.7430	0.7321

5.4, the number of input and output parameters are determined by *problem-specific constraints*, i.e. the constraints set by the modelled system and the goals of the modelling process. This leaves the topology of the hidden layer(s) to be adapted.

Section 5.4 used a simple *number of hidden nodes* $= 2 +$ *number of inputs* formula for determining the number of hidden nodes used (in one hidden layer). According to Hagan et al. (1996), no general algorithm for determining the optimal number of hidden nodes exist.

In order to find the optimal hidden layer topology, the number of hidden nodes are varied from 1 to 14 in one hidden layer, and from 2 to 28 in two hidden layers. A maximum of two hidden layers is used, considering the statement made by Hagan et al. (1996), namely that one or two hidden layers are sufficient for most problems. The optimal set of input parameters selected in Section 5.4

$$\Big[rm_{10}(B^{imf}),\, rm_{10}(V_p),\, rm_{10}(N_p),\, LTC,\, rstd_{10}(N_\alpha)\Big]$$

is used to predict $\omega_{x,y}$ and $log(\omega_{x,y})$. Thus, the same testing and training data sets as defined in Section 5.2 are used. The network configuration is kept constant, except for the number of hidden nodes which are varied. As is the case in the previous section, the correlation coefficient (equation 5.5) is used as the measure of performance. Figure 5.5 and Table 5.5 illustrate the results.

As the number of hidden nodes are increased from 1 to 5 (in one hidden layer), the correlation between measured and predicted measurements increases steadily from approximately

0.71 to 0.7480. With 6 to 14 hidden nodes in one layer, the performance remains approximately constant around 0.74. In the double hidden layer configuration the prediction performance rises from approximately 0.71 to 0.744 for 2 to 10 hidden nodes. As the number of hidden nodes is increased from 14 to 26 (7:7 to 13:13), the correlation remains approximately constant around 0.73. At 12 and 28 hidden nodes, the correlations are low (0.7202 and 0.7181) compared to other configurations. The reason for this is not clear.

Overall, one hidden layer seems to yield slightly better prediction results. Specifically, the use of 5 hidden nodes proves successful. Based on these results, one hidden layer containing more than 3 nodes is sufficient.

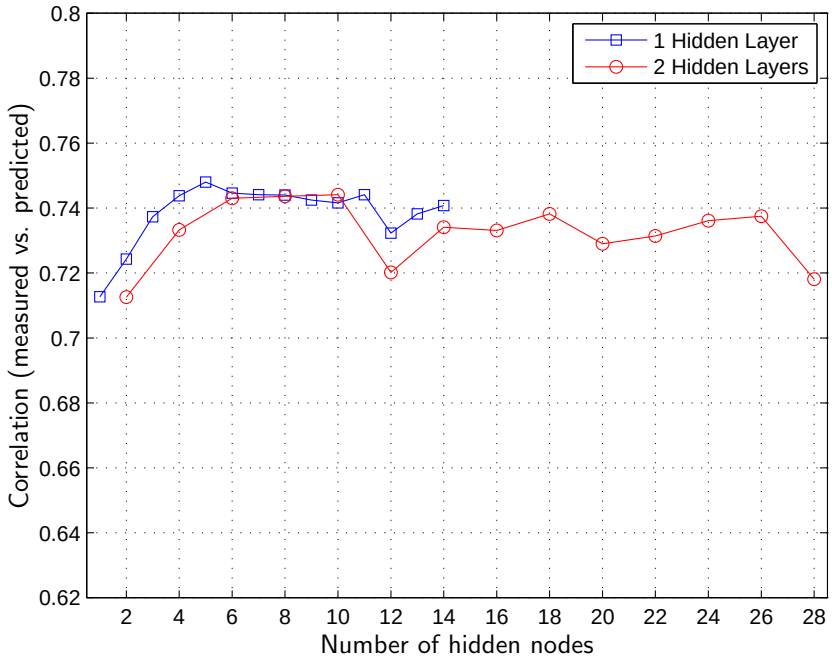

Figure 5.5: Prediction performance versus increasing number of hidden nodes. One (blue squares) and two (red circles) hidden layers are used. The correlation coefficients are listed in Table 5.5.

Table 5.5: Predictive performance for different numbers of hidden nodes in one (first and second columns) and two (third and fourth columns) hidden layers. The coefficients listed here are illustrated in Figure 5.5.

