Published paper

Boynton, R.J, Balikhin, M.A, Billings, S.A, Wei, H.L and Ganushkina, N (2011) *Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere.* Journal of Geophysical Research: Space Physics, 116. pp. 1-8.

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. ???, XXXX, DOI:10.1029/,

Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere

R. J. Boynton,[1] M. A. Balikhin,[1] S. A. Billings,[1] H. L. Wei,[1] N. Ganushkina,[2]

[1]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom.

[2]Finnish Meteorological Institute, Helsinki, Finland

Abstract.

The NARMAX OLS-ERR algorithm, which is widely used in the study of systems dynamics, is able to determine the causal relationship between the input and output variables for nonlinear systems. This technique has been applied to measurements of the solar wind from ACE at L1 and the D_{st} index in order to find the best solar wind-magnetosphere coupling function, i.e, which combination of solar wind parameters provide the best predictive capabilities of the D_{st} index. The data deduced coupling functions were then compared to those suggested in previous analytical and data based studies. The most appropriate coupling function was found to be $n^{1/2}V^{\alpha}B_T\sin^6(\theta/2)$, where the power of velocity, α, was inconclusive but should be in the range 2 - 3.

1. Introduction

A coupling function based on solar wind parameters, to predict the magnetospheric dynamics, has been sought since Chapman and Ferraro [1931]. These authors assumed the dynamic pressure would provide the best potential for forecasts. However, through measurements taken in the solar wind, it has been shown that on its own the solar wind pressure has a limited capability for predicting the magnetosphere dynamics [Crooker and Gringauz, 1993]. Dungey [1961] proposed that the magnetic merging between the interplanetary magnetic field (IMF) and the geomagnetic field would have a greater influence than the viscous forces on the dynamics of the magnetosphere. Hence the north-south component of the IMF, B_z, would provide a better predicting capability than the dynamic pressure. However, like dynamic pressure, on its own the north-south IMF does not have a large influence over the magnetosphere dynamics. Burton et al. [1975] introduced a half-wave rectifier, based on the dawn-dusk component of the interplanetary electric field that is set to be zero below a critical threshold. This value is effectively a product of the velocity and the southward component of the IMF, $I_B = VB_s$. Later Perreault and Akasofu [1978] suggested $\varepsilon = VB^2 \sin^4(\theta/2)$ which in contrast to I_B has a continuous dependence on the clock angle, $\theta = \tan^{-1}(B_y/B_z)$, of the IMF. The theoretical derivation of the ε parameter was addressed by Kan and Lee [1979]. Arguments based on dimensionality were used by Vasyliunas et al. [1982] to suggest $I_V = n^{1/6}V^{4/3}B_T G(\theta)$ and other coupling functions, where n is the density, G is a function of the clock angle and B_T is the tangential IMF, $B_T = \sqrt{B_y^2 + B_z^2}$. These analytically deduced coupling functions, in particular I_B, have often been used as inputs to forecasting data derived models [Klimas et al., 1996, Klimas et al., 1999, Balikhin et al., 2001, Boaghe et al., 2001 and Zhu et al., 2006].

Data based studies have also been devoted to the quest of determining the most appropriate coupling functions. Previous experimental studies were based on correlations between geomagnetic indices and combinations of solar wind parameters [Newell et al., 2007]. The correlation function indicates the linear dependence between data sets. Its application to nonlinear systems can be misleading. This can be illustrated by considering a simple example of a quadratic stochastic system with a zero mean input $X(t)$, shown in Figure 1 and output $Y(t) = X^2(t-1) + \zeta(t)$ shown in Figure 2, where $\zeta(t)$ are the noise and measurement errors and are assumed to have a zero mean. The correlation between X and Y is shown for 20 time lags in Figure 3. A lag at τ in Figure 3 represents the correlation coefficient between $Y(t)$ and $X(t-\tau)$. Even though $X(t-1)$ is the only input to the system, the correlation between Y and X is roughly zero for all of the time lags, including a time lag of one, which is the correlation coefficient between $Y(t)$ and $X(t-1)$. As a result the correlation between X and Y produces the misleading result that X does not have a causal relationship with Y, despite the known fact that X is the input and Y is the output for the simple quadratic system.

This example emphasizes that for nonlinear systems, only techniques that can take into account nonlinearities can be applied successfully. One possibility is to apply a methodology based on a nonlinear autoregressive moving average model with exogenous inputs (NARMAX) and an orthogonal least squares (OLS) algorithm [Leontaritis and Billings, 1985, Billings et al., 1989] to study the nonlinear dependences of the dynamics of the magnetosphere. In this approach the output at time t is a scalar value and is assumed to be a function of previous values of inputs $u(t)$, output $y(t)$ and error terms $e(t)$, as described by equation 1).

$$
\begin{aligned}
y(t) = & F[y(t-1), ..., y(t-n_y), \\
& u_1(t-1), ..., u_1(t-n_{u_1}), ..., \\
& u_m(t-1), ..., u_m(t-n_{u_m}), \\
& e(t-1), ..., e(t-n_e)] + e(t)
\end{aligned} \quad (1)
$$

where $F[\cdot]$ is some nonlinear function (e.g. polynomial, rational, B-Spline, radial basis function), y, u, and e are the output, input and error respectively, m is the number of inputs to the system and n_y, n_{u_1},...,n_{u_m} and n_e are the maximum time lags of the output, the m inputs and error respectively.

The NARMAX OLS-ERR methodology consists of three stages, namely model structure selection, parameter estimation and model validation. The model structure selection stage determines the most influential model terms by analyzing all possible cross-coupled combinations of past inputs and past outputs. The parameter estimation stage then determines the coefficients for each of the selected terms in the model. Finally the model validation stage justifies the final model. In this study, only the model structure selection stage will be used to identify the most significant solar wind-magnetosphere coupling functions.

