

A study on precursors leading to geomagnetic storms using artificial neural network

GAURAV SINGH and A K SINGH*

Atmospheric Research Laboratory, Department of Physics, Banaras Hindu University, Varanasi 221 005, India.

**Corresponding author. e-mail: abhay_s@rediffmail.com*

Space weather prediction involves advance forecasting of the magnitude and onset time of major geomagnetic storms on Earth. In this paper, we discuss the development of an artificial neural network-based model to study the precursor leading to intense and moderate geomagnetic storms, following halo coronal mass ejection (CME) and related interplanetary (IP) events. IP inputs were considered within a 5-day time window after the commencement of storm. The artificial neural network (ANN) model training, testing and validation datasets were constructed based on 110 halo CMEs (both full and partial halo and their properties) observed during the ascending phase of the 24th solar cycle between 2009 and 2014. The geomagnetic storm occurrence rate from halo CMEs is estimated at a probability of 79%, by this model.

1. Introduction

Space weather prediction involves forecasting of the magnitude and time of the commencement of a geomagnetic storm, based on solar and interplanetary observations (Srivastava 2005). During a geomagnetic storm, severe changes occur both in interplanetary (IP) space and the terrestrial environment such as acceleration of charged particles and enhancement of electric currents, auroras, and magnetic field variations on the Earth's surface, which can endanger human life or health (Siscoe and Schwenn 2006). Geomagnetic storms (GMS) represent typical features of space weather. They occur as a result of energy transfer from the solar wind (SW) to Earth's magnetosphere via magnetic reconnection. However, despite the prominent role played by CMEs in producing GMS, their prediction cannot be based only on CME observations (Uwamahoro *et al.* 2012). To achieve this goal, it is important to develop a prediction scheme based on both solar and IP properties of geo-effective CMEs (Srivastava 2005).

Artificial neural network (ANN) techniques have been described by various authors to be suitable for predicting transient solar-terrestrial phenomena (Lundstedt *et al.* 2005; Palocchia *et al.* 2006; Woolley *et al.* 2010). A very well-designed and trained network can improve a theoretical model by performing generalizations rather than simply curve fitting. By changing the ANN 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). ANN 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. The model developed by Lundstedt *et al.* (2002) consists of a recurrent neural network that requires hourly averages of the solar wind magnetic field component B_z , particle density n , and velocity V as inputs and predicts the *Dst* index in almost real-time (Srivastava 2005). In order to improve GMS forecasts, Dryer *et al.* (2004) suggested that models

Keywords. Space weather; coronal mass ejections; geomagnetic storm; artificial neural network.

should include both solar and near-Earth conditions. Recently, Uwamahoro *et al.* (2012) estimated the geo-effectiveness of halo CMEs from associated solar and IP parameters using neural networks. They presented an improved performance with an accuracy of 86% in the prediction of geomagnetic storm occurrence. However, for moderate storms ($-100 < Dst \leq -50$), the model is successful up to 75% only.

For the present study, a combination of solar and IP properties of halo CMEs is used in an ANN model to predict the probability of GMS occurrence following halo CMEs observed during the ascending phase of the 24th solar cycle between 2009 and 2014. Out of a total of 110 geomagnetic storms that occurred during the study period, 21 were intense storms and 89 were moderate. The results obtained by previous studies show the ability of the ANN model to produce a good estimate of the probability occurrence of only intense storms compared to moderate storms. However, the present model focuses mainly on the probability occurrence of moderate geomagnetic storms with an improved performance of about 80%.

2. Observational inputs for space weather prediction

2.1 Key solar parameters

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 (Uwamahoro *et al.* 2012). 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 this study, we considered halo CMEs as categorized by Gopalswamy *et al.* (2007), where full halo CMEs (F-type) have an apparent sky plane angular width (AW) of 360° , while partial halos (P-type) are those with an apparent AW in the range $120^\circ \leq W \leq 360^\circ$ as suggested by Uwamahoro *et al.* (2012). The angular width of a CME is a measure of the volume in the corona that is 'blown out' (Robbrecht *et al.* 2009).

During the ascending phase of the 24th solar cycle (January 2009–December 2014), the LASCO/SOHO catalogue list indicates 110 halo CMEs (both partial and full halo CMEs) which are geo-effective. Table 1 lists 110 halo CMEs ($Dst < -50$ nT) out of which 21 were intense storms and

89 were moderate storms. In addition to the AW, the CME speed represents another important property of geo-effective CMEs. Halo CMEs generally have a higher speed than the mean SW speed (470 km s^{-1}) and are useful parameters to predict the intensity of GMS (Srivastava 2005). For the model developed in this study, we used halo CME (AW and SW speed values of CMEs) data from the LASCO/SOHO catalogue list (available online at: <http://cdaw.gsfc.nasa.gov/CMEList>).

Another solar input used is the flare parameter expressing the flare activity association with CMEs. In their analysis, Wang *et al.* (2002) found that geo-effective 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 ANN model, we used the logarithmic value of the peak flux of the most significant flare that has occurred during the CME eruption as an input, quantifying the halo CME association with solar flares. The flare data archive used is available on the website: <http://hesperia.gsfc.nasa.gov/goes/goeseventlistings/goesxrayeventlist2014.txt>.

2.2 Key interplanetary parameters

The main inputs for any space weather prediction based on properties of interplanetary medium are known to be the solar wind speed at 1 AU, the total interplanetary magnetic field (BT) and the southward component of the interplanetary magnetic field (B_z) (Srivastava 2005). 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). 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, $E_y = VB_s$, 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). So for the ANN model developed in this study, halo CMEs ($AW \geq 120^\circ$), CME speed (V_{cme}), peak flux as well as IP peak values of negative B_z , E_y and SW speed (V_{sw}) were used as ANN numeric input (as shown in table 1) (table 2).