Hidden Nodes (1 layer)	$cc(\log\omega_x, \widetilde{\log\omega_x})$	Hidden Nodes (2 layers)	$cc(\log\omega_x, \widetilde{\log\omega_x})$
1	0.7127	1:1	0.7126
2	0.7243	2:2	0.7333
3	0.7373	3:3	0.7430
4	0.7438	4:4	0.7436
5	0.7480	5:5	0.7441
6	0.7446	6:6	0.7202
7	0.7441	7:7	0.7341
8	0.7440	8:8	0.7331
9	0.7425	9:9	0.7382
10	0.7416	10:10	0.7290
11	0.7441	11:11	0.7314
12	0.7323	12:12	0.7361
13	0.7382	13:13	0.7375
14	0.7408	14:14	0.7181

5.6 Prediction Results

To summarise, this chapter described the development of a neural network prediction model. The geomagnetic quantities $log(\omega_x)$ and $log(\omega_y)$, defined in Section 4.2.1 are predicted with a 30-minute lead time from a subset of the input parameters listed in Section 4.1.5. Choices regarding the type of network employed, construction of testing and training data sets, input parameters and network configuration were made.

In the prediction of $log(\omega_x)$, a [5:7:4] Elman neural network, with input parameters

$$rm_{10}(B^{imf}),\ rm_{10}(V_p),\ rm_{10}(N_p),\ LTC \text{ and } rstd_{10}(N_\alpha)$$

achieved 0.74 prediction accuracy (Fig. 5.6), where *prediction accuracy* refers to the correlation coefficient between measured and predicted output parameter values. Correspondingly, a [5:7:4] Elman network with input parameters

$$rm_{10}(B^{imf}),\ rm_{10}(V_p),\ rstd_{10}(N_p),\ LTC,\ LTS$$

achieved 0.73 accuracy in predicting $log(\omega_y)$ with a 30-minute lead (Fig. 5.7).

Recalling from Section 5.2, six sequences of input and output parameter measurements are used as the testing data sets for the predictions of $log(\omega_x)$ and $log(\omega_y)$, respectively.

The best predictions of $log(\omega_x)$ and $log(\omega_y)$, according to Section 5.4, are plotted with their corresponding measured values in Figures 5.6 and 5.7. The predictions are made with a 30-minute prediction lead (described in Section 5.3).

The general structure of the target data set is predicted with reasonable accuracy, but the finer structures within the output variable are not modelled very well. Other authors (Wintoft, 2005; Koskela et al., 1996), employing recurrent neural networks for time series prediction have found the same behaviour. The inability of the model to predict the highly variable components may be due to the statistical nature of the training process, in which more rapid variations are smoothed out and inevitably lost.

