GEOPHYSICAL RESEARCH LETTERS, VOL. 26, NO. 10, PAGES 1353-1356, MAY 15, 1999

Forecasting Auroral Electrojet Activity from Solar Wind Input with Neural Networks

R.S. Weigel, W. Horton, T. Tajima

Institute for Fusion Studies, The University of Texas, Austin, TX 78712

T. Detman

NOAA Space Environment Center, Boulder, CO 80303

Abstract. Neural networks are developed for reconstructing the chaotic attractor in the nonlinear dynamics of the solar wind driven, coupled magnetosphere–ionosphere (MI) system. Two new methods which improve predictive ability are considered: a gating method which accounts for different levels of activity and a preconditioning algorithm which allows the network to ignore very short time fluctuations during training. The two networks are constructed using the Bargatze *et al.* [1985] substorm database that contains solar wind speed and interplanetary magnetic field (IMF) along with ionospheric electrojet index, AL. Both networks are found to produce improvements in predictability, and the significance of the performance increase of the gated network is demonstrated using the bootstrap model testing method.

1. Introduction

Forecasting the auroral region westward electrojet currents measured by the AL geomagnetic index with solar wind satellite data is of fundamental importance in space physics. The solar wind driver used in the better known substorm databases (VB_s) is the product of the southward component (B_s) of the interplanetary magnetic field (IMF), and the earthward component of the solar wind velocity (V). The Bargatze *et al.* [1985] database gives these input-output (VB_s–AL) time series sorted according to the strength of the geomagnetic disturbances at 2.5 min. time intervals.

The solar wind driven magnetosphere–ionosphere system is a complex dynamical system requiring a large number of ($\gg 10$) of physics parameters to describe from first principle calculations [*Papadopoulos et al., 1995*; *Horton et al.*, 1998]. Modern chaos science suggests that chaotic dynamics due to feedback loops in physical systems can be effectively modeled using a nonlinear mapping method. Artificial neural networks with inputs motivated by the state space reconstruction approach [*Packard et al.*, 1980] are well suited for constructing such nonlinear maps [*Kulkarni et al.*, 1998].

Earlier related approaches which use a mapping to predict the input-output relationship of the solar wind driven auroral westward electrojet index (AL-VB_s) data used local linear ARMA filters [*Price and Prichard*, 1993; *Vassiliadis et al.* 1994; *Detman and Vassiliadis*, 1997]. Another recent method is to model the magnetospheric dynamces as a low dimensional system driven by the solar wind. This is the approach taken in the nonlinear physics based model (WINDMI) developed by *Horton and Doxas* [1998]. This model takes into account conservation of energy and charge and has the advantage of having relatively few parameters compared to the number of weights required for a successful network model.

In this study we extend the work of *Hernandez et al.* [1993] by including a larger set of data when training and by developing a selection band on a new architecture which takes into account activity level by using a gated network which makes a prediction based on the outputs from networks trained on intervals with differing levels of activity [*Ramamurti and Ghosh*, 1998]. This architecture is able to account for the scaling problem intrinsic to neural networks with nonlinear activation functions. Finally, we consider preconditioning the input data. Future work will compare the predictive ability of these neural networks with that of physics based models.

2. Neural Networks and State Space Reconstruction

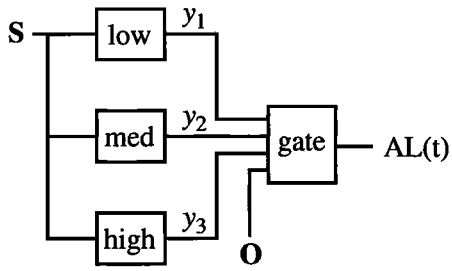
A neural network is a system with parallel processing units that has the ability to learn input–output relations and recognize patterns directly from a database. The basic architecture consists of an input layer, a number of hidden layers, and an output layer. When used for forecasting purposes, the goal is to produce an output which is the same as the output of the physical system under consideration, given an input which is believed to contain enough information to determine the output. One of the challenges of using a neural network for forecasting purposes is to determine the relevant input variables given measurements of only a few variables. For systems with many degrees of freedom the results of modern dynamical system theory are used as a guide.