The peak values (V_{sw} , B_z and E_y) correspond to the maxima (magnitude only) recorded during the time period of ICME passage. SW data are provided by the OMNI-2 dataset and available online (<http://www.nssdc.gsfc.nasa.gov/omniweb.html>).

Table 1. List of all 110 CME driven geomagnetic storm events along with solar and interplanetary parameters and their arrivals.

Dst (min.) [nT]	Halo CMES		Vcmes [km s ⁻¹]	Bz [nT]	Vsw [km s ⁻¹]	AW [deg.]	Ey [mV m ⁻¹]	Flare Class	Lnf	
	Date	Time								
-59	FH	12/02/2010	13:42:04	509	-12	379	360	3.9	B8.9	-13.932
-81	FH	03/04/2010	10:33:58	668	-6.9	814	360	4.52	B7.4	-14.117
-67	PH	08/04/2010	04:54:07	227	-8	468	156	3.34	B3.7	-14.81
-80	FH	23/05/2010	18:06:05	258	-8	468	360	4.86	B1.3	-15.856
-80	FH	24/05/2010	14:06:05	427	-13.8	385	360	4.86	B1.1	-16.023
-74	FH	01/08/2010	13:42:05	850	-10.5	598	360	6.28	B0.0	-16.118
-74	PH	01/08/2010	23:18:05	527	-10.5	598	120	6.28	B0.0	-16.118
-75	PH	06/10/2010	07:12:05	282	-11.6	447	152	4.15	B0.0	-16.118
-63	FH	01/02/2011	23:24	437	-15.9	647	360	7.03	B4.5	-14.614
-88	PH	26/02/2011	20:54	379	-12.1	687	172	4.33	B4.3	-14.66
-83	PH	07/03/2011	14:48	698	-10	405	261	3.76	M1.5	-11.108
-83	FH	07/03/2011	20:00	2125	-10	405	360	3.76	M1.5	-11.108
-83	PH	08/03/2011	04:12	732	-10	405	282	3.76	M1.5	-11.108
-60	PH	04/04/2011	15:57	2081	-8.7	579	120	4.35	B0.0	-16.118
-80	PH	24/05/2011	21:24	657	-10.7	752	122	5.44	B2.0	-15.425
-59	PH	02/07/2011	07:50	511	-8.7	414	196	3.57	B0.0	-16.118
-115	FH	03/08/2011	14:00:07	610	-19.3	611	196	3.57	M6.0	-9.7212
-115	FH	04/08/2011	04:12	1315	-19.3	611	360	11.66	M9.3	-9.2829
-75	FH	06/09/2011	02:24	782	-17.1	560	360	8.34	M5.3	-9.8452
-75	FH	06/09/2011	23:05	575	-17.1	560	360	8.34	X2.1	-8.4684
-75	PH	07/09/2011	18:48:05	924	-17.1	560	188	8.34	B9.1	-13.91
-75	PH	07/09/2011	23:05:58	792	-17.1	560	167	8.34	X1.8	-8.6226
-75	PH	08/09/2011	22:12:06	983	-17.1	560	281	8.34	C1.1	-13.72
-72	PH	14/09/2011	00:00	408	-7.6	549	242	3.79	B0.0	-16.118
-72	PH	14/09/2011	20:12	375	-7.6	549	131	3.79	C9.2	-11.596
-72	PH	15/09/2011	00:00	530	-7.6	549	130	3.79	C4.6	-12.289
-72	PH	16/09/2011	23:48	669	-7.6	549	148	3.79	C2.6	-12.86
-118	FH	24/09/2011	12:48	1915	-24.1	704	360	14.34	M7.1	-9.5528
-118	FH	24/09/2011	19:36	972	-24.1	704	360	14.34	M3.0	-10.414
-147	FH	22/10/2011	01:25	593	-13.1	516	360	6.67	B0.0	-16.118
-147	FH	22/10/2011	10:24	1005	-13.1	516	360	6.67	M1.3	-11.251
-66	PH	29/10/2011	13:48	295	-7.9	436	147	3.39	C1.0	-13.816
-73	FH	19/01/2012	14:36	1120	-11.9	459	360	4.78	M3.2	-10.35
-80	PH	23/01/2012	03:12	684	-8.9	607	221	5.4	B0.0	-16.118
-80	FH	23/01/2012	04:00	2175	-8.9	607	360	5.4	M8.7	-9.3496
-62	FH	10/02/2012	20:00	533	-8.3	414	360	3.19	B0.0	-16.118
-54	FH	16/02/2012	06:36	538	-9.7	489	360	3.66	B0.0	-16.118
-85	PH	04/03/2012	05:00	584	-15.3	592	160	6.53	B0.0	-16.118
-85	PH	04/03/2012	11:00	1306	-15.3	592	360	6.53	M2.0	-10.82
-85	FH	05/03/2012	04:00	1531	-15.3	592	360	6.53	X1.1	-9.115
-143	FH	07/03/2012	00:24	2684	-16.4	737	360	11.71	X5.4	-7.5239
-143	FH	07/03/2012	01:30	1825	-16.4	737	360	11.71	X1.3	-8.948
-51	FH	10/03/2012	18:00	1296	-11.1	723	360	5.54	M8.4	-9.3847
-80	FH	13/03/2012	17:36	1884	-9.2	787	360	6.84	M7.9	-9.446
-55	PH	26/03/2012	06:48	783	-9.1	464	133	3.67	C1.4	-13.479
-55	FH	26/03/2012	23:12	1390	-9.1	464	360	3.67	C2.7	-12.822
-56	PH	02/04/2012	02:12	350	-8.9	361	135	3.01	B7.9	-14.051
-104	PH	19/04/2012	15:12	540	-15.4	601	142	5.89	C1.8	-13.228
-86	PH	13/06/2012	13:25	632	-16.5	519	253	6.93	M1.2	-11.331
-86	FH	14/06/2012	14:12	987	-16.5	519	360	6.93	M1.9	-10.871
-69	FH	06/07/2012	23:24	1828	-11.4	401	360	4.45	X1.1	-9.115
-133	PH	11/07/2012	10:36	748	-18.7	667	166	11.37	B0.0	-16.118

Table 1. (Continued.)

