



# Estimating the geoeffectiveness of halo CMEs from associated solar and IP parameters using neural networks

J. Uwamahoro<sup>1,2</sup>, L. A. McKinnell<sup>2,3</sup>, and J. B. Habarulema<sup>2</sup>

<sup>1</sup>Department of Mathematics and Physics, Kigali Institute of Education [KIE], P.O. Box 5039 – Kigali, Rwanda

<sup>2</sup>South African National Space Agency [SANSA], Space Science, 7200 Hermanus, South Africa

<sup>3</sup>Department of Physics and Electronics, Rhodes University, Grahamstown 6140, South Africa

Correspondence to: J. Uwamahoro (mahorojpacis@gmail.com)

Received: 11 July 2011 – Revised: 28 March 2012 – Accepted: 16 May 2012 – Published: 12 June 2012

**Abstract.** Estimating the geoeffectiveness of solar events is of significant importance for space weather modelling and prediction. This paper describes the development of a neural network-based model for estimating the probability occurrence of geomagnetic storms following halo coronal mass ejection (CME) and related interplanetary (IP) events. This model incorporates both solar and IP variable inputs that characterize geoeffective halo CMEs. Solar inputs include numeric values of the halo CME angular width (AW), the CME speed ( $V_{\text{cme}}$ ), and the comprehensive flare index (cfi), which represents the flaring activity associated with halo CMEs. IP parameters used as inputs are the numeric peak values of the solar wind speed ( $V_{\text{sw}}$ ) and the southward Z-component of the interplanetary magnetic field (IMF) or  $B_s$ . IP inputs were considered within a 5-day time window after a halo CME eruption. The neural network (NN) model training and testing data sets were constructed based on 1202 halo CMEs (both full and partial halo and their properties) observed between 1997 and 2006. The performance of the developed NN model was tested using a validation data set (not part of the training data set) covering the years 2000 and 2005. Under the condition of halo CME occurrence, this model could capture 100 % of the subsequent intense geomagnetic storms ( $\text{Dst} \leq -100 \text{ nT}$ ). For moderate storms ( $-100 < \text{Dst} \leq -50$ ), the model is successful up to 75 %. This model's estimate of the storm occurrence rate from halo CMEs is estimated at a probability of 86 %.

**Keywords.** Magnetospheric physics (Solar wind-magnetosphere interactions)

## 1 Introduction

Explosive events occurring on the Sun are the main causes of space weather affecting space- and ground-based technology as well as life on Earth in a number of ways (e.g. Siscoe and Schwenn, 2006). The predictability of space weather is therefore one way to minimize its effects. However, space weather prediction is still relatively inaccurate given that the underlying physics of the main drivers (e.g. CMEs and associated X-ray flares) is not yet sufficiently well understood (Schwenn et al., 2005).

Geomagnetic storms (GMS) represent typical features of space weather. They occur as a result of the energy transfer from the solar wind (SW) to the Earth's magnetosphere via magnetic reconnection. The main solar sources of GMS are (a) the CMEs from the Sun (Gopalswamy et al., 2007), and (b) the corotating interaction regions (CIRs) that result from the interaction between the fast and slow SW originating from coronal holes (Zhang et al., 2007). The two phenomena evolve into geoeffective conditions in the SW producing moderate to intense GMS when there is an enhanced and long lasting IMF in the southward direction (Richardson et al., 2002; Richardson, 2006; Gonzalez et al., 2004). However, despite the prominent role played by CMEs in producing GMS, their prediction cannot only be based on CME observations. As noted by Wang et al. (2002), the properties of CMEs that lead to magnetic storms are still a subject of intense research. Hence, improving the prediction of GMS requires an identification of key solar and IP geoeffective parameters of CMEs (Srivastava, 2005).

Currently, magnetic storm prediction models include statistical, empirical and physics-based methods. However,

despite previous attempted theoretical models to forecast the magnetic storm occurrence (Dryer, 1998; Dryer et al., 2004), physics-based models are still difficult to achieve. This is due to the complex, non-linear chaotic system of the solar-terrestrial interaction, with its physics still to be well understood (Fox and Murdin, 2001; Schwenn et al., 2005). Space weather forecasters often prefer empirical approaches based on observable data (Kim et al., 2010). Various functional relationships have been proposed for magnetic storm predictions. An algorithm for predicting the disturbance storm time (Dst) index from SW and the IMF parameters was first proposed by Burton et al. (1975). Empirical models for predicting GMS using CME-associated parameters at the Sun have been developed, including a recent work by Kim et al. (2010). Other authors prefer statistical methods, e.g. Srivastava (2005), who used a combination of solar and IP properties of geoeffective CMEs in a logistic regression model to predict the occurrence of intense GMS.

Empirical methods also include NN methods that are input-output models and have proven to be efficient in capturing the linear as well as the non-linear processes (Kamide et al., 1998). NN techniques have been described by various authors to be suitable for predicting transient solar-terrestrial phenomena (Lundstedt et al., 2005; Pallochchia et al., 2006; Woolley et al., 2010). A very well-designed and trained network can improve a theoretical model by performing generalization rather than simply curve fitting. By changing the NN input values, it is possible to investigate the functional relationship between the input and the output and therefore, be able to derive what the network has learned (Lundstedt, 1997). NN models for predicting magnetic storms using SW data as inputs have been developed (Lundstedt and Wintoft, 1994), with the ability to estimate the level of geomagnetic disturbances as measured by the Dst index. In particular, the use of Elman NN-based algorithms has achieved improved Dst forecasts (Lundstedt et al., 2002). In a NN-based model developed by Valach et al. (2009), geoeffective solar events such as solar X-ray flares (XRAs) and solar radio bursts (RSRs) were used to predict the subsequent GMS. In order to improve GMS forecasts, Dryer et al. (2004) suggested that models should include both solar and near-Earth conditions.

For this study, a combination of solar and IP properties of halo CMEs is used in a NN model to predict the probability of GMS occurrence following halo CMEs. Unlike the work by Srivastava (2005) that produced the intense and super-intense storm prediction model, the present NN model attempts to also explore the predictability of moderate storms ( $-100 \text{ nT} < \text{Dst} \leq -50 \text{ nT}$ ). Note that input parameters used are directly associated with halo CMEs, and therefore, the developed model cannot predict the probability occurrence of GMS that are non-CME-driven such as those caused by the CIRs. In developing the NN model described in this paper, a procedure was followed similar to the one used by McKinnell et al. (2010) for predicting the probability of spread-F occurrence over Brazil.