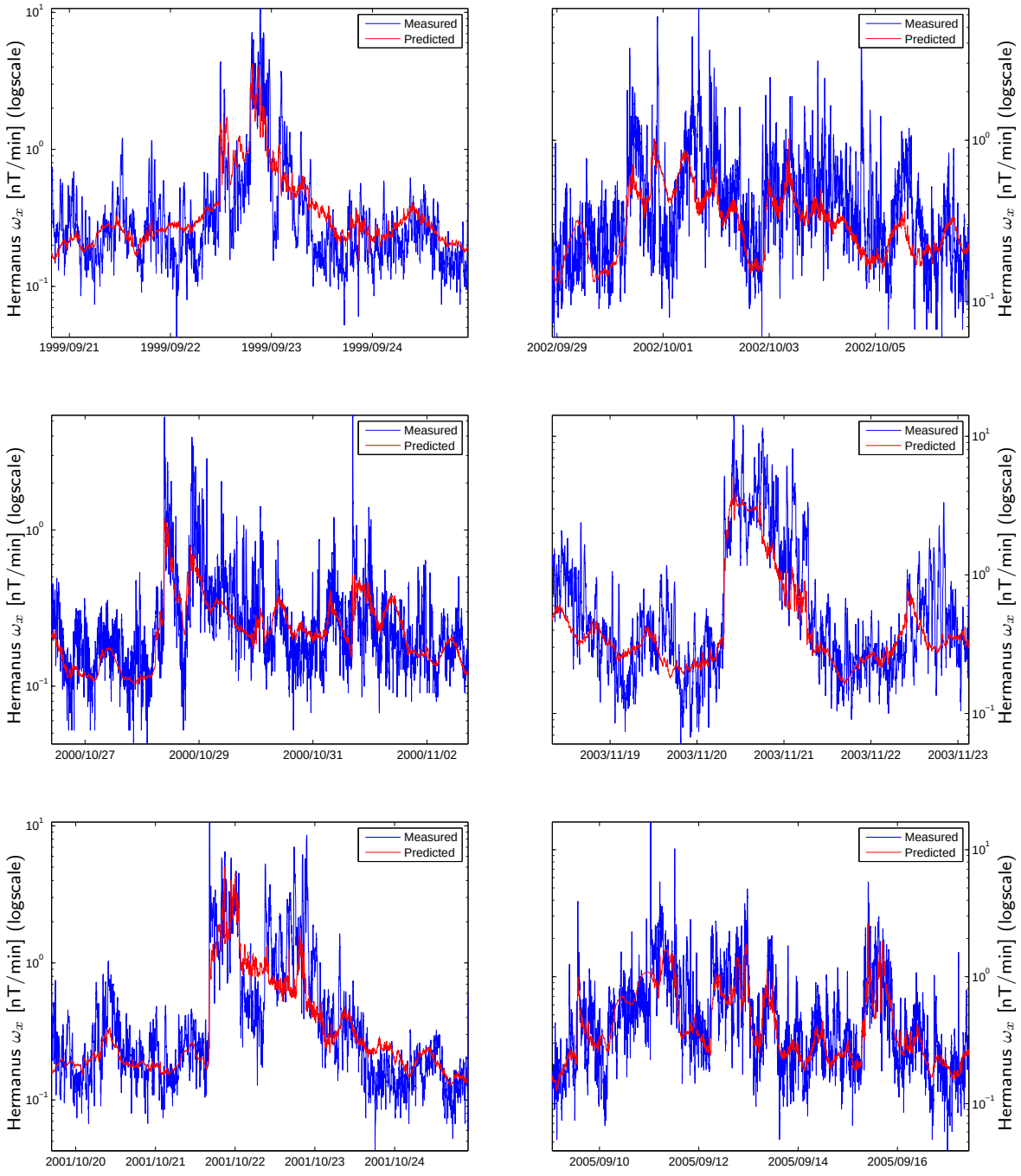

Figure 5.6: Measured and predicted values of $log(\omega_x)$ for six sequences of testing data recorded at the Hermanus Magnetic Observatory. The overall prediction performance is 0.74.