Dst (min.) [nT]	Halo CMES		Vcmes [km s ⁻¹]	Bz [nT]	Vsw [km s ⁻¹]	AW [deg.]	Ey [mV m ⁻¹]	Flare Class	Lnf	
	Date	Time								
-133	FH	12/07/2012	16:48	885	-18.7	667	360	11.37	B0.0	-16.118
-78	PH	30/08/2012	15:12	480	-11.3	545	130	4.89	B0.0	-16.118
-78	FH	31/08/2012	20:00	1442	-11.3	545	360	4.89	C8.4	-11.687
-133	FH	28/09/2012	00:12	947	-19.2	410	360	7.53	B0.0	-16.118
-133	FH	28/09/2012	10:36	768	-19.2	410	360	7.53	B0.0	-16.118
-106	PH	03/10/2012	02:48	383	-14.4	447	153	5.39	B0.0	-16.118
-111	PH	05/10/2012	20:48	284	-15.1	466	284	5.87	B0.0	-16.118
-74	FH	28/10/2012	16:48	317	-11.7	373	360	4.24	B7.3	-14.13
-74	PH	28/10/2012	08:48	487	-11.7	373	158	4.24	C1.7	-13.285
-109	PH	10/11/2012	14:12	460	-17.4	482	191	6.85	B0.0	-16.118
-109	PH	11/11/2012	00:48	690	-17.4	482	154	6.85	C1.3	-13.553
-53	PH	15/01/2013	19:48	798	-12.3	458	162	4.74	C1.8	-13.228
-51	PH	21/01/2013	05:36	304	-10.6	531	128	4.21	B0.0	-16.118
-51	PH	23/01/2013	03:12	501	-10.6	531	122	4.21	B0.0	-16.118
-51	PH	23/01/2013	14:12	530	-10.6	531	185	4.21	B0.0	-16.118
-55	PH	27/02/2013	04:00	622	-13.4	638	138	6.32	B8.3	-14.002
-132	FH	15/03/2013	07:12	1063	-14.4	725	360	9.71	M1.1	-11.418
-50	FH	20/04/2013	06:00	741	-12.8	562	153	4.43	B0.0	-16.118
-50	FH	21/04/2013	07:24	919	-12.8	562	360	4.43	B8.7	-13.955
-50	FH	21/04/2013	16:00	857	-12.8	562	130	4.43	C2.9	-12.751
-67	PH	26/04/2013	18:24	271	-9.9	484	150	3.97	C2.3	-12.983
-67	PH	28/04/2013	13:25	400	-9.9	484	121	3.97	B0.0	-16.118
-57	FH	14/05/2013	01:25	2625	-7.9	439	360	3.42	X3.2	-8.0472
-54	PH	21/05/2013	02:24	562	-6.6	757	185	3.85	B0.0	-16.118
-54	FH	22/05/2013	13:25	1466	-6.6	757	360	3.85	M5.0	-9.9035
-119	PH	27/05/2013	19:24	528	-17.4	722	126	7.06	B0.0	-16.118
-97	FH	25/06/2013	11:12	349	-11.9	552	360	4.58	B6.6	-14.231
-79	PH	03/07/2013	07:24	807	-12.5	369	267	4.44	M1.5	-11.11
-54	PH	21/10/2013	05:12	385	-10	481	228	4.07	C1.6	-13.35
-67	FH	29/09/2013	22:12	1179	-8.8	629	360	5.54	C1.2	-13.63
-62	PH	04/10/2013	19:07	392	-6.2	639	229	2.74	B0.0	-16.12
-62	FH	05/10/2013	07:09	964	-6.2	639	360	2.74	B0.0	-16.12
-50	FH	28/10/2013	02:24	695	-8.1	436	360	2.86	X1.0	-9.21
-50	FH	28/10/2013	15:36	812	-8.1	436	360	2.86	M4.4	-10.03
-54	PH	05/11/2013	08:24:06	850	-8.6	383	197	3.11	M2.5	-10.6
-81	PH	05/11/2013	22:36:05	562	-11.4	611	195	4.8	X3.3	-8.016
-81	FH	07/11/2013	00:00:06	1033	-11.4	611	360	4.8	B0.0	-16.12
-81	FH	07/11/2013	10:36:05	1405	-11.4	611	360	4.8	C2.1	-13.07
-70	PH	06/11/2013	14:24:26	347	-6.2	539	122	3.34	M3.8	-10.18
-66	FH	07/12/2013	07:36:05	1085	-9.8	639	360	4.12	M1.2	-11.33
-112	FH	16/2/2014	10:00:05	634	-12.9	474	360	6.09	M1.1	-11.418
-112	PH	16/2/2014	12:48:05	659	-12.9	474	243	6.09	C0.0	-13.816
-73	PH	17/2/2014	05:12:05	582	-8.5	691	121	5.87	B0.0	-16.118
-86	FH	18/2/2014	01:36:21	779	-8.3	618	360	5.13	C4.7	-12.268
-56	PH	21/2/2014	01:25:51	648	-8.3	498	143	3.82	B0.0	-16.118
-56	FH	21/2/2014	16:00:05	1252	-8.3	498	360	3.82	B0.0	-16.118
-56	PH	22/2/2014	12:12	1033	-8.3	198	233	3.82	B0.0	-16.118
-99	PH	24/2/2014	11:36:05	495	-12	483	193	5.5	M1.2	-11.331
-99	PH	24/2/2014	23:24:07	790	-12	483	188	5.5	B0.0	-16.118
-80	FH	08/04/2014	23:12:12	514	-8.9	387	360	3.06	B9.6	-13.856
-57	PH	12/04/2014	07:24:06	1016	-7.9	346	139	2.69	C5.0	-12.206
-67	PH	25/04/2014	00:48:03	456	-9.4	309	296	2.86	X1.3	-8.948
-80	FH	24/08/2014	12:36:05	551	-13	333	360	3.89	M5.9	-9.738
-65	PH	25/08/2014	20:48:05	711	-10.2	328	177	3.27	M3.9	-10.152