## 2 Data: determination of input and output parameters

### 2.1 Halo CMEs

The Solar and Heliospheric Observatory/Large Angle Spectrometric Coronagraph (SOHO/LASCO) (Bruckner et al., 1995) has been detecting the occurrence of CMEs on the Sun for more than a decade. Halo CMEs are those that appear to surround the occulting disk of the observing coronagraphs. It has been observed that halo CMEs originating from the visible solar disc and that are Earth-directed have the highest probability to impact the Earth's magnetosphere (Webb et al., 2000), and hence are useful for the prediction of GMS. In their study, Webb et al. (2000) and Cyr et al. (2000) used  $140^\circ$  and  $120^\circ$  respectively as a threshold apparent angular width (AW) to define halo CMEs, while a study by Wang et al. (2002) considered a halo CME as the one with an apparent AW greater than  $130^\circ$ . In this study, we considered halo CMEs as categorized by Gopalswamy et al. (2007), where full halo CMEs (F-type) have an apparent sky plane AW of  $360^\circ$ , while partial halos (P-type) are those with an apparent AW in the range  $120^\circ \leq W \leq 360^\circ$ .

During the first 11-year period of solar cycle (SC) 23 (from January 1996 to December 2006), the LASCO/SOHO catalog list indicates 393 full halo CMEs, representing 3.4 % of all 11683 CMEs recorded. During the same period, the number of partial halo CMEs was 840. Hence, in total, LASCO observed 1233 (10.5 %) halo CMEs. A correlation coefficient of 0.75 was found between full halo CMEs occurrence rate per year and the occurrence rate of geomagnetic disturbances (disturbed day frequency per year with  $\text{Dst} \leq -50 \text{ nT}$ ) from 1996 to 2006. However, not all halo CMEs are associated with GMS, and some non-halo CMEs can also cause intense GMS if they arrive at Earth with an enhanced southward component of the magnetic field with high speed (Gopalswamy et al., 2007). A number of GMS events have been identified without any link to frontside halo CMEs (Schwenn et al., 2005), and various studies, such as an analysis by Cane and Richardson (2003), have suggested that about half of the observed halo CMEs are not geoeffective. Indeed, both intense and moderate GMS can also be caused by CIRs resulting from the interaction between fast and slow SW in the IP medium (Richardson et al., 2006; Zhang et al., 2007). For the model developed in this study, we used halo CME (AW values of CMEs) data from the LASCO/SOHO catalog list (available online at: [http://cdaw.gsfc.nasa.gov/CME\\_list](http://cdaw.gsfc.nasa.gov/CME_list)).

### 2.2 Halo CME geoeffective properties: solar input parameters

In addition to the AW, the CME speed represents another important property of geoeffective CMEs. Halo CMEs have generally higher speed than the mean SW speed ( $470 \text{ km s}^{-1}$ ) and are useful parameters to predict the intensity of GMS (Srivastava, 2005). For this study, the

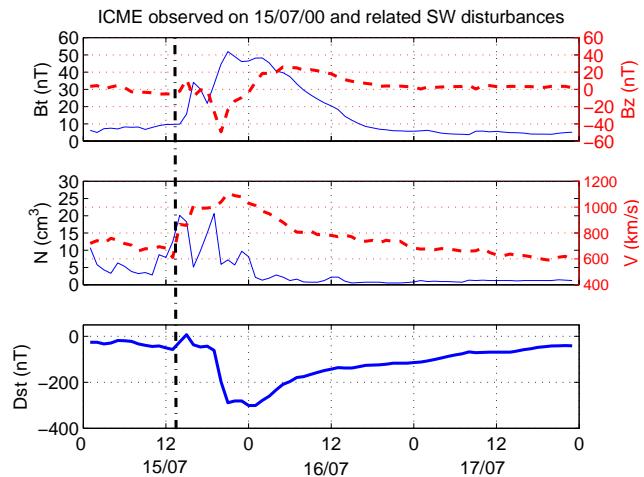
CME linear speed measured in the LASCO-C2 field of view has been used. Another solar input used is the cfi expressing the flare activity association with CMEs. In their analysis, Wang et al. (2002) found that geoeffective halo CMEs were mostly associated with flare activity. Furthermore, Srivastava and Venkatakrishnan (2004) observed that fast and full halo CMEs associated with large flares drive large geomagnetic disturbances. For our NN model, we used the cfi index as an input quantifying the halo CME association with solar flares. The minimum flare activity corresponds to 0 as a value of cfi, and the highest value of cfi (144) observed in SC 23 occurred during the *Halloween event* on 28 October 2003. The cfi data archive used is available on the website [ftp://www.ngdc.noaa.gov/STP/SOLAR\\_DATA/SOLAR\\_FLARES/FLARES\\_INDEX/Solar\\_Cycle/23/daily.plt](ftp://www.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FLARES/FLARES_INDEX/Solar_Cycle/23/daily.plt).

### 2.3 IP input parameters

In the IP medium, CMEs are manifested as shocks and interplanetary coronal mass ejection (ICME) structures, which couple to the magnetosphere to drive moderate to major storms (Webb, 2000; Echer et al., 2008). In situ observations of plasma and magnetic field properties are used to identify the arrival of ICMEs near Earth. Occurrence of shock waves and possible associated ICMEs can be characterized by a simultaneous increase of the SW speed, density, abnormal proton temperature as well as an increase in magnetic field magnitude. Plasma and magnetic field signatures indicating the presence of ICMEs are fully described in Cane and Richardson (2003) and Schwenn et al. (2005). As indicated by Gonzalez and Tsurutani (1987), the intensity of the storm following the passage of shock-ICME structures is well correlated with two parameters namely: (1) the IMF negative  $B_z$ -component ( $B_s$ ) and (2) the electric field convected by the SW,  $E_y = V B_s$ , where  $V$  is the SW velocity. Recent findings have also confirmed that the convective electric field has the best correlation with the Dst index (Echer et al., 2008).

For the NN model developed in this study, halo CMEs ( $AW \geq 120^\circ$ ), CME speed ( $V_{cme}$ ), cfi as well as IP peak values of negative  $B_z$  and SW speed ( $V_{sw}$ ) were used as NN numeric input (as shown in Table 1). The peak values ( $V_{sw}$ ,  $B_s$ ) correspond to the maxima recorded during the time period of ICME passage. SW data are provided by the OMNI-2 data set and available online (<http://www.nssdc.gsfc.nasa/omniweb.html>).

Shocks and ICME events that trigger SW geoeffective conditions are observed in situ by the Solar Wind Electron Proton Monitor (SWEPAM) and the Magnetic Field Experiment (MAG) instruments on board the Advanced Composite Explorer (ACE) spacecraft (Stone et al., 1998). The listing of ICMEs by Richardson and Cane (2008) and associated properties are available on the website <http://www.ssg.sr.unh.edu/mag/ace/ACElists/ICMEtable.html>.