From (1), the function $F[\cdot]$ can be taken as linear-in-the-parameters to a specified power. Therefore $F[\cdot]$ is a polynomial in which the monomials comprise of all the possible cross-coupled combinations of the components to the specified power. Then (1) becomes

$$y(t) = \theta_1 p_1(t) + \theta_2 p_2(t) + \cdots + \theta_M p_M(t) + e(t) \tag{2}$$

where $p_i(t)$ is the i^{th} monomial or regressor, θ_i is the coefficient of the i^{th} regressor and M is the total number of monomials. Equation (2) can be written using a scalar, $y(t)$, or by evaluating over the data to construct the vector notation, $\mathbf{y}$.

$$y(t) = \sum_{i=1}^{M} p_i(t)\theta_i + e(t) \qquad or \qquad \mathbf{y} = \mathbf{P}\theta + \mathbf{e} \tag{3}$$

where

$$\mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}, \mathbf{P} = \begin{bmatrix} p_1(1) & p_2(1) & \cdots & p_M(1) \\ p_1(2) & p_2(2) & \cdots & p_M(2) \\ \vdots & \vdots & & \vdots \\ p_1(N) & p_2(N) & \cdots & p_M(N) \end{bmatrix},$$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{bmatrix}$$

and N is the data length. The columns of the matrix $\mathbf{P}$ are made orthogonal to each other using the Gram-Schmidt procedure to give the matrix $\mathbf{W}$. The application of the Gram-Schmidt process to the columns of matrix $\mathbf{P}$ yields $\mathbf{y} = \mathbf{P}(\mathbf{R^{-1}R})\theta + \mathbf{e}$, so the matrix $\mathbf{W} = \mathbf{PR^{-1}}$ and vector $\mathbf{g} = \mathbf{R}\theta$, where

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1M} \\ 0 & 1 & r_{23} & \cdots & r_{2M} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & & \cdots & 1 & r_{(M-1)M} \\ 0 & & \cdots & 0 & 1 \end{bmatrix}$$

and is called the upper right triangular matrix. Since the columns of $\mathbf{W}$ are orthogonal, for $i \neq j$ the multiplication of the columns $\mathbf{w_i^T w_j} = 0$. (3) then becomes the auxiliary equation

$$y(t) = \sum_{i=1}^{M} w_i(t) g_i + e(t) \qquad or \qquad \mathbf{y} = \mathbf{Wg} + \mathbf{e} \tag{4}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1(1) & w_2(1) & \cdots & w_M(1) \\ w_1(2) & w_2(2) & \cdots & w_M(2) \\ \vdots & \vdots & & \vdots \\ w_1(N) & w_2(N) & \cdots & w_M(N) \end{bmatrix} \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_M \end{bmatrix}$$

here g_i is the i^{th} auxiliary coefficient of the i^{th} orthogonalized regressor, $w_i(t)$. An estimate of the auxiliary coefficients can be found from the fact that the multiplication of different columns of $\mathbf{W}$ equals zero. So $\mathbf{w_i^T w_j} = 0$ for $i \neq j$ where $\mathbf{w_i} = [w_i(1)\ w_i(2)\ \cdots\ w_i(N)]^T$ and $\mathbf{w_j} = [w_j(1)\ w_j(2)\ \cdots\ w_j(N)]^T$. Multiplying (4) by $\mathbf{w_n^T}$ yields

$$\mathbf{w_n^T y} = \left(\mathbf{w_n^T} \sum_{i=1}^{M} \mathbf{w_i} g_i + \mathbf{w_n^T e} \right) \quad (5)$$

For $\mathbf{w_n^T w_i}$ to be non-zero, n must equal i, leaving

$$\mathbf{w_i^T y} = \mathbf{w_i^T w_i} g_i + \mathbf{w_i^T e} \quad (6)$$

Assuming that the noise, $e(t)$, has a zero mean, is ergodic and is uncorrelated with all the regressors then $\mathbf{w_i^T e} = 0$. Equation 6 is now $\mathbf{w_i^T y} = \mathbf{w_i^T w_i} g_i$, which is an orthogonalized solution of minimizing the Least Squares. It should be noted that this method is for estimating the unknown coefficients in a linear regression model. Although the terms are nonlinear with respect to the inputs, the terms are linear-in-the-parameters (Equation 2). This enables the unknown i^{th} auxiliary coefficient, $\hat{g}_i$, to be estimated as

$$\hat{g}_i = \frac{\mathbf{w_i^T y}}{\mathbf{w_i^T w_i}} \quad (7)$$

The contribution to the dependent variable variance by each regressor can be found from multiplying $\mathbf{y^T}$ by (4)

$$\mathbf{y^T y} = \sum_{i=1}^{M} \left(g_i^2 \mathbf{w_i^T w_i} + g_i \mathbf{w_i^T e} + g_i \mathbf{e^T w_i} \right) + \mathbf{e^T e} \quad (8)$$

From the above mentioned properties of $e(t)$, $\mathbf{w_i^T e} = 0$ and $\mathbf{e^T e}$ represents the variance of the noise, σ_e^2, so that (8) may be rewritten as

$$\mathbf{y^T y} = \sum_{i=1}^{M} g_i^2 \mathbf{w_i^T w_i} + \sigma_e^2 \tag{9}$$

For the i^{th} regressor the dependent variable variance will be $g_i^2 \mathbf{w_i^T w_i}$. Dividing this by $\mathbf{y^T y}$ will determine the proportion of the dependent variable variance explained by the i^{th} regressor. This is called the error reduction ratio (ERR) and is defined by

$$[ERR]_i = \frac{\hat{g}_i^2 \mathbf{w_i^T w_i}}{\mathbf{y^T y}} \tag{10}$$

The algorithm was applied to the previous example of $Y(t) = X^2(t-1) + \zeta(t)$, using Y as the output, X as the input, 5 time lags for both input and output and a nonlinearity of degree of four. Thus the algorithm will search all the possible cross-coupled past output and input functions to the power of four, to determine the most influential model terms. The results in Table 1 show that the algorithm was able to determine the parameter X^2 as the most significant function, with an error reduction ratio of 99.93%.

The main goal of this study was to use the model structure selection procedure of the NARMAX OLS-ERR algorithm to identify the most significant solar wind-magnetosphere coupling function for the D_{st} index, i.e., determine which combination of solar wind parameters results in the best predictive capabilities of the D_{st} index. The algorithm is able to detect nonlinear dependencies on the output and thus assess the prediction capability of the coupling functions. In essence, the algorithm is used in a way similar to the application of the correlation function by other authors. However, unlike the correlation function which can only assess linear dependencies, the NARMAX algorithm is able to find nonlinear dependencies.