Given a scalar time series, $O(t)$, produced from a measurement of an autonomous system with many degrees of freedom, it was shown by *Packard et al.* [1980] that if the dynamics of the system lies on a low-dimensional attractor, then the attractor can be reconstructed by creating a delayed coordinate vector

$$\mathbf{O}(t) = O(t-\tau), O(t-2\tau), ..., O(t-m\tau). \qquad (1)$$

The quality of the attractor reconstruction depends on m and τ, and when Eq. 1 is used for predictive purposes, the value of m for which the network gives optimal predictions is

Paper number 1999GL900280.
0094-8276/99/1999GL900280$05.00



$\mathbf{S} = VB_s(t\text{-}\tau), ..., VB_s(t\text{-}l\tau), AL(t\text{-}\tau), ..., AL(t\text{-}l\tau)$
$\mathbf{O} = AL(t\text{-}\tau), ..., AL(t\text{-}l\tau)$

Figure 1. Schematic of gated network. Each individual "expert" network is individually trained using a different level of activity, and their AL predictions y_1, y_2, and y_3 become input to the gate net along with **O**. Given y_1, y_2, y_3 and **O**, the gate net predicts AL(t).

chosen. In this work we use the prediction time τ =15 min. in order to compare with *Hernandez et al.*

In the case of the solar wind driven response of the magnetosphere, we have an input-output system, rather than an autonomous system. Therefore, to specify the state of the system for a network, we must include the vector which contains information about the input at past times [*Casdagli*, 1992]. The state of the system, **S**, as viewed from the input of the neural network is then given by the $l+m$ dimensional vector

$$\mathbf{S}(t) = I(t-\tau), I(t-2\tau), ..., I(t-l\tau), \quad O(t-\tau), O(t-2\tau), ..., O(t-m\tau) \tag{2}$$

where $I(t)$ is the system driving variable, and m and l are adjustable.

In this work we use a multilayer perceptron in which the input layer has as many units as the dimension $(m + \ell)$ of $\mathbf{S}(t)$. The activation function for the output is linear, and there is a single output, giving the predicted AL. The neural net is given by

$$y = \sum_{j=1}^{M} w_j^{(2)} g\left(\sum_{i=1}^{l+m} w_{ji}^{(1)} S_i\right) \tag{3}$$

where the activation function is given by $g(x) = tanh(x)$, and M is is the number of hidden units. An intrinsic problem with the nonlinear mapping in (3) is that for a fixed set of weights, $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, the activation function may saturate and the output will be clipped when its argument becomes large. The effect of this saturation is seen in Figure 2c of *Hernandez et al.* [1993].

To avoid the clipping problem and to account for the fact that the magnetosphere has different levels of activity, we propose a new architecture suggested by the physics of the problem which requires a nonlinear response [*Bargatze et al., 1985*] to the driving amplitude. To model this, we separate the data into three groups corresponding to low, medium, and high activity levels, and for each activity level a separate network is trained. The low, medium, and high activity levels are chosen to be the intervals (1-10), (11-20), and (21-31) of the Bargatze database. Next, all three networks are combined, and their individual outputs are connected to a gating network as shown in Figure 1. In addition to the AL predictions of the three networks (y_1, y_2, y_3), we connect the vector $(y_4, ..., y_{(m+3)}) = \mathbf{O}(t)$, so that the gate network is

$$Y = \sum_{p=1}^{P} W_p^{(2)} g\left(\sum_{i=1}^{m+3} W_{pi}^{(1)} y_i\right) \tag{4}$$

and the weights $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$ are determined by training with intervals from each activity level with the previously determined weights, $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, held fixed. The number of hidden units in the gate network is given by P. The saturation effect will be avoided if the gate network recognizes when the individual networks will become saturated, and then bases its predictions on other information. In effect, the gate network is trained to make its prediction given the predictions of three expert networks, and the past activity levels; it does not pass the prediction of only one expert. Although at some times the gate may use weighted predictions from one or two expert nets which gave (presumably) poor predictions, we find that this cost is offset by the benefit of allowing a gate to weight the output of the experts given the activity level.