**Fig. 1.** Plot showing the variation of the IMF total field  $B_t$ , the SW density  $N$  (solid lines), the  $B_z$  component of the IMF and the SW velocity  $V$  (dashed lines), following the passage of an ICME, observed by the WIND spacecraft on 15/16 July 2000. The vertical solid dashed line labels the shock ahead of the ICME. This ICME event has also been reported in Messerotti et al. (2009).

Figure 1 shows measured IP disturbances associated with the shock (and driver ICME) arrival at 1 AU on 15 July 2000, driving a storm on 16 July 2000 with peak minimum Dst reaching  $-301$  nT. This storm was driven by a very fast ( $1674\text{ km s}^{-1}$ ) full halo CME on 14 July at 10:54 UT and was associated with an  $X5.7$  flare (cfi = 59.13) originating at N22W07. In the IP medium,  $B_s$  reached a peak value of  $49.4$  nT, and  $1040\text{ km s}^{-1}$  was the maximum SW during the passage of ICME. Note that this event corresponds to the solar explosive event that triggered a radiation storm around Earth nicknamed the *Bastille event*.

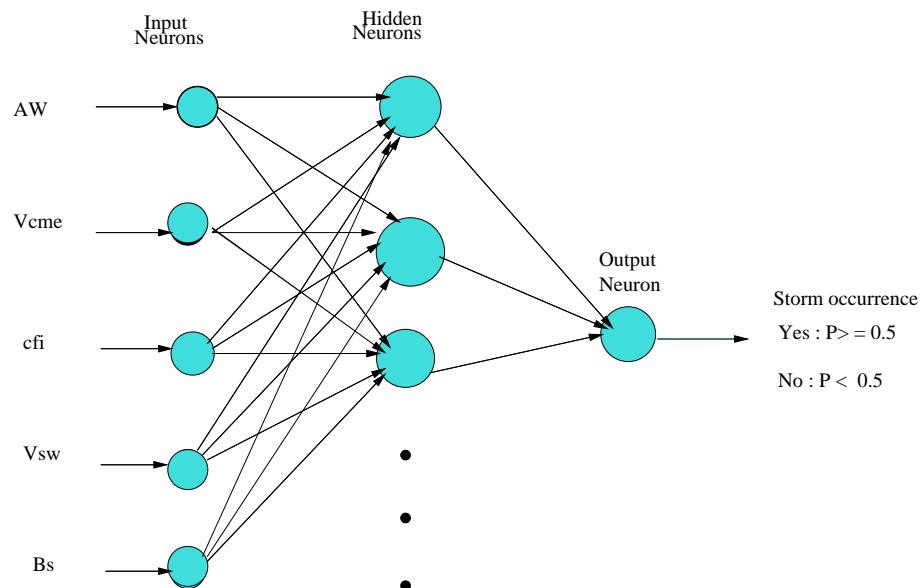
### 2.4 Geomagnetic response

There are various indices that indicate the level of geomagnetic disturbance. For this study, the disturbance storm time (Dst) was preferred since it is the widely used index for measuring the intensity of geomagnetic storms (Zhang et al., 2007). The Dst indicates the average change in the horizontal component of the Earth's magnetic field measured at four low latitude stations (see <http://swdcwww.kugi.kyoto-u.ac.jp/dstdir/dst2/onDstindex.html> for more details).

When the ICME structure in the IP medium presents an intensified southward component of the IMF ( $B_z$ ), it reconnects with the Earth's magnetic field. This magnetosphere–solar wind coupling induces the build-up of the ring current (Gonzalez et al., 1994; Gopalswamy, 2009), and therefore, the Dst index variation is a response to the build-up and decay of the ring current. Based on the minimum Dst values, Loewe and Prölss (1997) classify weak GMS ( $-30$  to  $-50$  nT), moderate ( $-50$  to  $-100$  nT), intense ( $-100$  to  $-200$  nT), severe ( $-200$  to  $-350$  nT) and great ( $< -350$  nT).

**Table 1.** Characteristics of the NN input and output parameters.

Model parameter type	Parameter name	Variable type	Measure	Value
Inputs	CME AW	Numeric	$\geq 120^\circ$	–
	CME speed	Numeric	Value in $\text{km s}^{-1}$	–
	cfi	Numeric	–	–
	$V_{\text{sw}}$	Numeric	Value in $\text{km s}^{-1}$	–
	$B_s$	Numeric	Value in nT	–
Outputs	No storm occurrence	Binary	Dst $> -50 \text{ nT}$	0
	Storm occurrence	Binary	Dst $\leq -50 \text{ nT}$	1

**Fig. 2.** A simplified illustration of the three-layered FFNN architecture as developed and used in this study.

For simplicity, in this analysis we followed the classification by Gopalswamy et al. (2007) to categorize two kinds of events: moderate storms ( $-100 \text{ nT} < \text{Dst} \leq -50 \text{ nT}$ ) and intense storms  $\text{Dst} \leq -100 \text{ nT}$ . As shown in Table 1, the storm occurrence (a row of NN outputs before training) is represented by binary values: 1 in the case of a moderate to intense storm occurrence ( $\text{Dst} \leq -50 \text{ nT}$ ) and 0 in the presence of a minor (or absence of a) storm ( $\text{Dst} > -50 \text{ nT}$ ). GMS events are defined here as storms periods with  $\text{Dst} \leq -50 \text{ nT}$ , which may last from a few hours to a couple of days.

### 3 Neural networks

In this work, NNs have been used as a tool in the development of a model to predict the probability of GMS occurrence from the observed solar and IP properties of halo CMEs. In summary, a NN is an assembly of interconnected computing elements called units or neurons. For the model developed in this work, we used a three-layered feed forward artificial NN. Feed forward neural networks (FFNN) repre-

sent the simplest and most popular type of NN, which has been widely used with success in the prediction of various solar-terrestrial time series (Lundstedt and Wintoft, 1994; Macpherson et al., 1995; Conway, 1998; Uwamahoro et al., 2009). In a FFNN arrangement, neurons (units) between layers are connected in a forward direction. Neurons in a given layer do not connect to each other and do not take inputs from subsequent layers. The input units, which are set to the previous values of the time series, send the signals to the hidden units. These hidden units process the received information and pass the results to the output units, which produce the final response to the input signals.