It should be noted that for different conditions the best coupling functions can vary. For example, in the case of a northern IMF, a viscous solar wind-magnetosphere interaction is expected. This should differ from the case of a purely southward IMF direction when reconnection is expected.

2. Data sets and methodology

Data from OMNI web, for the period from the start of 1998 to the end of 2008, have been used in this study. The hourly averaged solar wind data for the period occasionally has data gaps, which breaks the consecutive data into many sections. The NARMAX algorithm needs a continuous time series data set of about 1000 data points or greater. The initial 11-year data set was divided into 1000 point subsets and the data sets with data gaps were removed from this study. This procedure resulted in 64 continuous subsets. The NARMAX OLS-ERR algorithm was run for the 64 data sets, returning 64 models, each consisting of 20 model terms.

The 20 model terms, or coupling functions, were selected from all the possible cross-coupled combinations of the inputs, in the order of the functions ERR. The reason for limiting the NARMAX algorithm to select only the top 20 terms was to reduce the time taken for the algorithm to run. If the ERR were to be calculated by simply using (10) for every single candidate term, it may lead to an incorrect calculation of the ERR. This is because the ERR may depend on the order in which each candidate term enters the equation. Therefore orthogonalizing the candidate coupling functions into an orthogonal equation in the order in which the coupling functions happen to be written down may produce the wrong ERR. To prevent this from happening, the ERR is calculated by using a forward regression procedure [Billings et al., 1988].

The first step involves calculating the ERR of each of the M candidate function, so for $i = 1, \ldots, M$

$$\mathbf{w_1}^{(i)} = \mathbf{p_i}$$

$$g_1{}^{(i)} = \frac{(\mathbf{w_1}^{(i)})^T \mathbf{y}}{(\mathbf{w_1}^{(i)})^T \mathbf{w_1}^{(i)}}$$

$$[ERR]_1{}^{(i)} = \frac{g_1{}^{(i)2} (\mathbf{w_1}^{(i)})^T \mathbf{w_1}^{(i)}}{\mathbf{y}^T \mathbf{y}}$$

then the index of the function with the highest ERR is found

$$h_1 = \arg[\max\{[ERR]_1{}^{(i)}, 1 \le i \le M\}]$$

which is the index of the first and most significant function of the model, so $\mathbf{w_1} = \mathbf{w_1}^{(h_1)} = \mathbf{p_{h_1}}$.

The upper triangular matrix $\mathbf{R}$ is used so that the subsequent calculations are made in a subspace orthogonal to $\mathbf{w_1}$. To do this the elements of $\mathbf{R}$ are calculated in the subsequent steps. For the n^{th} step, to find the n^{th} most significant term, the ERR is calculated for each candidate function, apart from the $n-1$ functions already selected. So for $i = 1, \ldots, M$, where $i \ne h_1, i \ne h_2, \ldots, i \ne h_{n-1}$, the elements of $\mathbf{R}$ are calculated for $j = 1, \ldots, n-1$

$$r_{jn}{}^{(i)} = \frac{\mathbf{w_j}^T \mathbf{p_i}}{\mathbf{w_j}^T \mathbf{w_j}}$$

this can then be used to calculate $\mathbf{w_n}$ so that it will be orthogonal to all the other columns of $\mathbf{W}$.

$$\mathbf{w_n}^{(i)} = \mathbf{p_i} - \sum_{j=1}^{n-1} r_{jn}{}^{(i)} \mathbf{w_j}$$

$$g_n{}^{(i)} = \frac{(\mathbf{w_n}^{(i)})^T \mathbf{y}}{(\mathbf{w_n}^{(i)})^T \mathbf{w_n}^{(i)}}$$

$$[ERR]_n{}^{(i)} = \frac{g_n{}^{(i)2}(\mathbf{w_n}{}^{(i)})^T \mathbf{w_n}{}^{(i)}}{\mathbf{y}^T \mathbf{y}}$$

then the index of the function with the highest ERR is found as

$$h_n = \arg[\max\{[ERR]_n{}^{(i)}, 1 \leq i \leq M, i \neq h_1, \ldots, i \neq h_{n-1}\}]$$

So $\mathbf{p_{h_n}}$ will be the n^{th} most significant function of the model.

Calculating the ERR for every possible candidate function is computationally expensive and since the majority of the ERR is explained by the top functions, only 20 model terms or coupling functions were calculated.

The estimates of the regressor coefficients, $\hat{\theta}$, can then be computed backwards from the number of model terms that the algorithm is set to select, M_s, (in this case 20) using $\mathbf{g} = \mathbf{R}\hat{\theta}$.

$$\hat{\theta}_{M_s} = g_{M_s}$$

$$\hat{\theta}_i = g_i - \sum_{n=i+1}^{M_s} r_{in}\hat{\theta_n}$$

The previous hour value of $D_{st}(t-1)$ had substantially the highest ERR, in all the 64 models produced by the algorithm, with a mean ERR of 95.5% and a standard deviation of 2.13%. All the other terms had a much lower ERR in comparison. It is well known that the autoregressive term of $D_{st}(t-1)$ greatly helps in the prediction of the next value. It has been shown to have the highest ERR in other NARMAX studies of the D_{st} index, e.g. Boaghe et al. [2001]. The coefficient of the $D_{st}(t-1)$ term had an mean value of 1.126 for the first stage of the study with a standard deviation of 0.14. The small standard deviation of the $D_{st}(t-1)$ ERR and coefficient, indicates that the algorithm is consistent.