Table 1. (Continued.)

Dst (min.) [nT]	Halo CMES		Vcmes [km s ⁻¹]	Bz [nT]	Vsw [km s ⁻¹]	AW [deg.]	Ey [mV m ⁻¹]	Flare Class	Lnf	
	Date	Time								
-75	PH	10/09/2014	17:24:05	1071	-11.3	726	134	7.12	X1.6	-8.7403
-75	FH	10/09/2014	18:00:05	1267	-11.3	726	360	7.12	X1.6	-8.7403
-57	PH	07/11/2014	18:08:34	795	-5.4	503	293	2.62	B0.0	-16.118
-71	FH	19/12/2014	01:04:42	1195	-14.9	429	360	5.04	B0.0	-16.118

Table 2. The input and output parameters and their corresponding measures.

Parameters	Parameter name	Variable type	Measure
Inputs	CME AW	Numeric	$\leq 120^\circ$
	CME Speed	Numeric	Values in km s ⁻¹
	F_{class}	Numeric	-
	V_{sw}	Numeric	Values in km s ⁻¹
	B_z	Numeric	Values in nT
	E_y	Numeric	Values in mV m ⁻¹
Outputs	Dst	Numeric	Values in nT

3. Artificial neural network (ANN) and principal component analysis (PCA)

In this study, ANN has been used as a tool in the development of a model to study the precursors leading to GMS from the observed solar and IP properties of halo CMEs. Neural network 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 ANN. Feed Forward Neural Networks (FFNN) represent the simplest and most popular type of ANN, 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). The ANNs generally have a layered structure in which each input passes through same number of nodes to produce an output. The first layer is the layer of input data, termed the input layer whereas the output layer is the layer of nodes which produces the output. The intermediate layers of processing units are called hidden layers.

Figure 1 illustrates the three-layered ANN architecture used in the present study. In a three layered FFNN with six input neurons, one hidden layer with different neurons and one output neuron. The activation functions we have used in hidden layer and output layer nodes are tan-sigmoid and linear functions, respectively. Such activation functions on the network outputs play an important role in allowing the outputs to be given a probabilistic interpretation (Bishop 1995). Indeed, ANNs provide an estimate of the posterior probabilities using

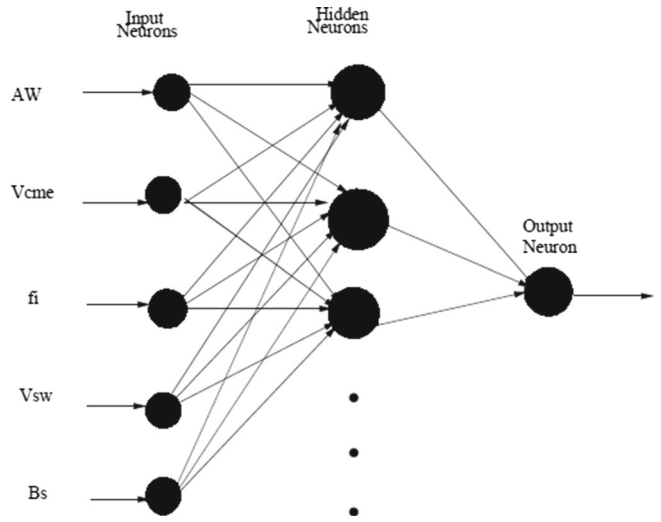


Figure 1. The three-layered FFNN architecture which was used in the study.

the least squares optimization and are sensitive to sample size. A larger database provides better estimates (Richard and Lippmann 1991; Hung *et al.* 1996). If there exists a relation between the input and the output, the network 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 dataset is generally split randomly into training, testing and validation datasets in order to avoid the training results becoming biased towards a particular section of the database. For the ANN trained while developing this model, data were split into 60% for the training set and 20% for the testing, and rest 20% for the validation dataset, in order to determine how the ANN has learned the behaviour in the input-output patterns, 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 and F_{class} , values of AW and SW speed in hundreds, CME speed in thousands), they were first normalized through weight initialization.