Figure 2 illustrates the three-layered NN architecture used in the work presented in this paper. In a three-layered FFNN with  $d$  input neurons, one hidden layer of  $M$  neurons and one output neuron, the output of the network, can be written in the form (Bishop, 1995):

$$y_k = g \left( \sum_{j=0}^M w_{kj} g \left( \sum_{i=0}^d w_{ji} x_i \right) \right), \quad (1)$$

where  $w_{ji}$  and  $w_{kj}$  represent the weights from the input to hidden layer and hidden to the output layer, respectively.  $M$  and  $d$  represent the number of hidden and input units, respectively, and  $x_i$  represents the input vectors used. The letter  $g$  represents the non-linear activation function. Activation functions are needed to introduce the non-linearity into the network. In this work, a logistic sigmoid activation function used for both hidden and output neurons is given by the relation

$$g(a) = \frac{1}{1 + \exp(-a)}. \quad (2)$$

The above function is a monotonically increasing function, which is defined for all real numbers. Such activation functions on the network outputs play an important role in allowing the outputs to be given a probabilistic interpretation (Bishop, 1995). Indeed, NNs provide an estimate of the posterior probabilities using the least squares optimization and are sensitive to sample size. A larger database provides better estimates (Richard and Lippmann, 1991; Hung et al., 1996). During the training process, inputs are shown to the net together with the corresponding known outputs. If there exists a relation between the input and the output, the net learns by adjusting the weights until an optimum set of weights that minimizes the network error is found and the network then converges.

Before training, the data set is generally split randomly into training and testing data sets in order to avoid the training results becoming biased towards a particular section of the database. For the NN trained while developing this model, data were split into 70 % for the training set and 30 % for the testing set. In order to determine how the NN has learned the behaviour in the input-output patterns, a validation data set consisting of the data not involved in the network training process was selected. Given that input variables have different numerical ranges (negatives values of  $B_z$ , small values of cfi, values of AW and SW speed in hundreds, CME speed in thousands), they were first normalized through weight initialization. Next, a suitable (optimal) learning parameter was selected by repeatedly trying different values. For the development and the training process of this model, we used the Stuttgart Neural Network Simulator (SNNS), developed by the Institute for Parallel and Distributed High Performance Systems, University of Tübingen, and the Wilhem-Schickard-Institute for Computer Science in Germany (<http://www.ra.cs.uni-tuebingen.de/SNNS/>). Details about the SNNS can be found in Zell et al. (1998).

### 3.1 NN model development: input/output data preparation

The first step in developing this model was to prepare the database for the NN training based on the criteria that halo CMEs were, or were not, followed by a storm ( $\text{Dst} \leq -50 \text{ nT}$  or  $\text{Dst} > -50 \text{ nT}$ ) within a 5-day window from the launch

**Table 2.** Determination of an optimum NN architecture over the validation data set. Optimized NN architectures are highlighted for 3 and 5 inputs, respectively.

Inputs	NN architecture	RMSE
3 input: AW,V,cfi	3:3:1	0.51261
	3:4:1	0.51377
	3:5:1	0.51471
	3:6:1	0.51553
5 input: AW,V,cfi $V_{\text{sw}}, B_s$	5:5:1	0.3225
	5:6:1	0.3396
	5:7:1	0.3366
	5:8:1	0.3376

of a halo CME. Halo CME (and associated solar and IP parameters) data covering the period from September 1997 to December 2006 were used in the model. This is the period corresponding to the availability of ICMEs and related shock structures (see the listing by Richardson and Cane, 2008). Note that there were missing CME data records for July, August and September 1998 as well as January 1999. In total, 1202 halo CMEs and associated geoeffective properties were included in the training, testing and validation of the NN model. The data covering 6 months in 2000 and 12 months in 2005 were set aside as the validation data set and were not used in the training. These unseen data provide an indication of the model's ability to predict the output in a general way.

Note that a positive response (code 1) was assigned as output for all inputs (described above) that were followed by GMS events within a 5-day window. Therefore, the number of input events that were associated with a positive response in the one column of output data set is actually larger than the total number of isolated GMS events (around 225 investigated, including about 90 intense storms (Echer et al., 2008)). The reason is that there were many cases where one isolated storm was common to more than one halo CME.

Table 3 shows 43 (five of the 48 listed had no halo CME background in the time window) halo CME driven storm events as well as their solar and IP characteristics as considered for the validation data set. Note that Table 3 is simplified and does not indicate many cases where more than one halo CME was the source of one geomagnetic storm. A good example is a storm of the 24 August 2005 with peak minimum Dst of  $-216 \text{ nT}$ . Although one full halo CME is indicated in Table 3 (event number 44) as the storm driver, there were actually two high speed ( $V > 1000 \text{ km s}^{-1}$ ) full halo CMEs that were probable sources of the storm. In fact, the two halo CMEs involved were all front-sided, associated with M-class solar flares and were followed by an ICME also observed on 24 August 2005. Therefore, for this particular example, there were two rows of input events made of AW = 360: with  $V_{\text{cme}}$  of  $1194 \text{ km s}^{-1}$  and  $2379 \text{ km s}^{-1}$ , respectively. The two

**Table 3.** Magnetic storm events and associated halo CME characteristics used for the validation data set. Only 43 of the 48 storm events were halo CME-driven. FH and PH in column 4 indicate full and partial halo CME respectively.