This three level network system is similar in context to the local linear filter approach used by *Vassiliadis et*

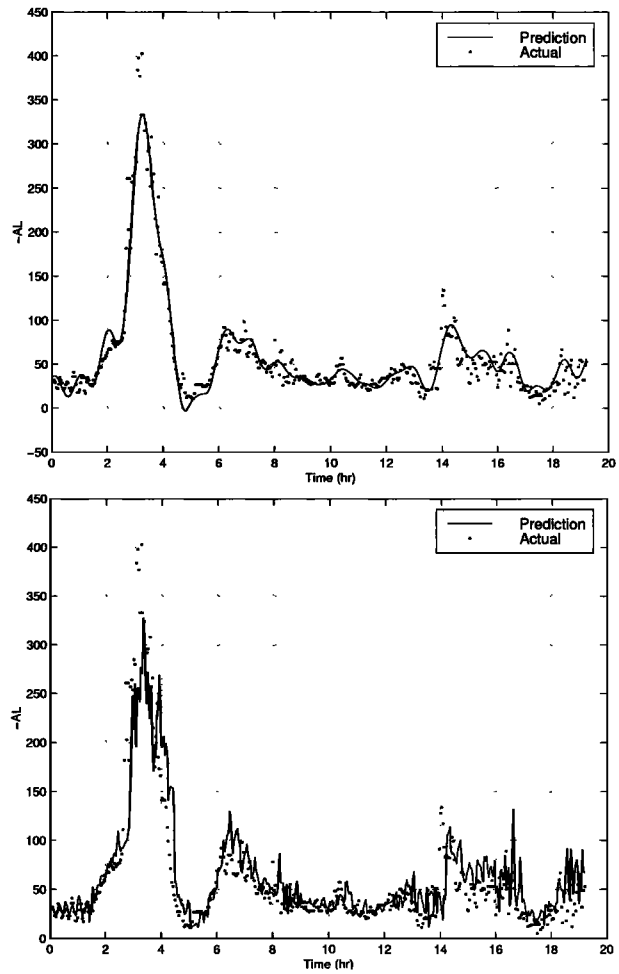


Figure 2. Top panel: 15 min. prediction of interval 8 with network trained on filtered data (ARV = 0.18). Bottom Panel: 15 min. prediction of same network trained with unfiltered data (ARV = 0.34). In both panels the dotted line is the actual value of the AL, and the solid line is the predicted AL.

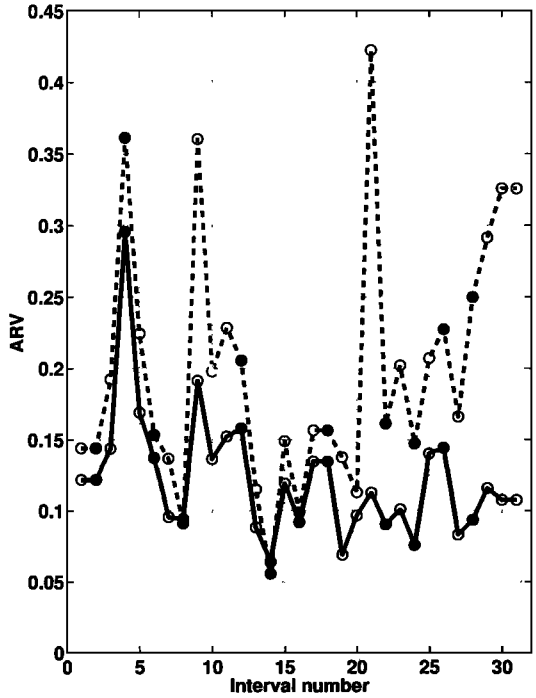


Figure 3. Comparison of ARV generated with gated network (solid line) with that without (dashed line). The training intervals have filled circles. The mean ARV of the non-training intervals is 0.12 for the gated network, and 0.20 for the ungated network.

al. [1994] in that both methods use different filters for different activity levels. The difference is that only three levels (low, medium, and high) are used for prediction here.