The aim of this study is to find the most significant solar wind magnetosphere coupling function to aid in the understanding of the underlying physics. Hence we do not wish to discuss the $D_{st}(t-1)$ term and therefore it will not be included in the results. However since it is an important part of the model, the term was kept in the algorithm as a candidate term to avoid any ill-conditioning. Instead of removing $D_{st}(t-1)$ as a candidate term, the ERR of all the candidate coupling functions were normalized using the difference between the whole ERR of each model and the ERR of the D_{st}. If the sum of every terms ERR is represented by W and the ERR of the $D_{st}(t-1)$ is represented by D, then the remaining ERR, or explained variable variance, is $W - D$. therefore, a candidate term, with an ERR of c, has a normalized ERR (NERR) of $c/(W - D) \times 100\%$. This NERR effectively yields the dependent variable variance of the output $D_{st}(t) - \alpha D_{st}(t-1)$, where α is the decay term and is automatically calculated for each model in the parameter estimation stage of the NARMAX algorithm.

In each stage of the study, the algorithm used 5 time lags of the output and inputs as candidate terms. The second to fifth lags of D_{st} were still kept as candidate functions, as well as the nonlinear coupling functions of all the past values of D_{st} (including $D_{st}(t-1)$), coupled with the inputs to the selected degree. The NERR for each function is then averaged over the 64 models displayed in Tables 2-5 to quantify the effects of a particular term in the evolution of the D_{st} index.

3. NERRs of second order basic solar wind parameters

It was anticipated that $I_B = VB_s$ by Burton et al. [1975], would be the coupling function with best predicting capability. This hypothesis influenced the choice of solar wind inputs and the degree at which the algorithm was set for the first stage of the study. The basic solar wind parameters used as the inputs to the algorithm were the solar wind velocity V, density n, dynamic pressure p, ion temperature T_i, the x, y, z components of the IMF in GSM coordinates B_x, B_y, B_z, and the z component of the IMF split into its north and south components B_n and B_s ($B_n = 0$ for southward IMF and $B_s = 0$ for northward IMF). The ion temperature has not been known to have any effect on the D_{st} index, because the solar wind thermal distribution is lost as it penetrates terrestrial bow shock. However, T_i was included to provide extra validation of the algorithm. Only nine solar wind parameters were used, due to the current limitation of the software. The degree of nonlinearity was limited to second order, due to the expectation of I_B having the best predicting capability.

Table 2 lists the top four terms in order of NERR and also shows how many times each individual term was selected.

The results show that the coupling function with the best predicting capability was, in fact, VB_s. Out of all the possible linear and quadratic cross-coupled combinations of the inputs, the half-wave rectifier was selected as the best input confirming our initial hypothesis. The second best coupling function was pB_s, and since $p = \frac{1}{2}nV^2$, this term also can be represented as a product of nV and VB_s and so it is effectively a fourth order nonlinear term. Since it has about half the NERR of I_B, this points to the fact that the limitation to a second order nonlinearity is too restrictive.

The epsilon function, $\varepsilon = VB^2 \sin^4(\theta/2)$, [Perreault and Akasofu, 1978] was added to the inputs, in place of the northward IMF, and the algorithm was run again. This was done to see how this coupling function would compare to the half-wave rectifier and other quadratic and linear terms. The epsilon function was selected as the fourth best function with a NERR of 3.89%.

The second lag of D_{st} was the third best and was selected in 51 out of the 64 models. As a result, it appears in more models than the VB_s function but has a lower NERR. This implies that the second lag of D_{st} must only have a minor influence in each of the 51 models that it appears in, compared to the top three functions which are selected less, but when selected have a much higher NERR and hence predicting capability.

4. Comparison of NERRs for previously proposed coupling functions

The aim of the second stage of the study was to differentiate between the predicting capabilities of previously proposed coupling functions. This aim was similar to the goal of Newell et al. [2007], however, the major difference is that a methodology appropriate for nonlinear systems is used in the present study. The NARMAX OLS-ERR algorithm can theoretically combine the solar wind parameters to reproduce these functions, however, as the order of nonlinearity of these coupling functions is high, up to 9, such a direct approach would be too computationally demanding.

The coupling functions used as inputs to the algorithm were similar to those chosen by Newell et al. [2007]. These were $I_B = VB_s$ [Burton et al., 1975], the epsilon parameter $\varepsilon = VB^2 \sin^4(\theta/2)$ [Perreault and Akasofu, 1978], $I_W = VB_T \sin^4(\theta/2)$ [Wygant et al., 1983], $I_{SR} = p^{1/2} VB_T \sin^4(\theta/2)$ [Scurry and Russell, 1991], $I_{TL} = p^{1/2} VB_T \sin^6(\theta/2)$ [Temerin and Li, 2006], $I_N = V^{4/3} B_T^{2/3} \sin^{8/3}(\theta/2)$ [Newell et al., 2007] and $I_V = n^{1/6} V^{4/3} B_T G(\theta)$ [Vasyliunas et al., 1982]. $\sin^4(\theta/2)$ was used as the clock angle function in I_V, $G(\theta)$. Unlike the previous section, the algorithm was set to have a degree of nonlinearity equal to one, implying a linear relationship between the inputs and output. The NERR was then used to assess the predicting capabilities of the coupling functions on the D_{st} index. The results for the top 5 coupling functions with highest NERR are presented in Table 3.

From Table 3, the coupling function I_{TL} [Temerin and Li, 2006] had the highest NERR, just over twice that of the next best function. The half wave rectifier, which was selected as the best term in the previous section, was the second best coupling function when competing with more complex functions. The third best function was I_V [Vasyliunas et al., 1982], using a linear dependence on B_T and $\sin^4(\theta/2)$ as the function of the clock angle and the I_{SR} function fourth. Both the I_V and I_{SR} functions are similar to the I_{TL} function. All three functions include the solar wind parameters of density, velocity, tangential IMF and a continuous function of the IMF clock angle. Since $n^{1/6}V^{1/3} = p^{1/6}$, the I_V, I_{SR} and I_{TL} functions only differ by the power of the pressure and the I_{TL} having a factor of $\sin^6(\theta/2)$ instead of $\sin^4(\theta/2)$. The second lag of D_{st} had the fifth highest NERR, again being selected in many of the models but only having a minor influence in each one.