A typical batch learning procedure can be implemented through the following algorithm:

- (1) Read the input data
- (2) Initialize the weights

- (3) Repeat
 - (a) Initialize all weight correction to zero
 - (b) For each input point
 - (i) Calculate net output
 - (ii) Calculate gradient of error corresponding to each weight
 - (iii) Accumulate the weight corrections
 - (c) Modify the weights using average of the accumulated weight corrections until the error reduces to desired level or the iterations exceed the allowed maximum number.
- (4) Print the net output and net error
- (5) End

3.1 Principal component analysis

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a dataset by finding a new set of variables, smaller than the original set of variables that retains most of the sample information (Jolliffe 2002). It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Principal component analysis is appropriate when there are measures on a number of observed variables and we wish to develop a smaller number of artificial variables. Principal component analysis is a variable reduction procedure useful when data have a large number of variables and there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Principal component can be defined as a linear combination of optimally-weighted observed variables. The goal is to account for the variation in a sample in fewer variables. Steps involved in calculation of PCA are given below:

- *Giving data*
The data consists of 110 CMEs which need to be reduced so that it takes minimum time for classification, providing more accurate results.
- *Subtracting the mean*
For PCA to work properly, the mean is subtracted from each of the data dimensions. The mean subtracted, is the average across each dimension. This produces a dataset whose mean is zero.
- *Calculating the covariance matrix*
Since the data is 2 dimensional, the covariance matrix will be calculated. Covariance shows

the variation in the data which will ultimately account for pattern classification.

- *Calculating the eigenvectors and eigenvalues of the covariance matrix*

From the covariance matrix, calculate the eigenvectors and eigenvalues for this matrix. These are important as they contain useful information about our data. An eigenvalue represents the amount of variance that is accounted for by a given component. It is important to notice that these eigenvectors are both unit eigenvectors, i.e., their lengths are both 1. This is very important for PCA that when asked for eigenvectors, it must give unit eigenvectors. These are important as they provide with information about the patterns in the data.

- *Choosing components and forming a feature vector*

It turns out that the eigenvector with the highest eigenvalue is the principal component of the dataset. Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives the components in order of significance. The components with lesser significance can be ignored. Originally data have n dimensions and n eigenvectors and eigenvalues but final dataset can have only p dimensions. This is constructed by taking the eigenvectors that are kept from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

$$\text{Feature vector} = (eig1\ eig2\ eig3 \dots\ eign)$$

- *Deriving the new dataset*

Once components (eigenvectors) are chosen and formed a feature vector, take the transpose of the vector and multiply it on the left of the original dataset, transposed as:

$$\text{Final data} = \text{Row feature vector} \times \text{row data adjust}$$

where *row feature vector* is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and *row data adjust* is the mean-adjusted data transposed, i.e., the data items are in each column, with each row holding a separate dimension. Final data is the final dataset, with data items in columns and dimensions along rows.

The PCA plot of the data is shown in figure 2 which shows the highly nonlinear behaviour of the CME data (110 CMEs) when the data dimension is reduced from six variables to a set of orthonormal basis (principal components). When the first three components are plotted, they give the maximum information about the data.

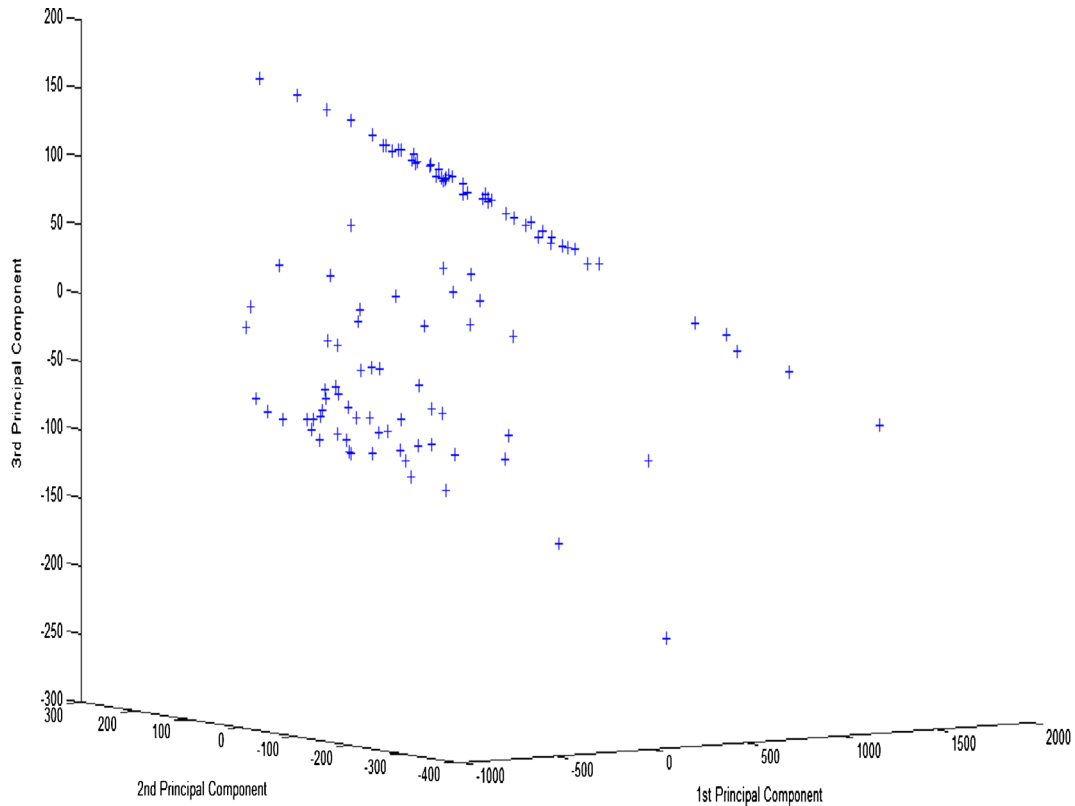


Figure 2. The PCA plot showing the linearity in the CME data.

4. Result and discussion

We have first discussed the interdependency of various input parameters and then discussed our results from ANN by various plots showing the network output.