No. event	Date/time	Dst (min.) [nT]	Halo CMEs [FH or PH]	$V_{\text{cme}}$ [km s $^{-1}$ ]	X-Ray flare
1	08/06/00 – 19:00	-90	FH: 06/06 [15:54]	1119	X2.3
2	26/06/00 – 17:00	-76	PH: 25/06 [07:54]	1617	M1.9
3	16/07/00 – 00:00	-301	FH: 14/07 [10:54]	1674	X5.7
4	20/07/00 – 09:00	-93	–	–	–
5	23/07/00 – 22:00	-68	PH: 22/07 [11:54]	1230	M3.7
6	29/07/00 – 11:00	-71	FH: 25/07 [03:30]	528	M8.0
7	06/08/00 – 05:00	-56	PH: 03/08 [8:30]	896	C1.4
8	11/08/00 – 06:00	-106	PH: 08/08 [15:54]	867	C1.4
9	12/08/00 – 09:00	-235	FH: 09/08 [16:30]	702	C2.3
10	29/08/00 – 06:00	-60	PH: 25/08 [14:54]	518	M1.4
11	02/09/00 – 14:00	-57	PH: 01/09 [04:06]	603	C1.1
12	12/09/00 – 19:00	-73	PH: 09/09 [08:56]	554	M1.6
13	16/09/00 – 23:00	-68	FH: 12/09 [11:54]	1550	M1.0
14	18/09/00 – 23:00	-201	FH: 16/09 [05:18]	1215	M5.9
15	26/09/00 – 02:00	-55	FH: 25/09 [02:50]	587	M1.8
16	30/09/00 – 14:00	-76	PH: 27/09 [01:50]	820	C5.2
17	05/10/00 – 13:00	-182	FH: 02/10 [20:26]	569	C8.4
18	14/10/00 – 14:00	-107	PH: 11/10 [06:50]	799	C2.3
19	29/10/00 – 03:00	-127	FH: 25/10 [08:26]	770	C4.0
20	07/11/00 – 21:00	-159	FH: 03/11 [18:26]	291	C3.2
21	10/11/00 – 12:00	-96	FH: 08/11 [04:50]	474	–
22	29/11/00 – 13:00	-119	FH: 25/11 [01:31]	2519	M8.2
23	23/12/00 – 04:00	-62	PH: 20/12 [21:30]	609	C3.5
24	01/01/05 – 19:00	-57	FH: 30/12 [20:30]	832	B2.8
25	08/01/05 – 02:00	-96	FH: 05/01 [15:30]	735	–
26	12/01/05 – 10:00	-57	PH: 09/01 [09:06]	870	M2.4
27	18/01/05 – 08:00	-121	FH: 15/01 [06:30]	2049	M8.6
28	22/01/05 – 06:00	-105	FH: 19/01 [08:29]	2020	X1.3
29	07/02/05 – 21:00	-62	PH: 05/02 [13:31]	711	–
30	18/02/05 – 02:00	-86	FH: 17/02 [00:06]	1135	–
31	06/03/05 – 16:00	-65	–	–	–
32	05/04/05 – 05:00	-85	PH: 04/04 [11:06]	421	–
33	12/04/05 – 05:00	-70	PH: 09/04 [08:26]	329	B2.6
34	08/05/05 – 18:00	-127	FH: 05/05 [20:30]	1180	C7.8
35	15/05/05 – 08:00	-263	FH: 13/05 [17:12]	1689	M8.0
36	20/05/05 – 08:00	-103	PH: 17/05 [03:06]	449	M1.8
37	30/05/05 – 13:00	-138	FH: 26/05 [15:06]	586	B7.5
38	13/06/05 – 00:00	-106	PH: 08/06 [07:48]	179	–
39	15/06/05 – 12:00	-54	–	–	–
40	23/06/05 – 10:00	-97	–	–	–
41	09/07/05 – 18:00	-60	FH: 05/07 [15:30]	772	C1.3
42	10/07/05 – 20:00	-94	FH: 09/07 [22:30]	1540	M2.8
43	18/07/05 – 06:00	-76	FH: 14/07 [10:54]	2115	X1.2
44	24/08/05 – 11:00	-216	FH: 22/08 [17:30]	2378	M5.6
45	31/08/05 – 19:00	-131	FH: 29/08 [10:54]	1600	–
46	11/09/05 – 09:00	-147	FH: 09/09 [19:48]	2693	X6.2
47	15/09/05 – 16:00	-86	FH: 13/09 [20:00]	1866	X1.5
48	31/10/05 – 19:00	-75	–	–	–

halo CMEs occurred on the same day, and therefore had the same value of  $c_{fi}$ , which was 10.31. The last two NN inputs are in situ measured peak values of  $B_s$  and  $V_{sw}$  (38.3 nT and

710 km s $^{-1}$ , respectively) identified in a window of at least five days after halo CME occurrence.

The outputs corresponding to the two input events described above were represented by a binary value of 1,

because there was storm occurrence ( $\text{Dst} \leq -50 \text{ nT}$ ). In cases where halo CMEs were not followed by a storm ( $\text{Dst} > -50 \text{ nT}$ ), the corresponding output events were represented by a binary value of 0. We notice that the output of the NN model after training is a numerical value ranging between 0 and 1. The input and output parameters are shown in Table 1 and Fig. 2. Therefore, the model developed behaves like a function that estimates the probability of storm occurrence and can be written as

$$P = f(\text{AW}_{\text{cme}}, V_{\text{cme}}, \text{cfi}, B_s, V_{\text{sw}}). \quad (3)$$

For this NN model, we followed the example as in Srivastava (2005) and considered 0.5 as a threshold value (probability) for determining the prediction output classification. Therefore, any prediction output with value  $\geq 0.5$  was considered likelihood of occurrence of a storm event following a halo CME eruption.

### 3.2 NN optimization

The network was repeatedly trained by changing the number of iterations and by systematically varying the number of nodes in the hidden layer. During the training process, the mean square error variation of the testing pattern was monitored in order to stop the training at the right time and avoid overtraining. The best NN architecture was obtained by considering the minimum root mean square error (RMSE) value computed over the entire validation set:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (P_{\text{obs}} - P_{\text{pred}})^2}, \quad (4)$$

where  $P_{\text{obs}}$  (e.g. 0 or 1) and  $P_{\text{pred}}$  represent the observed and predicted probability values respectively and  $N$  represents the number of data points in the validation data set. The validation data set was made of 267  $P_{\text{obs}}$ , including 118 ones and 149 zeros. Note that the same optimisation criteria were used to determine a suitable input space for the model (as shown in Table 2). An optimized NN architecture was reached after 400 iterations using 0.005 as the learning rate. In order to be able to evaluate the prediction performance of the model on the training data set, we also computed the least RMSE over the training data set, which was 0.3422 (not shown in Table 2).

## 4 Results and discussion

The optimum network architecture was found to be that with 5 inputs (i.e. Eq. 3) using 5 hidden nodes (configuration: 5:5:1). The network with only three solar input parameters was found to perform poorly when tested on the validation data set (as shown in Table 2). This indicates the importance of considering IP parameters ( $B_s$  and  $V_{\text{sw}}$ ) for improving the prediction performance of the model.

Figure 3a–d is just an example to illustrate the model estimate of the probability, by which halo CMEs might be followed by a storm. The x-axes indicate days in a month for which there were one or multiple halo CMEs, and the y-axes indicate the predicted value expressing the probability of halo CMEs to be geoeffective. All the predicted values above 0.5 indicate a correct prediction of GMS occurrence (following a halo CME). The maximum in each of the four cases presented in Fig. 3 indicates the probability by which a particular storm is predicted by the model. Note that not all halo CMEs are indicated on the plots for representation purposes due to the fact that some dates had multiple halo CMEs, while no halo CMEs were observed for other days. Here, four typical examples are described, for all of which the model demonstrates a very high probability ( $P \simeq 1$ ) of storm occurrence.