5. NERR of solar wind parameters from the best coupling functions

In the first stage of the study, an arbitrary set of basic solar wind parameters, V, n, p, T_i, B_x, B_y, B_z, B_n and B_s, were used as inputs to the algorithm. However, the coupling functions deduced in previous studies contained combinations of parameters with fractional powers or high powers. Since NARMAX cannot detect fraction powers and the search for higher powers is computationally demanding, the parameters with a fractional or high power, can be used as an input to the NARMAX algorithm. The aim of the third stage of this study was very similar to the first, the only difference being that instead of an arbitrary set of basic solar wind parameters, the factors of the best coupling functions from Table 3, have been used as building blocks to assemble the most appropriate coupling function. These factors were the square root of the pressure $p^{1/2}$, the sixth root of the density $n^{1/6}$, the velocities V and $V^{4/3}$, the southward IMF B_s, and the tangential IMF with the different functions of clock angle $B_T \sin^4(\theta/2)$ and $B_T \sin^6(\theta/2)$. Table 4 shows the top five coupling functions with the highest NERR.

The two most significant coupling functions in Table 4 are $p^{1/2}V^{4/3}B_T \sin^6(\theta/2)$ and $p^{1/2}V^2B_T \sin^6(\theta/2)$, which differ only by the fractional power of velocity. The coupling function with the third highest NERR, also has a similar form that contains a density, velocity, tangential IMF and clock function but with different fractional powers. The second lag of D_{st} was found to have the fourth highest NERR, being selected in over half of the models despite only having a small NERR in each one. The function with the fifth highest NERR is also of a similar form to the top three functions. The results suggested that the coupling function is a factor of B_T, n to the power of between 1/6 and 1/2, most likely 1/2, $\sin(\theta/2)$ most likely to the power of 6 and V to the power of somewhere between 2 and 3, when considering the pressure to be composed of velocity and density.

In Table 4, the NERR of the best coupling functions are small, compared to those in Table 3. This is due to the algorithm having to select a model from more candidate model terms than those used in the previous section. The algorithm used in the previous section generated a total 46 candidate model terns, whilst in this section 3571 candidate model terms were generated. Since the latter study resulted in many more model terms, many of which were very similar to each other, a small ERR would be attributed to terms that have no influence on the output. This would cause the influential terms to have a reduced

ERR. Therefore it is not possible to directly compare the ERR from the two different runs of the algorithms, which significantly differ by the number of inputs or degree.

6. NERR of the best overall coupling function including the two most significant from the previous section

The aim of the last stage of this study was to confirm that the results of the previous section do indeed result in better coupling functions than those suggested in previous studies. To do this, the top two coupling functions determined in the previous section were included as inputs along with those used in Section 4. Thus the functions used as inputs to the algorithm were I_B, ε, I_W, I_V, I_{SR}, I_{TL}, $p^{1/2}V^{4/3}B_T\sin^6(\theta/2)$ and $p^{1/2}V^2B_T\sin^6(\theta/2)$.

Table 5 unexpectedly shows that the function $p^{1/2}V^2B_T\sin^6(\theta/2)$ was selected as the best coupling function, followed by $p^{1/2}V^{4/3}B_T\sin^6(\theta/2)$. The subsequent functions, I_{TL}, I_B and I_V, are in the same order as those in Table 3. The top three functions are very similar, only differing by the power of the velocity and again showing that the most appropriate coupling function should be composed of density, velocity, tangential IMF and clock angle factors.

7. Discussion and conclusions

The main objective of this study was to apply the model structure selection procedure from the NARMAX OLS-ERR system identification algorithm, to identify, directly from data, the best coupling functions that describe the solar wind-magnetosphere interaction, i.e., the combinations of solar wind parameters that provide the best predictive capabilities for the magnetospheric dynamics, that are related to the D_{st} index.

The physical processes that are involved in the solar-wind magnetosphere interaction depend upon the parameters of IMF and the solar wind. Two cases are expected to exist. The first occurs when the IMF is directed exactly northward, in the stable flow of the solar wind. In this case, no reconnection should take place and the main factor affecting the magnetosphere should be the viscous forces of the solar wind shearing against the magnetosphere. In the second case, of a exactly southward IMF regime, the merging between magnetic field lines of the IMF and geomagnetic field should be the dominant factor. The most appropriate coupling functions in these two extreme examples should be different. Without a comprehensive model, based on first principles, to describe the interaction between the solar wind and the magnetosphere, it is only possible to suggest mechanisms for the interaction of the two extreme cases mentioned above. It is not possible to determine the sensitivity of their interaction to changes in the IMF orientation and magnitude.

The initial data sets were arbitrary subdivided into 64 data intervals, without any information about the direction of the IMF and other solar wind parameters. Therefore the most appropriate coupling function for each IMF scenario would most likely not coincide with the data sets. In the future, different regimes of the magnetosphere will be studied in more detail, similar to the studies of Vassiliadis et al. [1999], Valdivia et al. [1996] and Newell et al. [2008].

It should be noted that it is a well known fact that the previous value of D_{st} already provides a fair estimate in the prediction of the one hour ahead D_{st} values. This was confirmed by the results of the NARMAX algorithm. In each of the 64 models produced by the algorithm, the $D_{st}(t-1)$ function had the highest ERR.

The coupling functions with the highest NERR in Tables 3, 4 and 5 only differ by the power of velocity. The majority of the coupling functions in these three Tables contain the parameters of density, velocity, tangential IMF and clock angle function. In the final section, the two functions in Table 4 with the highest NERR, which were found by the NARMAX algorithm, were used to determine how they would compete against the perviously proposed coupling functions. Table 5 shows these two coupling functions, $p^{1/2}V^2B_T\sin^6(\theta/2)$ and $p^{1/2}V^{4/3}B_T\sin^6(\theta/2)$, to have the highest NERR but in reverse order when compared to Table 4. This is followed by $I_{TL} = p^{1/2}VB_T\sin^6(\theta/2)$ [Temerin and Li, 2006], $I_B = VB_s$ [Burton et al., 1975] and $I_V = n^{1/6}V^{4/3}B_T\sin^4(\theta/2)$ [Vasyliunas et al., 1982]. Again, four of the five functions with the highest NERR are composed of a density, pressure, tangential IMF and a clock angle term. Therefore, according to these Tables, the most appropriate coupling functions should be of the form

$$n^\alpha V^\beta B_T^\gamma \sin^\delta\left(\frac{\theta}{2}\right) \tag{11}$$

From Tables 3 - 5, it can be seen that the function that has the highest NERR, have values of α, γ and δ equal to 1/2, 1 and 6 respectively. The value for β is inconclusive but should be in the range $2 - 3$.