To determine the weights, we use the scaled conjugate gradient method [*Bishop*, 1995] to minimize the mean square error between the actual output y and the network output $\tilde{y}$ over the N forecast/observation pairs. Also, we introduce a regularization term Ω which is intended to penalize networks with very large values of weights. The error which is minimized is thus

$$E = \frac{1}{N}\sum_{j=1}^{N}(y_j - \tilde{y}_j)^2 + \nu\Omega. \tag{5}$$

The first term on the right hand side is the usual mean square error, and the second term is a constant ν multiplied by a weight factor

$$\Omega = \frac{1}{2}(\mathbf{w}^{(1)} \cdot \mathbf{w}^{(1)} + \mathbf{w}^{(2)} \cdot \mathbf{w}^{(2)}). \tag{6}$$

We use $\nu = 0.1$ in the networks constructed here. To determine the quality of the network, we use the average relative variance defined by

$$ARV = \frac{\sum_{j=1}^{N}(y_j - \tilde{y}_j)^2}{\sum_{j=1}^{N}(y_j - \bar{y})^2} \tag{7}$$

where the neural network output is represented by $\tilde{y}_j$; the actual and average measurements are represented by y_j and $\bar{y}$, respectively. The ARV can be interpreted as the fraction of the variance of the data which is not predicted by the model. If the ARV is 1, then the model predictions are no better than a prediction which uses just the average value of the measurements for the predictions ($\tilde{y}_j = \bar{y}$ for all j). Since the ARV will vary between test data intervals, we consider a large number of data sets and calculate the mean ARV ($\overline{\text{ARV}}$) of all the test sets.

In addition to the average relative variance, we calculate the correlation coefficient

$$\rho = \frac{\frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y})(\tilde{y}_j - \bar{\tilde{y}})}{\sigma_y \sigma_{\tilde{y}}} \tag{8}$$

which is a measure of the how closely the predicted and actual measurements are related. If ρ is near 1, then for a small range of predictions, there will be a small range of observations.

To compare the performance of different neural network methods, we calculate ARV and ρ for 16 test intervals and determine their average. To test the null hypothesis that the difference between these averages was due to chance, we use the bootstrap method as an alternative to using the t-test which assumes the fluctuations around the mean follow an approximately normal distribution. The bootstrap method will reproduce the t-test significance levels for approximately normal probability distributions, and produces reliable significance levels for most other probability distributions [*Efron and Tibshirani*, 1993].

To determine the significance level of the null hypothesis that the difference $\theta = \overline{\text{ARV}}_a - \overline{\text{ARV}}_b > 0$ between methods a and b was due to chance, we construct $K = 10^4$ bootstrap ARV vectors for each model. For example, if the original 16 element ARV vector for method "$a(b)$" is $\mathbf{ARV}_{a(b)} = (\text{ARV}_1, \text{ARV}_2, ..., \text{ARV}_{16})_{a(b)}$ then one of the 16 element bootstrap vectors determined from sampling with replacement from $\mathbf{ARV}_{a(b)}$ could be

$$\mathbf{ARV}^*_{a(b)} = (ARV_1, ARV_3, ARV_3, ARV_{12}, ..., ARV_7)_{a(b)}. \tag{9}$$

The simplest estimate of the bootstrap significance level is determined by looking at the histogram of $\theta^* = \overline{\text{ARV}}^*_a - \overline{\text{ARV}}^*_b$ of the K bootstrap vectors, and then taking the ratio of the number of times $\theta^* < 0$ to the number of bootstrap samples K. If θ^* was never less than zero, then we quote the significance level as $p < 1/K$. The significance level for the difference in the correlation coefficients, $\phi = \bar{\rho}_a - \bar{\rho}_b$, is determined in exactly the same way as θ.