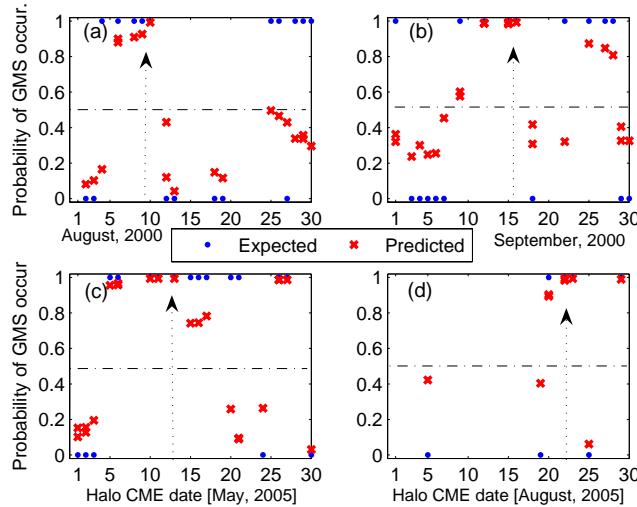
There were two intense storms that occurred on the 11 and 12 August 2000 reaching the Dst peak minima of  $-106 \text{ nT}$  and  $-235 \text{ nT}$ , respectively. As shown in Fig. 3a, the two storms were correctly predicted with more than 0.95 probability and were expected from the three halo CMEs that occurred on 8, 9 and 10 August 2000, respectively. The most probable cause of the 12 August 2000 storm ( $-235 \text{ nT}$ ) was a full halo CME which occurred on 9 August 2000 (see the arrow in the plot), indicated as number 9 in Table 3. Like in many observed moderate storm cases, this model fails to correctly predict the storm of 29 August 2000 ( $\text{Dst} = -60 \text{ nT}$ ) expected from a series of partial halo CMEs that occurred on 25–28 August 2000.

The example in Fig. 3b shows how the model correctly predicts the two storms of 16 and 18 September 2000, respectively. The 16 September 2000 moderate storm ( $\text{Dst} = -68 \text{ nT}$ ) was expected from the fast and full halo CME of 12 September 2000. A strong storm that occurred on 18 September 2000 (with Dst peak minimum of  $-201 \text{ nT}$ ) is correctly predicted by the model and was expected from two halo CMEs (shown by an arrow in Fig. 3b), which occurred on 15 and 16 September 2000, respectively. However, it is most likely that this storm was caused by the very fast full halo CME of 16 September (see event number 14, Table 3) or its interaction with the partial halo CME of 15 September 2000. Prior to the 18 September 2000 magnetic storm, there was an ICME first observed on 17 September at 21:00 UT.

Figure 3c shows that the NN model correctly predicts the 15 May 2005 great magnetic storm (with Dst peak minimum of  $-263 \text{ nT}$ ). This storm was expected from a very fast and powerful flare-associated full halo CME of 13 May 2005 (as shown by the arrow in Fig. 3c and row number 35 in Table 3). Similarly, Fig. 3d illustrates clearly the correct prediction of the 24 August 2005 strong magnetic storm (with a Dst peak minimum of  $-216 \text{ nT}$ ). This storm was expected from two fast full halo CMEs, which occurred on 22 August 2005. The very fast halo CME ( $V = 2378 \text{ km s}^{-1}$ ), which might have been the most probable cause of the storm, is number 44 in

**Table 4.** Prediction performance of the NN model on both the training and validation data sets.

Data set	Storm category	Observed	Correct predictions	False alarms
Training	Intense storms	53	51 [96 %]	32
	Moderate storms	59	42 [71 %]	
	Total	112	93 [83 %]	
Validation	Intense storms	19	19 [100 %]	8
	Moderate storms	24	18 [75 %]	
	Total	43	37 [86 %]	



**Fig. 3.** Illustration of the prediction performance of the model on some storms in the validation data set. All predictions above a 0.5 probability value indicate a successful prediction (of storm occurrence), where the predicted value can be interpreted as the probability by which a particular halo CME may be followed by a storm. Example events described in the text are indicated by the arrows.

Table 3, and its predicted geoeffectiveness is represented by the arrow in Fig. 3d.

The average value of correct predictions (observed storm responses and predicted with  $P \geq 0.5$ ) calculated over the whole validation data set was found to be 0.87. This can be considered as the NN model approximated probability, by which a storm occurrence can be predicted as a result of a halo CME event. The performance of the developed NN model was tested on 43 CME-driven GMS (listed in Table 3) by calculating the percentage of correctly predicted storms for both the validation and training data sets:

$$\frac{PE}{OE} \times 100 \quad (5)$$

where  $PE$  is the number of correctly predicted GMS and  $OE$  is the total number of observed GMS. Table 2 shows the RMSE values computed over the validation set, indicating that the model produces a lower estimate of storms occurrence when only three solar inputs were used. With three

solar inputs, the NN model could estimate 52 % and 41.6 % of the observed intense and moderate storms, respectively. The results presented in Table 4 summarise the prediction performance of the NN model using 5 inputs (solar and IP combined), tested on both the training and the validation data sets. As indicated in Table 4, the NN model predicts 100 % of intense storms and 75 % of moderate storms. The overall NN model prediction ability of GMS ( $Dst < -50$  nT) based on the observed halo CME was estimated at 86 %. The number of GMS predicted by the NN model, but not observed (false alarms), is also indicated in Table 4 for both the validation (8 events) and the training (32 events) data sets.

The results obtained demonstrate the ability of the NN model to produce a good estimate of the probability occurrence of intense storms compared to moderate storms. This difference in performance is related to the characteristics of inputs. Observations of the data indicate that intense storms are generally preceded by full halo CMEs ( $AW = 360^\circ$ ), high values of CME speed and  $c_{fi}$  as well as high peak values of  $B_s$  and  $V_{sw}$  compared to those associated with moderate GMS. On the other hand, previous studies have indicated that partial halo CMEs produce mostly moderate storms and the majority of them are less energetic (have lower speed). Note that moderate storms are often driven by the non-halo CMEs or CIRs that have not been considered in this study (also discussed in the earlier section).

We would like to emphasize that the results presented in this study only serve as an indication that solar and IP parameter characteristics of geoeffective halo CMEs can be used in a NN to estimate the probability occurrence of the subsequent GMS. The estimated geoeffectiveness of solar events (halo CMEs in this case) can be compared to other predictions from various analyses. Valach et al. (2009) used a combination of X-ray flares (XRAs) and solar radio burst (RSPs) as input to the NN model and obtained a 48 % successful forecast for severe geomagnetic response. The NN model described in this paper shows an improved performance with an accuracy of 86 % in the prediction of GMS. On the other hand, this compares favourably to the 77.7 % obtained by Srivastava (2005) using the logistic regression model. The prediction performance of the NN model described in this paper is unique, as it also attempted to estimate the probability

occurrence of moderate storms, which have not been considered in previous studies.