With the presently available methodology, it is only possible to find the exact values of β by a time-consuming trial and error approach, to check various fractional numbers. In the future, we will develop a methodology for the automatic identification of both integer powers and fractional powers of the solar wind and IMF parameters. However, currently the only certain conclusion, which follows from Tables 3-5, is that coupling function should include the following factors, $n^{1/2}$, B_T and $\sin^6(\theta/2)$. The dependance on the velocity appears to include a power of somewhere between 2 and 3.

Three functions of the IMF clock angle, θ, were used throughout this study. These were the southward component ($-\cos(\theta)$ for $\pi/2 < \theta < 3\pi/2$ and zero for all other clock angles) and the functions $\sin^4(\theta/2)$ and $\sin^6(\theta/2)$. Although the three functions are very similar, the algorithm continuously selected the $\sin^6(\theta/2)$ function as the most appropriate for explaining the dependent variable variance of D_{st}, throughout the stages of the study. The relevance of the $\sin^6(\theta/2)$ function is discussed in more detail by Balikhin et al. [2010]

Although $D_{st}(t-1)$ was not included in our results, the second to fifth time lags of D_{st} were included. Nonlinear functions of D_{st} were also considered. In Section 3, all possible quadratic cross-coupled nonlinearities, that included the first to fifth time lags of D_{st}, were candidate functions, and in Section 5, all possible fourth order cross coupled nonlinearities, involving the first to fifth time lags of D_{st}, were considered as candidate functions. Out of all of these past D_{st} functions and nonlinear functions of D_{st}, only $D_{st}(t-2)$ appears to have a significant NERR, appearing as a top 5 function in Tables 2, 3 and 4. In each of these Tables, the second time lag of D_{st} is selected in the majority of the models but with only a minor influence. Therefore, although it is not a highly significant term, it should be included in any model.

Acknowledgments. The authors would like to acknowledge the financial support from EPSRC, STFC and ERC.

References

M. A. Balikhin, O. M. Boaghe, S. A. Billings, and H. S. C. K. Alleyne. Terrestrial magnetosphere as a nonlinear resonator. *Geophys. Res. Lett.*, 28(6):1123-1126, 2001.

M. A. Balikhin, R. J. Boynton, S. A. Billings, M. Gedalin, N. Ganushkina, D. Coca, and H. Wei. Data based quest for solar wind-magnetosphere coupling function. *Geophys. Res. Lett.*, 37(24):L24107–, 2010.

S. A. Billings, M. J. Korenberg and S. Chen,. Identification of non-linear output affine systems using an orthogonal least-squares algorithm. *Int. J. of Systems Sci.*, 19 :1559-1568, 1988.

S. A. Billings, S. Chen, and M. J. Korenberg. Identification of MIMO non-linear systems using a forward-regression orthogonal estimator *Int. J. Control*, 49(6):2157-2189, 1989.

O. M. Boaghe, M. A. Balikhin, S. A. Billings, and H. Alleyne. Identification of nonlinear processes in the magnetospheric dynamics and forecasting of dst index. *J. Geophys. Res.*, 106(A12):30,047-30,066, 2001.

R. K. Burton, R. L. McPherron, and C.T. Russull. An empirical relationship between interplanetary conditions and dst. *J. Geophys Res*, 80:4202-4214, 1975.

S. Chapman and V. C. A. Ferraro. A new theory of magnetic storms, part 1, the initial phase. *Terr. Magn, Atmos. Electr.*, 36:171, 1931.

N. U. Crooker and K. I. Gringauz. On the low correlation between long term averages of solar wind speed and geomagnetic activity after 1976. *J. Geophys. Res.*, 98:59, 1993.

J. W. Dungey. Interplanetary magnetic field and auroral zones. *Phys. Rev. Lett.*, 6:47, 1961.

J. R. Kan and L. C. Lee. Energy coupling function and solar wind-magnetosphere dynamo. *Geophys. Res. Lett.*, 6:577, 1979.

A. J. Klimas, D. Vassiliadis, D. N. Baker, and D. A. Roberts. The organized nonlinear dynamics of the magnetosphere. *J. Geophys. Res.*, 101(A6):13,089-13,113, 1996.

A. J. Klimas, D. Vassiliadis, D. N. Baker, and J. A. Valdivia. Data derived analogues of the Solar Wind-Magnetosphere Interaction. *Phys. Chem. Earth*, 24 :37 1999.

I. J. Leontaritis and S. A. Billings. Input-output parametric models for non-linear systems Part I: Deterministic non-linear systems. *Int. J. Control*, 41 (2):303-328, 1985.

P. T. Newell, T. Sotirelis, K. Liou, C. I. Meng, and F. J. Rich. A nearly universal solor wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *J. Geophys. Res.*, 112:A01206, 2007.

P. T. Newell, T. Sotirelis, K. Liou, and F. J. Rich. Pairs of solar wind-magnetosphere coupling functions: Combining a merging term with a viscous term works best *J. Geophys. Res.*, 113:A04218, 2008.

W. K. Perreault and S-I. Akasofu. A study of geomagnetic storms. *Geophys. J. R. Astrom. Soc.*, 54:547, 1978.

L. Scurry and C. T. Russell. Proxy studies of energy transfer to the magnetosphere. *J. Geophys. Res.*, 96(A6):9541-9548, 1991.

M. Temerin and X. Li. Dst model for 1995-2002. *J. Geophys. Res.*, 111:A04221, 2006.

J. A. Valdivia, A. S. Sharma, and K Papalopoudus. Prediction of magnetic storms by nonlinear models. *Geophys. Res. Lett.*, 23 :2899-2902, 1996.

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

D. Vassiliadis, A. J. Klimas, J. A. Valdivia, and D. N. Baker. The Dst geomagnetic response as a function of storm phase and amplitude and the solar wind electric field. *J. Geophys. Res.*, 104:24957-24976, 1999.