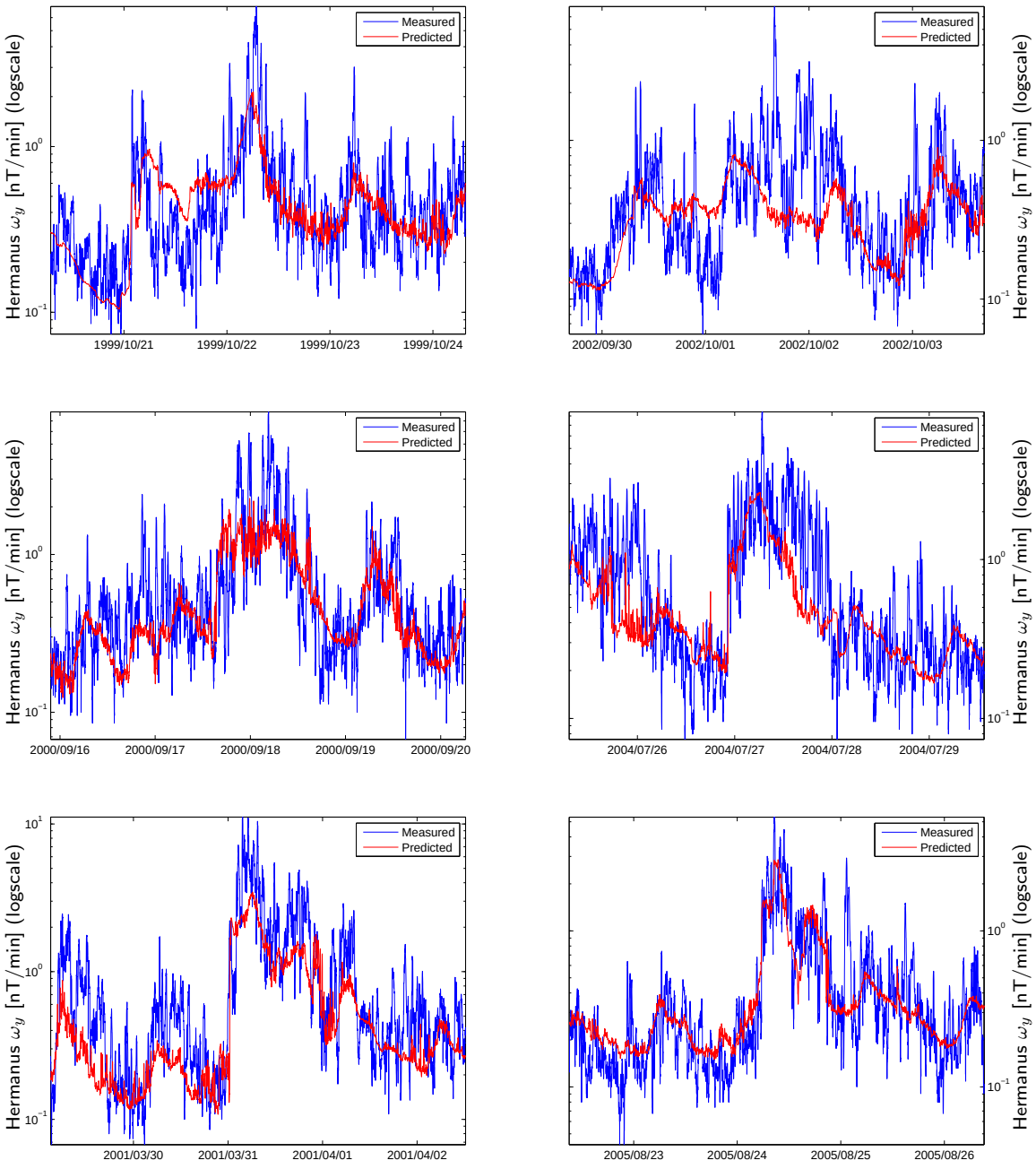

Figure 5.7: Measured and predicted values of $log(\omega_y)$ for six sequences of testing data recorded at the Hermanus Magnetic Observatory. The overall prediction performance is 0.73.

5.7 Discussion of Selected Input Parameters

Section 5.4 describes a process of input parameter selection with regards to the prediction of $log(\omega_{x,y}) = log\,[RRMS_{10}(\Delta B_{x,y})]$. Two sets of "optimal" input parameters have been determined (see previous section). According to this process, the total IMF magnitude (B^{imf}), the particle number density (N_p) and the solar wind bulk speed (V_p) are the most important solar wind parameters in the prediction of the geomagnetic field output parameters. These three parameters are related to two pressure terms. The IMF magnitude (B^{imf}) defines the magnetic pressure

$$p_B = B^2/2\mu_0 \tag{5.6}$$

in the solar wind, and the solar wind dynamic pressure, p_{sw}^{dyn} (equation 4.4), is related to the velocity (V_p) and density (N_p) parameters. The latter of these terms, p_{sw}^{dyn}, is related to the compression of the magnetosphere by the incoming solar wind (Wintoft, 2005; Russel et al., 1999).

The emergence of B^{imf} as an influential input parameter could be due to the coupling between charged particle density in the solar wind, and the magnetic field strength within the solar wind. The relation is understood through the frozen-in flux condition from magneto-hydrodynamics. Firstly, it is assumed that the frozen-in flux condition holds, i.e. each plasma particle is attached to a magnetic field line and remains attached to that field line. Now, an increase in particle density will result in a magnetic field line density increase. Although magnetic field lines are theoretical constructions, it is understood that field lines converge when approaching regions of strong magnetic field and diverge as a region of weak field is approached (Purcell, 1965, pg. 19). Hence, as the particle density within the solar wind plasma increases, the magnitude of the frozen-in magnetic field also increases.

A similar study conducted by Wintoft (2005) found a 0.79 correlation between measured and predicted $RMS(\Delta B_{x,y})$ and listed B_z^{imf}, V_p and the standard deviation of the density $std(N_p)$ as the most important input variables. Some differences between the Wintoft model and the one presented here exist. Our study is based on lower-middle latitude ($34.27^\circ\,S$) geomagnetic field measurements, whereas the Wintoft model is exclusively based on high-latitude measurements ($55.63^\circ\,N$ and $59.90^\circ\,N$). The number of "candidate" input parameters considered in this study outnumbers the selection considered by Wintoft (2005). Most notably, Wintoft did not consider the total IMF magnitude as a possible input, which proved the most influential parameter in our study. The different IMF components listed as influential parameters in this investigation and in the Wintoft (2005) model suggests that high latitude termporal geomagnetic field variations are due to sub-

storm activity (driven by reconnection which is dependent on B_z^{imf}) and lower-middle latitude variations are driven by the ring current effect, which is dependent on the total IMF strength, B^{imf}.

The selection criterion used in the selection of training and testing data sets is based on the variations in horizontal geomagnetic field with time, measured at Hermanus (see Section 5.2). In other words, no distinction with respect to the severity or source of the geomagnetic storms responsible for the temporal variations in $B_{x,y}$ is made. This could explain why B_z^{imf}, as a driver of substorm activity, is not identified by the input parameter selection process as a critical input to the prediction model.