3. Results

Visual inspection of the AL and VB_s time series reveals fluctuations on a very short time scale (minutes). Since we are more interested in long time predictions, we expect better predictions if we omit these short time scale fluctuations during training so that a network will train on longer term trends. To filter the data, we Fourier transform the time series and then set the highest 95% of the frequency components to zero, which corresponds to approximately 14% of the power. To test the effectiveness of this preconditioning method, we train one network with input and output data that has been filtered and another for which the raw time series has been used. Both networks have 8 input and hidden units, use the architecture given in Eqn. 3 with $\tau = 15$ min. and $l = m = 6$, and are trained for 15 min. predictions using the 15 even numbered intervals 2,...,30 of the Bargatze database. In addition to filtering the data as described

above, we scale both the AL and VB_s by 10^3. To calculate the ARV and correlation coefficient, we use the unfiltered database values. In Figure 2a we see the 15 min. predictions of the network which was trained on filtered input and output data. Figure 2b shows the predictions when unfiltered training data was used. As shown in Figure 2a, filtering the input and output when training gives a significant improvement in the average ARV over the 16 test intervals (intervals $1, 3, ..., 31$), $\overline{\text{ARV}}_{\text{filtered}} = 0.18$ vs. $\overline{\text{ARV}}_{\text{unfiltered}} = 0.34$ with $\theta = \overline{\text{ARV}}_{\text{unfiltered}} - \overline{\text{ARV}}_{\text{filtered}} = 0.16$ at a significance level of $p < 10^{-4}$. The correlation coefficients were also improved, with $\bar{\rho}_{\text{filtered}} = 0.91$, $\bar{\rho}_{\text{unfiltered}} = 0.81$ and $\phi = \bar{\rho}_{\text{filtered}} - \bar{\rho}_{\text{unfiltered}} = 0.10$ at a significance level of $p < 10^{-4}$. We also note that the network trained on the unfiltered data has a tendency to produce a time–delayed copy of the actual output which can be seen in Figure 2b. This is an indication that the network is not recognizing underlying dynamics, but is minimizing the ARV by producing a time–delayed copy of the output.

Next we consider a gated network which takes into account the dynamics of the solar wind – AL response by splitting the training data into three intervals corresponding to low (intervals 1-10), medium (11-20) and high (21-31) activity levels. The average of the maximum $|AL|$ values over these three intervals are 350nT, 550nT, and 900nT, respectively. For each activity interval an "expert" network is trained (8 input and hidden units, τ = 15 min. and $l = m = 6$) for 15 min. predictions using four of the filtered data sets from that interval. Finally, a fourth gating network is trained (7 inputs and 3 hidden units) with inputs from the output of each of the three "expert" networks and four additional inputs corresponding to the previous four activity levels of the time–series we are trying to predict. This network is shown in Figure 1. In Figure 3 we compare the effectiveness of the gating network with the outputs of a single network trained on the same intervals. We see that the 15 min. performance is dramatically improved, especially when the activity level is high (intervals 21-31), with $\overline{\text{ARV}}_{\text{gated}} = 0.12$, $\overline{\text{ARV}}_{\text{ungated}} = 0.20$ and $\theta = \overline{\text{ARV}}_{\text{ungated}} - \overline{\text{ARV}}_{\text{gated}} = 0.080$ at a significance level of $p < 10^{-4}$. The correlation coefficients were $\bar{\rho}_{\text{gated}} = 0.94$, $\bar{\rho}_{\text{ungated}} = 0.90$ and $\phi = \bar{\rho}_{\text{gated}} - \bar{\rho}_{\text{ungated}} = 0.04$ at a significance level of $p < 10^{-4}$.

In *Hernandez et al.* [1993], the performance of an ARMA and MA network were compared with a neural network. They found that the best ARMA and MA nets had a 15 min. prediction performance in the neighborhood of ARV = 0.20 – 0.25, and the neural networks had an ARV near 0.45. In comparison, the gated and ungated networks developed here had a 15 min. prediction ARV of 0.12 and 0.20, respectively.