Function	ERR (%)
$X^2(t-1)$	99.93
$Y(t-1)$	0.027
$X^4(t-1)$	0.013

Table 1. Functions selected by the NARMAX OLS-ERR algorithm, for the example system $Y(t) = X^2(t-1) + \zeta(t)$, using Y as the output, X as the input, 5 time lags for both input and output and a nonlinearity of degree four

Coupling Function	NERR (%)	Selected
$VB_s(t-1)$	30.77	49
$pB_s(t-1)$	15.95	25
$D_{st}(t-2)$	5.47	51
$B_s(t-3)$	2.74	16

Table 2. Coupling functions selected by the OLS-ERR algorithm using 9 basic solar wind parameters as inputs, showing the NERR and the number of times each function was selected in a model

Coupling Function	NERR (%)	Selected
$p^{1/2}VB_T\sin^6(\theta/2)(t-1)$	31.32	51
$VB_s(t-1)$	12.76	40
$n^{1/6}V^{4/3}B_T\sin^4(\theta/2)(t-1)$	10.30	32
$p^{1/2}VB_T\sin^4(\theta/2)(t-1)$	8.37	31
$D_{st}(t-2)$	7.23	45

Table 3. Coupling functions selected by the OLS-ERR algorithm using the previously proposed coupling functions, showing the NERR and the number of times each function was selected in a model

R. J. Wygant, R. B. Torbert, and F. S. Mozer. Comparison of S3-3 polar cap potential drops with the interplanetary magnetic field and models of magnetopause reconnection. *J. Geophys. Res.*, 88(A7):5727-5735, 1983.

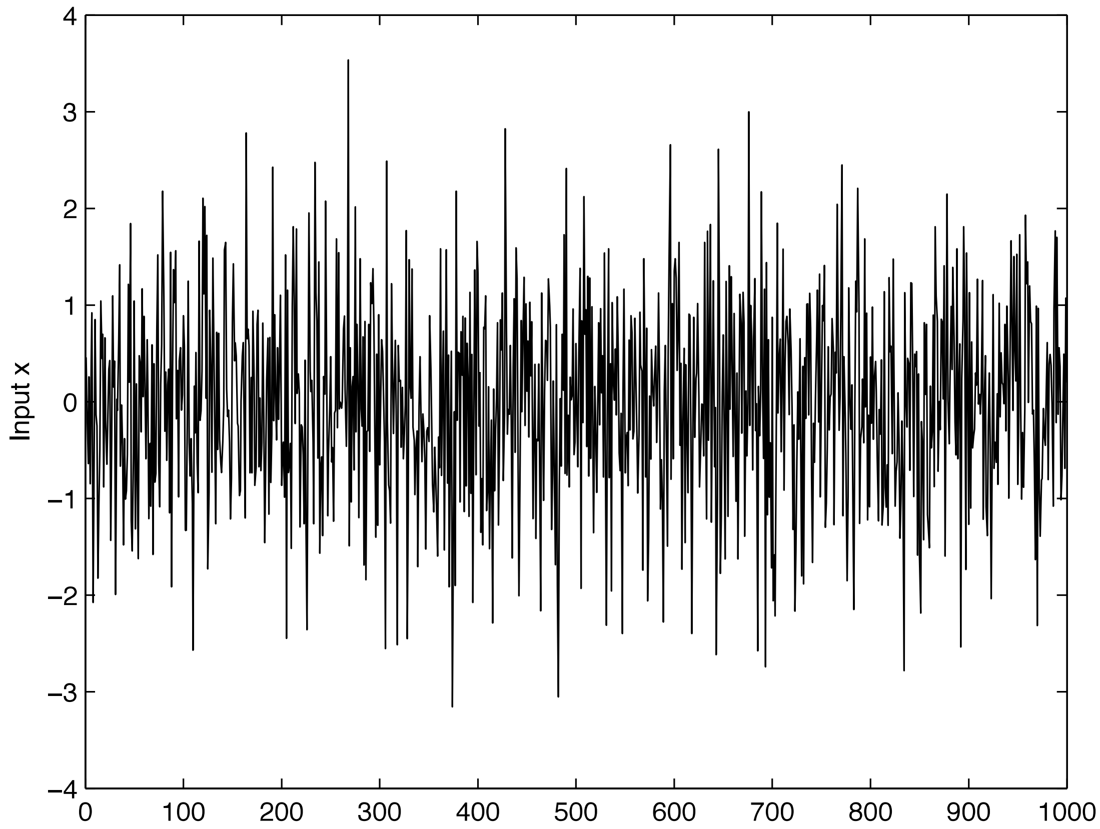
D. Zhu, S. A. Billings, M. A. Balikhin, S. Wing, and D. Coca. Data derived continuous time model for the Dst dynamics. *Geophys. Res. Lett.*, 33:L04101, 2006.