Figure 3 shows the geomagnetic storm of 7 March, 2012 with peak minimum $Dst = -143$ nT followed by two full halo CMEs ($AW = 360^\circ$) at 00:24 UT and 01:30 UT with high CME speed (V_{cme}) of about 2684 km s^{-1} and 1825 km s^{-1} , respectively, that were probable sources of the storm. Indeed, the two halo CMEs involved were associated with X-class solar flares and were followed by an ICME also observed on 7 March, 2012 (on the 67th day of the year). The red lines show that at the time of sudden storm commencement (SSC), there is a sudden increase in the magnitude of B_z and Dst while the wind speed (V_{sw}) and E_y decreases suddenly (on the 68th day of the year). Since the storm is followed by two full halo CMEs, there were two sudden storm commencement (SSC) regions depending upon their CME speeds. However, we get only one peak of Dst (min.) which is more pronounced and include the effects of both the CMEs shown by the black line (on 69th day of the year). This plot also depicts the interdependency of input parameters as well as with the output parameter. The B_z and Dst vary in

the same manner while the V_{sw} and E_y variations were out of phase, but their magnitudes vary in the same manner. So we have taken interplanetary parameters along with solar parameters as input variables while Dst (min) as an output variable for training, testing and validation of dataset in ANN.

The input given to the ANN for training in a manner such that rows contain the attributes and column contain the number of samples. The activation function of the hidden layer and output layer are logistic sigmoid function and pure linear function. The number of neurons in the hidden layer is determined empirically by experimenting with various network configurations. The ANNs were trained by the gradient descent back-propagation training algorithm. When the network is trained, the network gives the correct output upon testing and validation. The optimum network architecture was found to be that with six inputs using eight hidden nodes (configuration: 6:8:1). The network with only seven hidden nodes was found to perform poorly when tested on the validation dataset.

Uwamahoro *et al.* (2012) considered the ANN with only three solar input parameters, which performed poorly when tested on the validation dataset, indicating the importance of considering the IP parameters for improving the model performance.

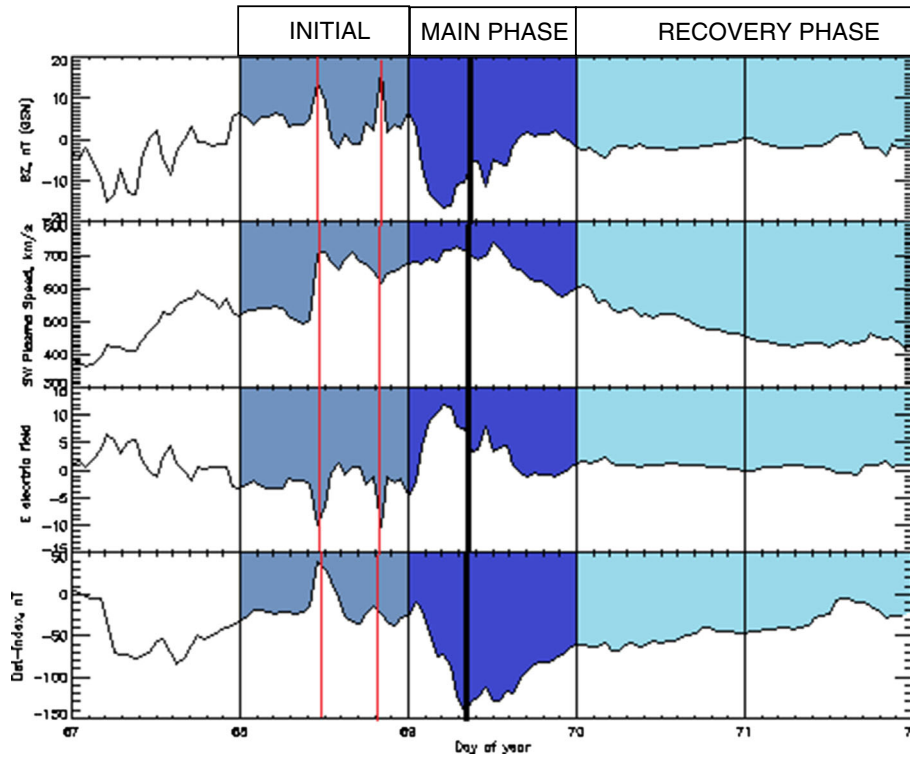


Figure 3. A 5-day window plot showing the variation of the interplanetary parameters, i.e., z-component of the interplanetary magnetic field (IMF) B_z , the southward wind speed V_{sw} , the electric field E_y with the output, Dst index, following the passage of the interplanetary coronal mass ejections (ICMEs), observed on 7 March, 2012.

Table 3. The ANN architecture used in the study with their corresponding best validation performances.

No. of hidden layers	Sample ratio	Best validation performance
19	6:7:1	325.6077
	6:8:1	254.905
	6:9:1	583.702
20	6:7:1	350.6579
	6:8:1	201.3559
	6:9:1	338.1322

Table 3 shows that with the increase in the hidden layer, the network performance improves and we found that our network performs best with 20 hidden neurons. Also on changing the sample ratio of the neurons, i.e., input : hidden : output, the network performance changes and the networks best validation performance was found to be for the sample ratio 6:8:1 with 20 hidden neurons.