Given the prediction results, it is argued that the selected sets of input parameters, especially

$$rm_{10}(B^{imf}),\ rm_{10}(V_p),\ rm_{10}(N_p) \text{ and } rstd_{10}(N_p)$$

could be considered as good precursors of horizontal field variation at middle-latitudes.

Chapter 6

Discussion and Concluding Remarks

6.1 Project Overview and Conclusion

The objective of this study is to determine if neural networks would provide a suitable tool for constructing a model capable of predicting GIC occurrence. The processes driving GIC's, through solar wind-magnetosphere-ionosphere interactions, are complex and highly non-linear. Due to the complexity and non-linearity of the modelled system, neural networks are identified as a viable basis for the prediction model.

Elman networks, a type of *recurrent* neural network, are employed due to their ability to correct for the inherent delay between input parameters (solar wind-based quantities measured at L1) and output parameters, which are based on ground magnetic field measurements. This is an example of how certain types of networks are better suited to a specific application. In this case the use of a simple feed-forward network would have demanded a *manual* correction for the delay in solar wind propagation time from L1 to the Earth's surface.

In order to develop the neural network-based model, input and output parameters are identified. The (predicted) output parameters selected are the 10-minute running root-mean-square of ΔB_x and ΔB_y, respectively. Predicting $log\,[RRMS_{10}(\Delta B_{x,y})]$ would enable the mean-square GIC magnitude to be known in advance, given that the coefficients which characterise the conductor network, a and b from equation 2.5, are accurate and available.

Input parameters are selected from a large set of "candidate" inputs, derived from *in-situ* solar wind measurements made by the ACE satellite. A robust and objective parame-

ter selection algorithm is used to identify the solar wind parameters bearing the most influence on the output parameters. The criteria for parameter selection are based on the magnitude of the correlation coefficient of measured and predicted output parameters yielded by using a particular set of inputs to make the prediction. The process is conducted separately for each output parameter, yielding two sets of "optimal" input parameters. The predictions resulting from the optimal input parameter selections result in a correlation between measured and predicted outputs of 0.74 for the $log\,[RRMS_{10}(\Delta B_x)]$ prediction and 0.73 for the $log\,[RRMS_{10}(\Delta B_y)]$ prediction. All predictions are made with a 30-minute prediction lead time.

Given the results presented here, it is reasonable to conclude that further investigation would yield a working neural network-based prediction model, capable of predicting GIC occurrence at least 30 minutes in advance, given the availability of suitable information regarding the configuration of the conductor network.

6.2 Future Work

Since the nature of this project is merely a feasibility study, it is clear that much needs to be done in order to construct a working GIC prediction model fit for integration into a regional space weather warning centre. A short discussion of possible future investigation follows.

A significant shortcoming in the development of this prediction model is the lack of available GIC measurements within the South African powerline network. Currently, measurements taken over a 3-day period, spanning the severe geomagnetic storm of October 2003, is the only available data set of South African GIC measurements. Clearly, a much larger data set is required to compare the predictions made with actual GIC occurrences. Negotiations are currently underway with the South African power utility (Eskom) regarding the availability of future measured GIC data.

It would be logical to expand the model to include more magnetometer stations. This study only considers measurements made at Hermanus, South Africa. Expansion entails collecting and processing geomagnetic field measurement data obtained from other stations over the last solar cycle, and repeating the modelling process for both output parameters. Depending on the data made available by Eskom, the additional magnetometer stations should ideally be situated near the relevant Eskom transformer substations, where GIC's are measured. Expanding the model to include more locations and comparing results with GIC measurements would provide more information as to the physical processes driving GIC's at mid-latitudes. It would also be worthwhile to use the same method to develop

a simple prediction model based on high-latitude geomagnetic data. That should yield valuable information as to the differences between high-latitude and mid-latitude occurrence of GIC's, especially in terms of the solar wind-bound drivers.

A significant shortcoming in the results presented in Section 5.6 is the inability of the model to predict the highly variable component of the output parameters. In *wavelet analysis* techniques exist that would allow the signal represented by the output parameter measurements to be decomposed into "resolution levels". Having decomposed the original signal, an attempt to predict the various resolution levels *separately* could be made. Since only solar wind-based inputs were considered in this model, the inclusion of internal magnetospheric parameters may also help resolve this problem.