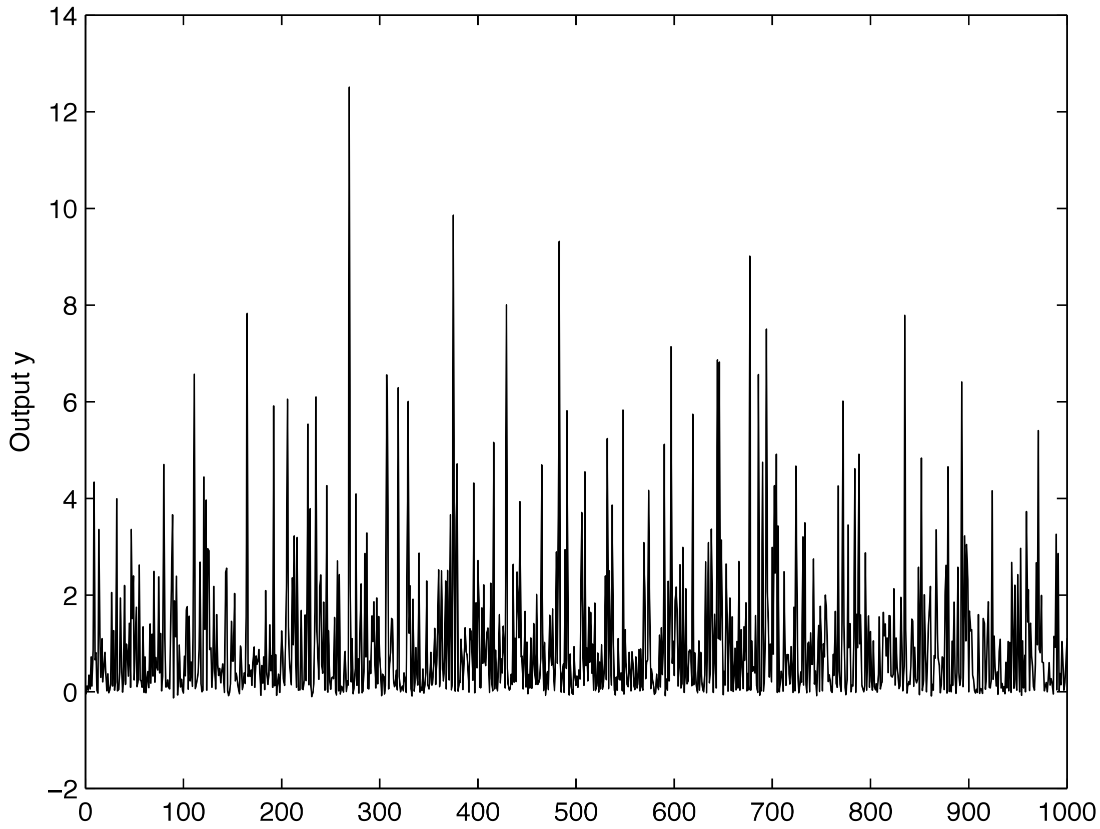
Coupling Function	NERR (%)	Selected
$p^{1/2}V^{4/3}B_T \sin^6(\theta/2)(t-1)$	5.46	7
$p^{1/2}V^2B_T \sin^6(\theta/2)(t-1)$	3.18	6
$n^{1/6}V^2B_T \sin^4(\theta/2)(t-1)$	3.15	4
$D_{st}(t-2)$	2.96	35
$p^{1/2}VB_T \sin^6(\theta/2)(t-1)$	2.77	4

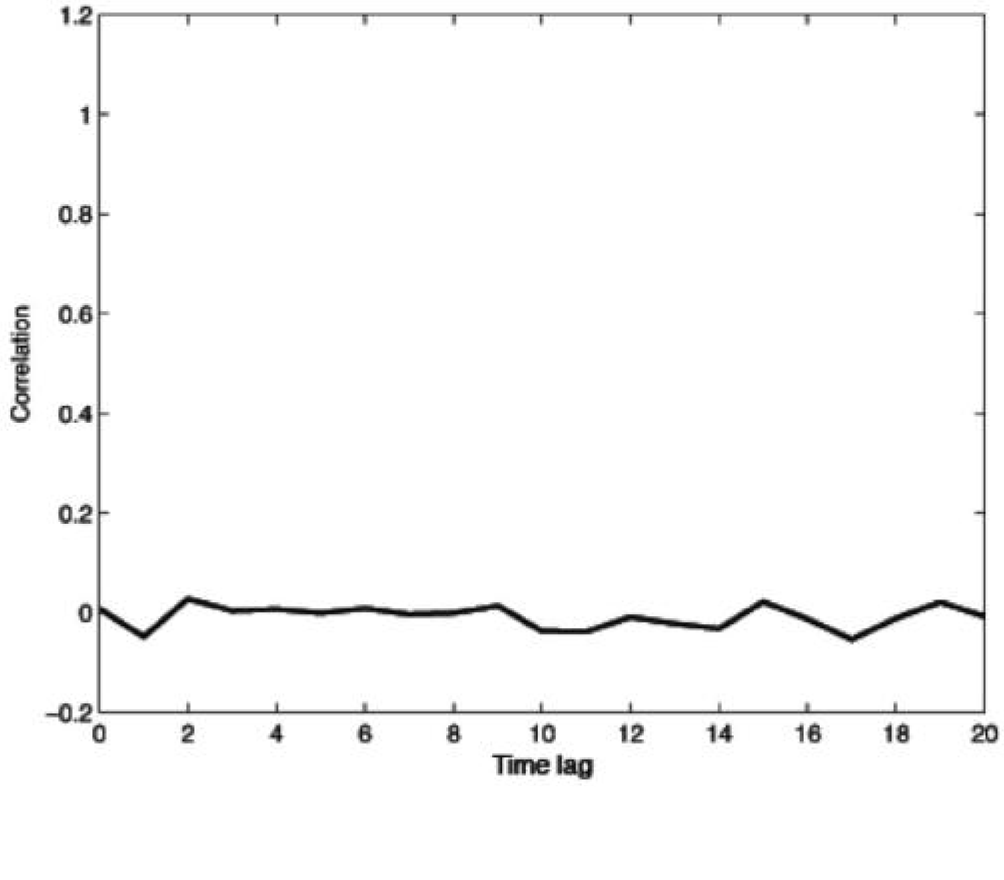
Table 4. Coupling functions selected by the OLS-ERR algorithm using the decomposed parameters from the best coupling functions, showing the NERR and the number of times each function was selected in a model

Coupling Function	NERR (%)	Selected
$p^{1/2}V^2B_T \sin^6(\theta/2)(t-1)$	14.0	39
$p^{1/2}V^{4/3}B_T \sin^6(\theta/2)(t-1)$	12.5	27
$p^{1/2}VB_T \sin^6(\theta/2)(t-1)$	12.1	34
$VB_s(t-1)$	8.91	41
$n^{1/6}V^{4/3}B_T \sin^4(\theta/2)(t-1)$	8.71	35

Table 5. Coupling functions selected by the OLS-ERR algorithm using the previously proposed coupling functions and best function from section 5, showing the NERR and the number of times each function was selected in a model


Input x
4
3
2
1
0
−1
−2
−3
−4
0
100
200
300
400
500
600
700
800
900
1000


Output y
14
12
10
8
6
4
2
0
−2
0
100
200
300
400
500
600
700
800
900
1000



1.2
1
0.8
0.6
0.4
0.2
0
−0.2
Correlation
0
2
4
6
8
10
12
14
16
18
20
Time lag