## 5 Summary

Predicting the occurrence of GMS on the basis of CME observations only is challenging and can sometimes lead to false alarms. In this study, a combination of solar and IP parameters has been used as inputs in a NN model with ability to estimate the probability occurrence of GMS resulting from halo CMEs. The results obtained indicate that the model performs well in estimating the occurrence of intense GMS as compared to moderate storms. In addition, this study shows that IP input parameters characterizing geoeffective halo CMEs and related ICME structures (i.e. increased peak values of  $B_s$  and  $V_{sw}$ ) contribute significantly in improving the predictability of GMS occurrence, confirming what is already known about the SW control of GMS phenomena. It was observed that the use of solar inputs only leads to a less accurate performance. However, such a model with only solar inputs is very useful for space weather, as the model provides a long warning time (1 to 4 days) compared to the NN model combining solar and IP inputs. The NN model described in this paper will contribute towards improving real-time space weather predictions. Locally, the model developed will be applied by the SANSA Space Weather Regional Warning Center (RWC) to improve various space weather models that involve consideration of storm conditions.

**Acknowledgements.** We acknowledge the authors of the LASCO/SOHO catalog list of CMEs, available online at [http://cdaw.gsfc.nasa.gov/CME\\_list](http://cdaw.gsfc.nasa.gov/CME_list) that was used for our study. We would also like to thank the National Geophysical Data Center (USA) for making available geomagnetic and solar data used via the following websites: <ftp://ftp.ngdc.noaa.gov/STP/GEOMAG/dst.html>, and [ftp://ftp.ngdc.noaa.gov/STP/SOLAR\\_DATA/SOLAR\\_FLARES/FLARES\\_INDEX](ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FLARES/FLARES_INDEX). The authors thank the SWEPAM and MAG teams for making available the listing of shocks and ICMEs used in this study.

This work was carried out with financial support from the National Astrophysics and Space Science Program (NASSP) and logistic support from the South African National Space Agency.

Topical Editor R. Nakamura thanks two anonymous referees for their help in evaluating this paper.

## References

- Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press Inc., New York, USA, 1995.
- Bruckner, G. E., Howard, R. A., Koomen, M. J., Korendyke, C. M., Michels, D. J., Moses, J. D., Socker, D. G., Dere, K. P., Lamy, P. L., Lleberia, A., Bout, M. V., Schwenn, R., Simnett, G. M., Bedford, D. K., and Eyles, C. J.: The Large Angle Spectroscopic Coronagraph, (LASCO), Solar Phys., 162, 357–402, 1995.
- Burton, R. K., McPherron, R. L., and Russell, C. T.: An empirical relationship between interplanetary conditions and Dst, J. Geophys. Res., 80, 4204–4214, 1975.
- Cane, H. V. and Richardson, I. G.: Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002, J. Geophys. Res., 108, A4, doi:10.1029/2002JA009817, 2003.
- Conway, A. J.: Time series, neural networks and the future of the Sun, New Astronomy Reviews, 42, 343–394, 1998.
- Cyr, O. C. S., Howard, R. A., Sheeley, N. R., Plunkett, S. P., Michels, D. J., Paswaters, S. E., Koomen, M. J., and Simnett, G. M.: Properties of coronal mass ejections: SOHO LASCO observations from January 1996 to June 1998, J. Geophys. Res., 105, 169–185, 2000.
- Dryer, M.: Multidimensional, magnetohydrodynamics simulation of solar-generated disturbances: Space weather forecasting of geomagnetic storms, AIAA Journal, 36, 365–370, 1998.
- Dryer, M., Smith, Z., Fry, C. D., Sun, W., Deehr, C. S., and Akasofu, S. I.: Real-Time shock arrival predictions during the Halloween 2003 epoch, Space Weather, 2, S09001, doi:10.1029/2004SW000087, 2004.
- Echer, E., Gonzalez, W. D., Tsurutani, B. T., and Gonzalez, A. L. C.: Interplanetary conditions causing intense geomagnetic storms ( $Dst \leq -100$  nT) during solar cycle 23 (1996–2006), J. Geophys. Res., 113, A05221, doi:10.1029/2007JA012744, 2008.
- Fox, N. and Murdin, P.: Solar-Terrestrial Connection: Space Weather Predictions, Encyclopedia, Astronomy and Astrophysics, 2416, doi:10.1888/0333750888/2416, 2001.
- Gonzalez, W. D. and Tsurutani, B. T.: Criteria of interplanetary parameters causing intense magnetic storms ( $Dst < -100$  nT), Planet Space Sci., 35, 1101–1109, 1987.
- Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Tsurutani, B. T., Vasiliunas, V. M., and Rostoker, G.: What is a Geomagnetic storm?, J. Geophys. Res., 99, 5771–5792, 1994.
- Gonzalez, W. D., Lago, A. D., de Gonzalez, A. L. C., Vieira, L. E. A., and Tsurutani, B. T.: Prediction of peak-Dst from halo CME/magnetic cloud-speed observations, J. Atmos. Sol. Terr. Phys., 66, 161–165, 2004.
- Gopalswamy, N.: Coronal mass ejection and space weather, in: Climate and Weather of the Sun-Earth System (CAWSES): Selected papers from the 2007 Kyoto Symposium, edited by: Tsuda, T., Fujii, R., Shibata, K., and Geller, M. A., pp. 77–120, ©TERRAPUB, TOKYO, 2009.
- Gopalswamy, N., Yashiro, S., and Akiyama, S.: Geoeffectiveness of halo coronal mass ejections, J. Geophys. Res., 112, A06112, doi:10.1029/2006JA012149, 2007.
- Hung, M. S., Hu, M. Y., Shanker, M. S., and Patuwo, B. E.: Estimating Posterior Probabilities In Classification Problems with Neural Networks, International Journal of Computational Intelligence and Organizations, 1, 49–60, 1996.
- Kamide, Y., Baumjohann, W., Gonzalez, W. D., Grande, M., Joselyn, J. A., McPheron, R. L., Phillips, J. L., Reeves, E. G. D., Rostoker, G., Sharma, A. S., Singer, H. J., and Vasiliunas, V. M. B. T. T.: Current understanding of magnetic storms: Storm-substorm relationships, J. Geophys. Res., 103, 705–728, 1998.
- Kim, R. S., Chao, K. S., Moon, Y. J., Dryer, M., Lee, J., Yi, Y., Kim, K. H., Wang, H., Park, Y. D., and Kim, Y. H.: An empirical model for prediction of geomagnetic storms using initially observed CME parameters at the Sun, J. Geophys. Res., 115, A12108, doi:10.1029/2010JA015322, 2010.