In this investigation only a 30-minute prediction lead time is considered. It would be possible to make use of various prediction lead times. For example, a 60-minute lead prediction could be made. Acknowledging that longer lead times would probably result in less accurate predictions, 45- and 30-minute predictions would subsequently be made. In this manner, longer, albeit less accurate, foresights would complement 30-minute predictions.

Although we believe the input parameter selection procedure to be objective, it might be worthwhile to commence the selection process with the z-component of the IMF. Since we expected magnetic reconnection to have an influence on the predicted parameters, the explicit selection of B_z^{imf} and the subsequently selected inputs would yield insight into the role of magnetic reconnection in the formation of temporal geomagnetic field variations at middle latitudes.

6.3 Possible Model Layout

This thesis concludes with a schematic diagram depicting the possible layout of a working prediction model, illustrated in Figure 6.1 below. It depicts how data from measurements made by ACE, ground-based magnetometer stations and GIC measurement sites could be processed and incorporated into a working prediction model.

Measurements of solar wind parameters are made by the ACE spacecraft and transmitted in real-time to NOAA ground stations. From NOAA the data is uploaded to the Goddard Space Flight Centre (GSFC, 2008) and NOAA Space Weather Prediction Centre (SWPC, 2003) websites.

Historic solar wind data downloaded from the GSFC (2008) website and geomagnetic field measurements obtained from magnetometer stations are collected and processed into

training and testing data sets. The training and testing sets are used to train the neural networks upon which the prediction algorithm is based. As more knowledge regarding the physical processes driving GIC's is gained (e.g. the suggestions mentioned in Section 6.2), more neural networks will be trained in order to incorporate the new information into the prediction model.

Solar wind data (say from K minutes in the past to the current time t) is taken in near real-time from the SWPC (2003) website, processed and used as input to the prediction algorithm. The prediction algorithm is used to predict the output parameters $\tau = 30$ minutes in advance, i.e. from $t - K + \tau$ to $t + \tau$, where τ is the prediction lead time. The predictions are regularly monitored and compared with the measurements made at magnetometer stations. After a GIC event the measured GIC magnitudes (obtained from Eskom substations) are compared to GIC values computed from predicted geomagnetic output parameters.

Obviously, this is only a prototype layout. As more time is spent developing such a model, other practical constraints will be encountered and (hopefully) dealt with.