4. Conclusion

Neural networks using various architectures and prefiltering algorithms have been created and compared in order to predict the activity level of the magnetosphere given an input of the solar wind. The problem reported in the past in which the output of the network is clipped is successfully resolved. The gated network which takes into account the varying input–output relation of the solar wind–ionosphere coupling performs significantly better on the *Bargatze et al.* [1985] data set when compared with an ungated network. The gated network also shows the most improvement at higher levels of activity where accurate predictions are of fundamental importance. We find that the gated neural network is a better 15 min. predictor than an equivalent ARMA system reported in *Hernandez et al.* [1993]. Also, it is shown that increasing the signal to noise ratio by preconditioning the data during training gives a network whose predictions are both more accurate for long range prediction and less sensitive to fluctuations between time steps.

Acknowledgments. We wish to acknowledge useful conversations with John Freeman. This work was supported by NSF Grant No. ATM-97262716 and NASA Grant NAGW-5193.

References

Bargatze, L.F., D.N. Baker, R.L., McPherron and E.W. Hones, Magnetospheric response to the IMF: substorms, *J. Geophys. Res., 90*, 6387, 1985.

Bishop, C.M., *Neural networks for pattern recognition*, Oxford University Press, 1995.

Casdagli, M., A dynamical systems approach to modeling input-output systems, In *Nonlinear Modeling and Forecasting*, M. Casdagli and S. Eubanks, eds., Addison-Wesley, 1992.

Detman and Vassiliadis, Review of Techniques for Magnetic Storm Forecasting, in Magnetic Storms, *Geophysical Monograph, 98*, American Geophysical Union, 1997.

Efron, B., Tibshirani, R., "An introduction to the bootstrap," *Monographs on Statistics and Applied Probability, 57*, Chapman and Hall, New York, 1993.

Hernandez, J.V., Tajima, T. and Horton, W., Neural net forecasting for geomagnetic activity, *Geophys. Res. Lett., 20*, 23, 2707, 1993.

Horton, W., I. Doxas, A low-dimensional dynamical model for the solar wind driven geotail-ionosphere system, *J. Geophys. Res., 103*, A3, 4561, 1998.

Horton, W., Pekker, M., Doxas, I. Magnetic energy storage and the nightside magnetosphere-ionosphere coupling, *Geophys. Res. Lett., 25*, 21, 4083, 1998.

Klimas, A.J., D. Vassiliadis, D.N. Baker, D.A. Roberts, The organized nonlinear dynamics of the magnetosphere, *J. Geophys. Res., 101*, 13,089, 1996.

Kulkarni, D.R., Pandya, A.S., Parikh, J.C., Modeling and predicting sunspot activity - state space reconstruction + artificial neural network methods, *Geophys. Res. Lett., 25*, 457, 1998.

Packard, N.H., J.P. Crutchfield, J.D. Farmer, and R.S. Shaw, Geometry from a time series, *Phys. Rev. Lett., 45*, 712, 1980.

Papadopoulos, K, J.G. Lyon, C.C. Goodrich, P.J. Cargill, A.S. Sharma, R. Kulkarni, C.L. Chang, A. Mankofsky, *Space Science Reviews, 71*, 671, 1995.

Price, C.P. and D. Prichard, The nonlinear response of the magnetosphere: October 30, 1978, *Geophys. Res. Lett., 20*, 771, 1993.

Ramamurti, V. and J. Ghosh, On the use of localized gating in mixture of experts networks, *Proc. SPIE Conf. on Applications of Science and Computational Intelligence*, SPIE Proc. Vol., Orlando, 1998.

Vassiliadis, D. A., A. J. Klimas, D. N. Baker, and D. A. Roberts, A description of solar wind-magnetosphere coupling based on nonlinear filters, *J. Geophys. Res., 99*, 3495, 1994.

Wray, J., and G.G.R. Green, Calculation of the Volterra kernels of nonlinear dynamic systems using an artificial neural network, *Biological Cybernetics, 71* (3), 187, 1994.

R.S. Weigel, W. Horton, and T. Tajima, Institute for Fusion Studies, The University of Texas, Austin, TX 78712 (e-mail: weigelr@kendall.ph.utexas.edu)

T. Detman, NOAA Space Environment Center, Boulder, CO 80303

(Received November 30, 1998; revised February 19, 1999; accepted March 1, 1999.)