- Loewe, C. A. and Prölss, G. W.: Classification and mean behaviour of magnetic storms, *J. Geophys. Res.*, 102, 14209–14213, 1997.
- Lundstedt, H.: AI Techniques in Geomagnetic Storm Forecasting, in: *Magnetic Storms*, edited by: Tsurutani, B. T., Gonzalez, W. D., Kamide, Y., and Arballo, J. K., vol. 98 of *Geophys. Monograph series*, pp. 243–252, AGU, Washington D.C., 1997.
- Lundstedt, H. and Wintoft, P.: Prediction of geomagnetic storms from solar wind data with the use of a neural network, *Ann. Geophys.*, 12, 19–24, doi:10.1007/s00585-994-0019-2, 1994.
- Lundstedt, H., Gleisner, H., and Wintoft, P.: Operational forecasts of the geomagnetic Dst index, *Geophys. Res. Lett.*, 29, 2181, doi:10.1029/2002GL016151, 2002.
- Lundstedt, H., Liszka, L., and Lundin, R.: Solar activity explored with new wavelet methods, *Ann. Geophys.*, 23, 1505–1511, doi:10.5194/angeo-23-1505-2005, 2005.
- Macpherson, K. P., Conway, A. J., and Brown, J. C.: Prediction of solar and geomagnetic activity data using neural networks, *J. Geophys. Res.*, 100, 735–744, 1995.
- McKinnell, L. A., Paradza, M. W., Cilliers, P. J., Abdu, M. A., and de Souza, J. R.: Predicting the probability occurrence of spread-F over Brazil using neural networks, *Adv. Space Res.*, 46, 1047–1054, 2010.
- Messerotti, M., Zuccarello, F., Guglielmo, S. L., Bothmer, V., Lilenstein, J., Noci, G., Storini, M., and Lundstedt, H.: Solar Weather Event Modelling and prediction, *Space Sci. Rev.*, 147, 121–185, 2009.
- Pallocchia, G., Amata, E., Consolini, G., Marcucci, M. F., and Bertello, I.: Geomagnetic  $D_{st}$  index forecast based on IMF data only, *Ann. Geophys.*, 24, 989–999, doi:10.5194/angeo-24-989-2006, 2006.
- Richard, M. D. and Lippmann, R. P.: Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities, *Neural Computation*, 3, 461–483, 1991.
- Richardson, I. G.: Major geomagnetic storms ( $Dst \leq -100$  nT) generated by corotating interactive regions, *J. Geophys. Res.*, 111, A07S09, doi:10.1029/2005JA011476, 2006.
- Richardson, I. and Cane, H.: Near-Earth Interplanetary Coronal Mass Ejections in 1996–2007, Tech. rep., Astrophysics Physics Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA, 2008.
- Richardson, I. G., Cane, H. V., and Cliver, E. W.: Sources of geomagnetic activity during nearly three solar cycles (1972–2000), *J. Geophys. Res.*, 107, A8, doi:10.1029/2001JA000504, 2002.
- Richardson, I. G., Webb, D. F., Zhang, J., Berdichersky, D. B., Biesecker, D. A., Kasper, J. C., Kataoka, R., Steinberg, J. T., Thompson, B. J., Wu, C.-C., and Zhukov, A. N.: Major geomagnetic storms ( $Dst \leq -100$  nT) generated by corotating interactive regions, *J. Geophys. Res.*, 111, A07S09, 2006.
- Schwenn, R., Dal Lago, A., Huttunen, E., and Gonzalez, W. D.: The association of coronal mass ejections with their effects near the Earth, *Ann. Geophys.*, 23, 1033–1059, doi:10.5194/angeo-23-1033-2005, 2005.
- Siscoe, G. and Schwenn, R.: CME disturbance forecasting, *Space Sc. Rev.*, 123, 453–470, 2006.
- Srivastava, N.: A logistic regression model for predicting the occurrence of intense geomagnetic storms, *Ann. Geophys.*, 23, 2969–2974, doi:10.5194/angeo-23-2969-2005, 2005.
- Srivastava, N. and Venkatakrishnan, P.: Solar and interplanetary sources of geomagnetic storms during 1996–2002, *J. Geophys. Res.*, 109, A10103, doi:10.1029/2003JA010175, 2004.
- Stone, E. C., Frandsen, A. M., Mewaldt, R. A., Christian, E. R., Margolies, D., Ormes, J. F., and Snow, F.: The Advanced Composition Explorer, *Space Sci. Rev.*, 86, 357–408, 1998.
- Uwamahoro, J., McKinnell, L. A., and Cilliers, P. J.: Forecasting solar cycle 24 using neural networks, *J. Atmos. Solar-Terr. Phys.*, 71, 569–574, 2009.
- Valach, F., Revallo, M., Bochnicek, J., and Hejda, P.: Solar energetic particle flux enhancement as a predictor of geomagnetic activity in a neural network-based model, *Space Weather*, 7, S04004, doi:10.1029/2008SW000421, 2009.
- Wang, Y. M., Ye, P. Z., Wang, S., Zhou, G. P., and Wang, J. X.: A statistical study on the geoeffectiveness of the Earth-directed coronal mass ejections from March 1997 to December 2000, *J. Geophys. Res.*, 107, A11, doi:10.1029/2002JA009244, 2002.
- Webb, D. F.: Coronal Mass Ejections: Origins, Evolution, and Role in Space Weather, *IEEE Trans. plasma sc.*, 28, 1795–1806, 2000.
- Webb, D. F., Cliver, E. W., Crooker, N. U., st Cyr, O. C., and Thompson, B. J.: Relationship of halo coronal mass ejections, magnetic clouds and magnetic storms, *J. Geophys. Res.*, 105, 7491–7508, 2000.
- Woolley, J. W., Agarwal, P. K., and Baker, J.: Modeling and Prediction of Chaotic Systems with artificial neural networks, *International Journal for Numerical Methods in Fluids*, 63, 2117, doi:10.1002/fld.2117, 2010.
- Zell, A., Mamier, G. M., Vogt, M., Mach, N., Hübner, R., Döring, S., Herrmann, K. U., Soyez, T., Schmalzl, M., Sommer, T., Hatzi Georgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., Wieland, J., and Gatter, J.: Stuttgart Neural Network Simulator (SNNS), User Manual, version 4.2, Universities of Stuttgart and Tübingen, Germany and the European Particle Research Lab, CERN, Geneva, Switzerland, 1998.
- Zhang, J., Richardson, I. G., Webb, D. F., Gopalswamy, N., Huttunen, E., Kasper, J. C., Nitta, N. V., Poomvises, W., Thompson, B. J., Wu, C. C., Yashiro, S., and Zhukov, A. N.: Solar and interplanetary sources of major geomagnetic storms ( $Dst \leq -50$  nT) during 1996–2005, *J. Geophys. Res.*, 112, A10102, doi:10.1029/2007JA012321, 2007.