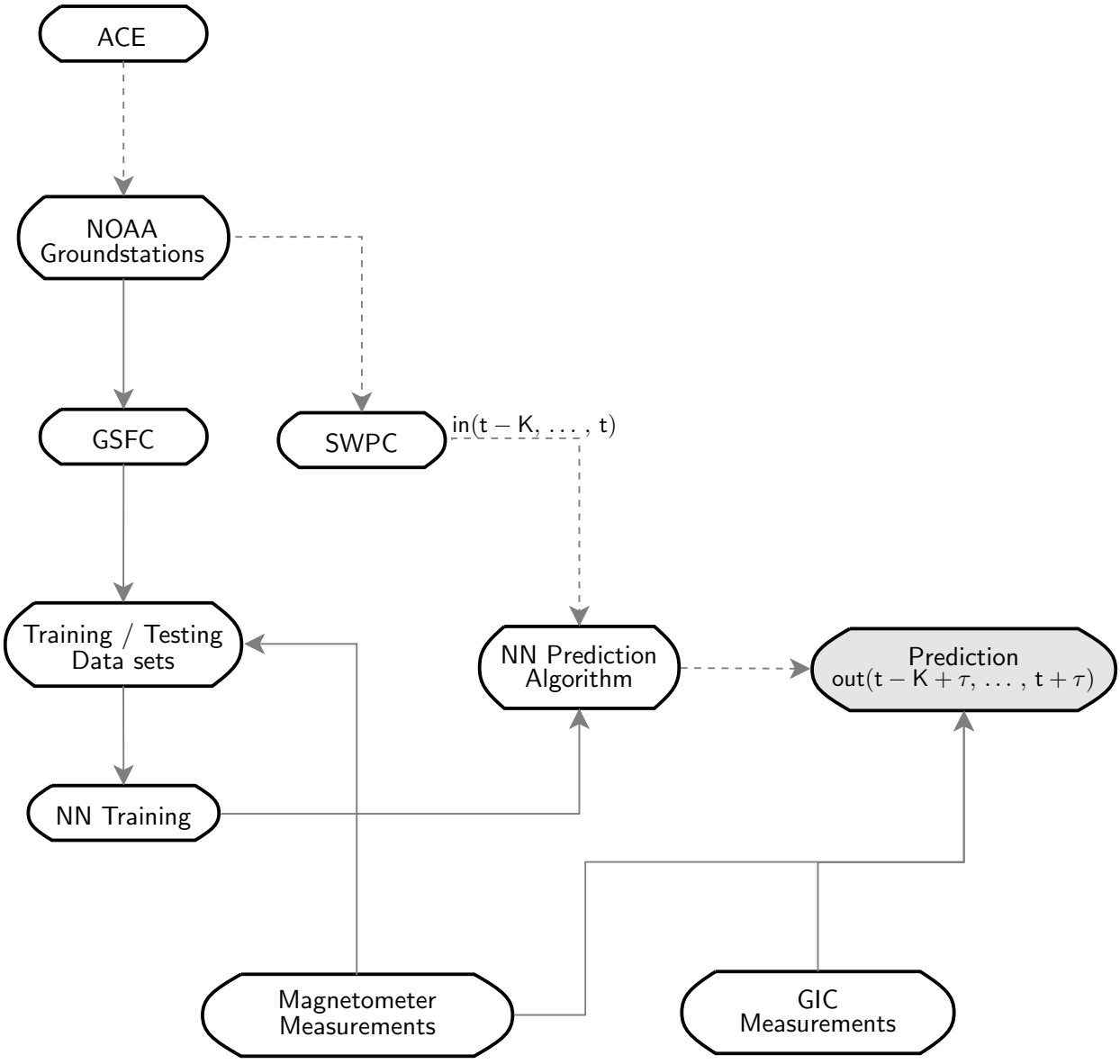

Figure 6.1: Schematic diagram depicting the possible layout of a working prediction model. Dashed lines denote flows that are meant to be executed in near real-time, as the timing of predictions directly depend on these steps. Input parameter data from K minutes in the past to the current time t is denoted by $\mathsf{in}(\mathsf{t}-\mathsf{K},\ldots,\mathsf{t})$, and $\mathsf{out}(\mathsf{t}-\mathsf{K}+\tau,\ldots,\mathsf{t}+\tau)$ denotes the predicted output, with a τ-minute prediction lead.

References

Bishop CM. *Neural Networks for Pattern Recognition*. Oxford University Press (1995).

Bolduc L. GIC observations and studies in the Hydro-Qébec power system. *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 64(16), pages 1793 – 1802 (2002).

De Silva CR, Ranganath S, and De Silva LC. Cloud basis function neural network: A modied RBF network architecture for holistic facial expression recognition. *Pattern recognition*, vol. 41, pages 1241 – 1253 (2008).

Elman JL. Finding structure in Time. *Cognitive Science*, vol. 14, pages 179 – 211 (1990).

Fausett L. *Fundamentals of Neural Networks: architectures, algoritms and applications*. Prentice Hall (1994).

Gosling JT. *Coronal mass ejections: An overview*, vol. 99, pages 9–16 (1997).

GSFC. Goddard Space Flight Center (Space Physics Data Facility) (2008).
URL: *http://spdf.gsfc.nasa.gov/*

Hagan MT, Demuth HB, and Beale M. *Neural Network Design*. PWS Publishing Company (1996).

Haykin S. *Neural Networks: A comprehensive foundation*. Macmillan College Publishing Company, Inc. (1994).

Hughes W. Magnetopause, Magnetotail, and Magnetic Reconnection. In MG Kivelson and CT Russel (editors), *Introduction to Space Physics*, pages 236 – 248. Cambridge University Press (1995).

Hundhausen A. The Solar Wind. In MG Kivelson and CT Russel (editors), *Introduction to Space Physics*, pages 91 – 128. Cambridge University Press (1995).

Kappenman JG. An overview of the impulsive geomagnetic field disturbances and power grid impacts associated with the violent Sun-Earth connection events of 29–31 October 2003 and a comparative evaluation with other contemporary storms. *Space Weather*, vol. 3 (2005).

Koen J. Geomagnetically induced currents in the Southern African electricity transmission network. Ph.D. thesis, University of Cape Town (2002).

Koskela T, Lehtokangas M, Saarinen J, and Kaski K. Time series prediction with Multilayer Perceptron, FIR and Elman neural networks. In *Proceedings of the World Congress on Neural Networks*, pages 491–496. INNS Press (1996).

Lanzerotti L. Space weather effects on technologies. In *Space Weather*, *Geophysical Monograph*, vol. 125. American Geophysical Union (2001).

Lundstedt H. A.I. techniques in geomagnetic storm forecasting. In BT Tsurutani, WD Gonzalez, Y Kamide, and JK Arballo (editors), *Magnetic Storms*, *Geophysical Monograph Series*, vol. 98, pages 243 – 252. American Geophysical Union (1997).