Figure 4 shows the best validation performance or the mean squared error (MSE) is 201.3559 at epoch 497 and the errors in the training, test and validation process decrease in the same manner. The regression plots are shown in figure 5 which clearly shows that the model validates up to 79% with an overall regression up to 77.36%. Thus the network is successful in predicting the occurrence of intense storms up to 77.36% and with a

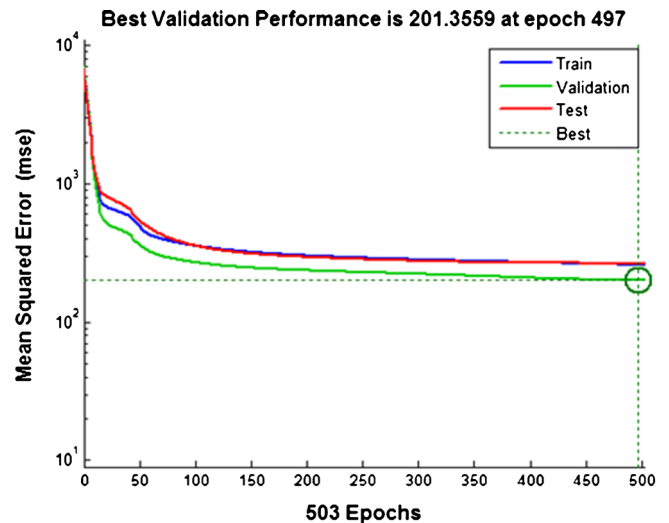


Figure 4. The performance plot with 19 hidden layer and 8 hidden neurons.

validation of 79% for the sample ratio 60:20:20 for training, validation and test respectively.

Uwamahoro *et al.* (2012) using the ANN model predicted 100% of intense storms and 75% of moderate storms. The overall ANN model prediction ability of GMS ($Dst < -50$ nT) based on the observed halo CME was estimated at 86%. The results obtained demonstrate the ability of the ANN 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 flare activity as well as high peak values of B_z 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 (possess lower speed). Note that moderate storms are often driven by the non-halo CMEs or CIRs that have not been considered in the present study.

The results presented in this study only serve as an indication that solar and IP parameter characteristics of geo-effective halo CMEs, can be used in an ANN to estimate the probability occurrence of the subsequent GMS. The estimated geo-effectiveness 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 ANN model and obtained a 48% successful forecast for severe geomagnetic response. The ANN model described in the present study shows an improved performance with an accuracy of 77.36% with a validation performance of 79%. We also want to emphasize that our topic of interest was to estimate the GMS occurrence following

halo CMEs observed during the ascending phase of 24th solar cycle between 2009 and 2014; however, a larger database provides better estimates (Richard and Lippmann 1991; Hung *et al.* 1996). 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 ANN model described in the present study is unique, as our main focus is to estimate the probability occurrence of both intense and moderate storms, which were not effectively done in previous studies.

5. Summary and conclusion

Predicting the occurrence of GMS on the basis of CME observations by ANN architecture is a very effective tool. In this study, a combination of solar and IP parameters have been used as inputs in an ANN model with the 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 geo-effective 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. It was observed

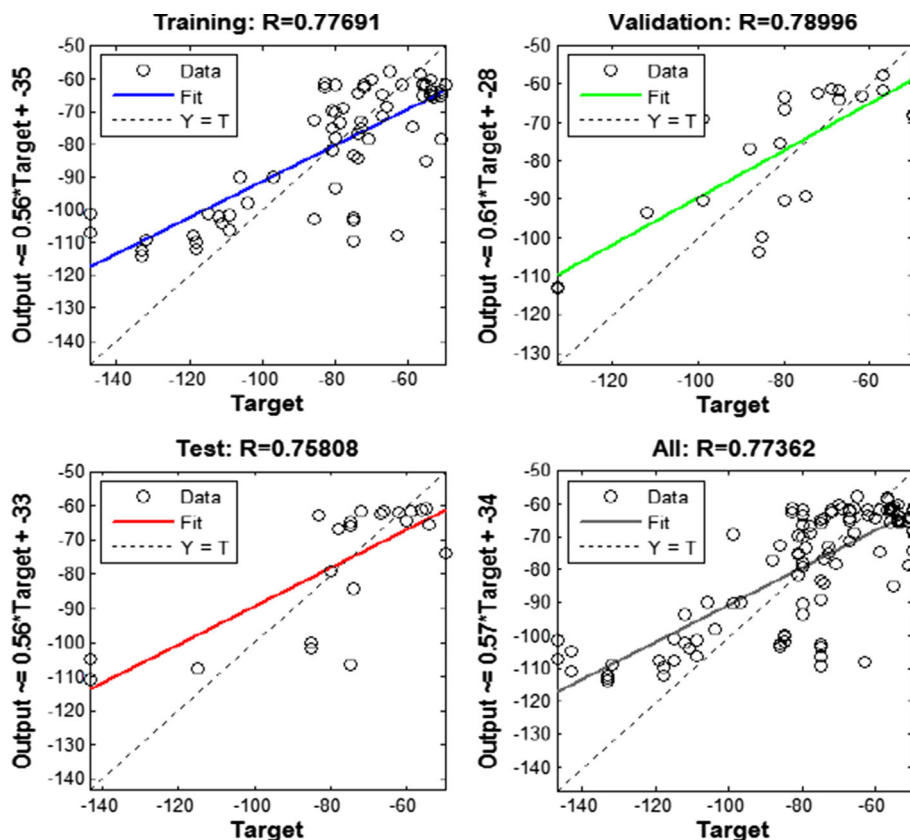


Figure 5. The regression plot showing the output-target fit.

by use of the PCA technique that the GMS data is highly nonlinear; however, our model can correctly predict up to 79% of the GMS resulting from CMEs. Thus, the ANN model described in this paper will contribute towards improving real-time space weather predictions and is used to minimize the effects on the radio as well as satellite communication.

Acknowledgements

This work is financially supported by ISRO, Bangalore under ISRO-SSPS to BHU. We acknowledge the authors of the LASCO/SOHO catalogue list of CMEs, available online at <http://cdaw.gsfc.nasa.gov/CMElist> which was used for our study. We would also like to thank the National Aeronautics and Space Administration for providing us the necessary data of the most significant flare, available online at http://hesperia.gsfc.nasa.gov/goes/goes_event_listings/goes_xray_event_list_2014.txt. We are also thankful to the reviewers for their suggestions to improve the quality of the manuscript.