McClary DW, Syrotiuk VR, and Lecuire V. Adaptive audio streaming in mobile ad hoc networks using neural networks. *Ad Hoc Networks*, vol. 6, pages 524 – 538 (2008).

NSWP. National Space Weather Program: Strategic Plan. Electronic (.pdf) format (1995). **URL**: *http://www.ofcm.gov/nswp-sp/pdf/NSWP-SP-1995-scan.pdf* (accessed 2008/10/15).

Pirjola R. Geomagnetically induced currents in the Finnish 400 kV power transmission system. *Physics of the Earth and Planetary Interiors*, vol. 53, pages 214 – 220 (1989).

Pirjola R. Geomagnetic effects on ground-based technological systems. *Review of Radio Science*, (21), pages 473 – 496 (2002).

Pirjola R, Kauristie K, Lappalainen H, and Viljanen A. Space weather risk. In *Space Weather*, vol. 3. American Geophysical Union (2005).

Prölss G. *Physics of the Earth's Space Environment (an Introduction)*. Springer-Verlag (2004).

Pulkkinen A. Geomagnetic induction during highly disturbed space weather conditions: Studies of ground effects. Ph.D. thesis, University of Helsinki (2003).

Pulkkinen A, Pirjola R, and Viljanen A. Determination of ground conductivity and system parameters for optimal modeling of geomagnetically induced current flow in technological systems. *Earth, Planets and Space*, vol. 59(9), pages 999 – 1006 (2007a).

Pulkkinen TI, Palmroth M, Tanskanen EI, Ganushkina NY, Shukhtina MA, and Dmitrieva NP. Solar wind – magnetosphere coupling: A review of recent results. *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 69, pages 256 – 264 (2007b).

Purcell EM. *Electricity and Magnetism*, *Berkerley Physics course*, vol. 2. McGraw-Hill Book Company (1965).

Russel CT, Zhou XW, Chi PJ, Kawano H, Moore TE, Peterson WK, Cladis JB, and Singer HJ. Sudden compression of the outer magnetosphere associated with an ionospheric mass ejection. *Geophysical Research Letters*, vol. 26(15), pages 2343 – 2346 (1999).

Stone EC, Frandsen AM, Mewaldt RA, Christian ER, Margolies D, Ormes JF, and Snow F. The Advanced Composition Explorer. *Space Science Reviews*, vol. 86(1), pages 1 – 22 (1998).

SWPC. ACE Real Time solar wind main page (2003). Accessed 2008/06/10.
URL: *http://www.swpc.noaa.gov/ace/*

Vallado DA. *Fundamentals of Astrodynamics and Applications*. McGraw-Hill (1997).

Viljanen A and Pirjola R. Statistics on geomagnetically induced currents in the Finnish 400 kV power system based on recordings of geomagnetic variations. *Journal of Geomagnetism and Geoelectrics*, vol. 41, pages 411 – 420 (1989).

Viljanen A and Pirjola R. Geomagnetically induced currents in the Finnish high-voltage power system. *Surveys in Geophysics*, vol. 15(4), pages 383 – 408 (1994).

Weigel RS, Vassiliadis D, and Klimas AJ. Coupling of the solar wind to temporal fluctuations in ground magnetic fields. *Geophysical Research Letters*, vol. 29(19), pages 1915 – 1918 (2002).

Weigel RS, Vassiliadis D, and Klimas AJ. Solar wind coupling to and predictability of ground magnetic fields and their time derivatives. *Journal of Geophysical Research*, vol. 108(A7) (2003).

Wind. Wind spacecraft website (1999). Accessed 2008/06/18.
URL: *http://www-spof.gsfc.nasa.gov/istp/wind/*

Wintoft P. Study of the solar wind coupling to the time difference horizontal geomagnetic field. *Annales Geophysicae*, vol. 23, pages 1949–1957 (2005).

Wintoft P, Wik M, Lundstedt H, and Eliasson L. Predictions of local ground geomagnetic field fluctuations during the 7 – 10 November 2004 events studied with solar wind driven models. *Annales Geophysicae*, vol. 23, pages 3095 – 3101 (2005).

Wolf RA. Magnetospheric Configuration. In MG Kivelson and CT Russel (editors), *Introduction to Space Physics*, pages 288 – 329. Cambridge University Press (1995).

Zell A, Mamier G, et al. *Stuttgart Neural Network Simulator User Manual, Version 4.2*. University of Stuttgart, University of Tübingen (1998).
URL: *http://www.nada.kth.se/~orre/snns-manual/*.