References

- Bishop C M 1995 *Neural Networks for Pattern Recognition*; Oxford University Press Inc. New York, USA.
- 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 1995 The Large Angle Spectroscopic Coronagraph (LASCO); *Sol. Phys.* **162** 357–402.
- Conway A J 1998 Time series, neural networks and the future of the Sun; *New Astron. Rev.* **42** 343–394.
- Dryer M, Smith Z, Fry C D, Sun W, Deehr C S and Akasofu S I 2004 Real-time shock arrival predictions during the Halloween 2003 epoch; *Space Weather* **2** S09001, doi: [10.1029/2004SW000087](https://doi.org/10.1029/2004SW000087).
- Echer E, Gonzalez W D, Tsurutani B T and Gonzalez A L C 2008 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](https://doi.org/10.1029/2007JA012744).
- Gonzalez W D and Tsurutani B T 1987 Criteria of interplanetary parameters causing intense magnetic storms (Dst of less than -100 nT); *Planet. Space Sci.* **35** 1101–1109, doi: [10.1016/0032-0633\(87\)90015-8](https://doi.org/10.1016/0032-0633(87)90015-8).
- Gopalswamy N, Yashiro S and Akiyama S 2007 Geoeffectiveness of halo coronal mass ejections; *J. Geophys. Res.* **112** A06112, doi: [10.1029/2006JA012149](https://doi.org/10.1029/2006JA012149).
- Hung M S, Hu M Y, Shanker M S and Patuwo B E 1996 Estimating posterior probabilities in classification problems with neural networks; *Int. J. Comput. Intelligence and Organizations*, pp. 149–60.
- Jolliffe I 2002 *Principal Component Analysis*; Springer Publication.
- Lundstedt H and Wintoft P 1994 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](https://doi.org/10.1007/s00585-994-0019-2).
- Lundstedt H 1997 AI techniques in geomagnetic storm forecasting; In: Magnetic Storms (eds) Tsurutani B T, Gonzalez W D, Kamide Y and Arballo J K, *Geophys. Monogr. Ser.* AGU, Washington D.C., **98** 243–252.
- Lundstedt H, Gleisner H and Wintoft P 2002 Operational forecasts of the geomagnetic Dst index; *Geophys. Res. Lett.* **29** 2181, doi: [10.1029/2002GL016151](https://doi.org/10.1029/2002GL016151).
- Lundstedt H, Liszka L and Lundin R 2005 Solar activity explored with new wavelet methods; *Ann. Geophys.* **23** 1505–1511, doi: [10.5194/angeo-23-1505-2005](https://doi.org/10.5194/angeo-23-1505-2005).
- Macpherson K P, Conway A J and Brown J C 1995 Prediction of solar and geomagnetic activity data using neural networks; *J. Geophys. Res.* **100** 735–744.
- Pallochia G, Amata E, Consolini G, Marcucci M F and Bertello I 2006 Geomagnetic Dst index forecast based on IMF data only; *Ann. Geophys.* **24** 989–999, doi: [10.5194/angeo-24-989-2006](https://doi.org/10.5194/angeo-24-989-2006).
- Richard M D and Lippmann R P 1991 Neural network classifiers estimate Bayesian *a Posteriori* probabilities; *Neural Computation* **3** 461–483.
- Robbrecht E, Berghmans D and Van der Linden 2009 RAM Automated Lasco CME catalog for solar cycle 23: Are CMEs scale invariant?; *Astrophys. J.* **691**(2) 1222.
- Siscoe G and Schwenn R 2006 CME disturbance forecasting; *Space Sci. Rev.* **123** 453–470.
- Srivastava N 2005 A logistic regression model for predicting the occurrence of intense geomagnetic storms; *Ann. Geophys.* **23** 2969–2974, doi: [10.5194/angeo-23-2969-2005](https://doi.org/10.5194/angeo-23-2969-2005).
- Srivastava N and Venkatakrishnan P 2004 Solar and interplanetary sources of geomagnetic storms during 1996–2002; *J. Geophys. Res.* **109** A10103, doi: [10.1029/2003JA010175](https://doi.org/10.1029/2003JA010175).
- Uwamahoro J, McKinnell L A and Cilliers P J 2009 Forecasting solar cycle 24 using neural networks; *J. Atmos. Sol.-Terr. Phys.* **71** 569–574.
- Uwamahoro J, McKinnell L A and Habarulema J B 2012 Estimating the geo-effectiveness of halo CMEs from associated solar and IP parameters using neural networks; *Ann. Geophys.* **30** 963–972.
- Valach F, Revallo M, Bochnicek J and Hejda P 2009 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](https://doi.org/10.1029/2008SW000421).
- Wang Y M, Ye P Z, Wang S, Zhou G P and Wang J X 2002 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](https://doi.org/10.1029/2002JA009244).
- Webb D F 2000 Coronal mass ejections: Origins, evolution, and role in space weather; *IEEE Trans. Plasma Sci.* **28** 1795–1806.
- Webb D F, Cliver E W, Crooker N U, St Cyr O C and Thompson B J 2000 Relationship of halo coronal mass ejections, magnetic clouds and magnetic storms; *J. Geophys. Res.* **105** 7491–7508.
- Woolley J W, Agarwal P K and Baker J 2010 Modeling and prediction of chaotic systems with artificial neural networks; *Int. J. Numerical Methods in Fluids* **63** 2117, doi: [10.1002/flid.2117](https://doi.org/10.1002/flid.2117).

MS received 3 September 2015; revised 23 February 2016; accepted 25 February 2016