Monograph:
Wei, H.L., Zhu, D.Q., Billings, S.A. et al. (2006) Forecasting the geomagnetic activity of the Dst Index using radial basis function networks. Research Report. ACSE Research Report no. 941 . Automatic Control and Systems Engineering, University of Sheffield

Forecasting the Geomagnetic Activity of the *Dst* Index Using Radial Basis Function Networks

H. L. Wei, D. Q. Zhu, S. A. Billings, and M. A. Balikhin

Research Report No. 941

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street, Sheffield
S1 3JD, UK

14 September 2006

Forecasting the Geomagnetic Activity of the *Dst* Index Using Radial Basis Function Networks

H. L. Wei[a], D. Q. Zhu[b], S. A. Billings[c], and M. A. Balikhin[d]

[a,b,c,d]*Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, United Kingdom*

Email: { s.billings@sheffield.ac.uk , w.hualiang@sheffield.ac.uk , d.zhu@sheffield.ac.uk, m.balikhin@shef.ac.uk }

Abstract

The *Dst* index is a key parameter which characterises the disturbance of the geomagnetic field in magnetic storms. Modelling of the *Dst* index is thus very important for the analysis of the geomagnetic field. A data-based modelling approach, aimed at obtaining efficient models based on limited input-output observational data, provides a powerful tool for analysing and forecasting geomagnetic activities including the prediction of the *Dst* index. Radial basis function (RBF) networks are an important and popular network model for nonlinear system identification and dynamical modelling. A novel generalised multiscale RBF (MSRBF) network is introduced for *Dst* index modelling. The proposed MSRBF network can easily be converted into a linear-in-the-parameters form and the training of the linear network model can easily be implemented using an orthogonal least squares (OLS) type algorithm. One advantage of the new MSRBF network, compared with traditional single scale RBF networks, is that the new network is more flexible for describing complex nonlinear dynamical systems.

1. Introduction

The magnetosphere is a complex input-output dynamical nonlinear system, where the solar wind and the associated parameters play the role of the inputs and the geomagnetic indices can be considered as the outputs. The *Dst* index is an important parameter to measure the disturbance of the geomagnetic field in a magnetic storm. Several approaches have been proposed to study the dynamics of the magnetosphere under the influence of the solar wind, and the existing methods can broadly be classified into two categories: first-principle-based modelling (or similar methods) (Burton et al., 1975; Baker et al., 1990; Goertz et al., 1993; Klimas et al., 1998; Pulkkinen and Baker, 1997; O'Brien and McPherron, 2000), and data-based modelling (Hernandez et al., 1993; Vassiliadis et al., 1995, 1999; Takalo and Timonen; 1997; Wu and Lundsted, 1997; McPherron, 1999; Boaghe et al., 2001; Watanabe et al., 2002; Wei et al., 2004). Generally, first-principle-based modelling approaches require a comprehensive physical insight of all the associated macro and micro events jointly occurring in space weather dynamics. It is not always easy to obtain a comprehensive mathematical model based on first principles due to the difficulty of understanding the inherent mechanism of magnetosheric events such as storms and substorms which represent chains of complex physical processes. Data-based modeling, or system identification, which is an alternative to first-principle-based modeling, is thus required to provide a desirable means for forecasting and understanding the complex magnetospheric dynamics. In data-based modeling, the dynamical system is considered to be structure-unknown, and a mathematical representation for the underlying dynamics will be identified

from available observational data. In such a case, the solar wind parameters will be considered as inputs and the geomagnetic indices will be treated as the outputs of the magnetosphere system.

In the literature many types of model structures have been proposed for nonlinear dynamical systems identification, where the inner structure of the underlying system is unknown but only the input and output observational data are available. The NARMAX models, neural networks, radial basis function networks, neurofuzzy networks, and wavelet networks and wavelet multiresolution models are among the classes of the most popular model types (Leontaritis and Billings, 1985; Billings et al., 1989, 1998; Chen et al., 1990; Chen and Billings, 1992; Boaghe et al., 2001; Liu, 2001; Harris et al., 2002; Lundstedt et al., 2002; Wei et al, 2004a; Billings and Wei, 2005a, 2005b; Sharifi et al., 2006).

Radial basis function (RBF) networks, as a special class of single hidden-layer feedforward neural networks, have been proved to be universal approximators (Hartman et al., 1990; Poggio and Girosi, 1990; Park and Sandberg, 1991) for arbitrary nonlinear functions. One advantage of RBF networks, compared to multi-layer perceptrons (MLP), is that the linearly weighted structure of RBF networks, where parameters in the units of the hidden layer can often be pre-fixed, can easily be trained with a fast speed without involving nonlinear optimization. Another advantage of RBF networks, compared with other basis function networks, is that each basis function in the hidden units is a nonlinear mapping which maps a multivariable input to a scalar value, and thus the total number of candidate basis functions involved in a RBF network model is not very large and does not increase when the number of input variables increases.

This study aims to propose a new direct approach for identifying a mathematical model for the magnetospheric dynamics without any a priori information of the physical processes of the magnetosphere system but only a limited observational data. To achieve this objective, a novel class of RBF networks is introduced to represent the underlying dynamics of the magnetosphere system. Unlike a conventional single scale (kernel width) RBF, where all the basis functions have a common single scale, or each basis function has a single individual scale, the new RBF network uses a number of multiscale basis functions, where each basis function has multiple scale parameters (kernel widths). The new network will be referred to as the *multiscale RBF network* (MSRBF). The construction procedure of such a MSRBF network is as follows. The positions (centres) of the basis functions in the MSRBF network are initially pre-clustered and selected using some unsupervised clustering algorithm say the *k*-means clustering method. For each selected centre, the associated scales (kernel widths) are determined heuristically, and the selected centres and scales are restricted to a fixed grid. Finally, an MSRBF network is converted into a linear-in-the-parameters model form. A forward orthogonal regression (FOR) algorithm (Billings et al., 1989; Chen et al. 1989; Wei et al., 2006), regularised by a Bayesian information criterion (BIC) (Schwarz, 1978; Efron and Tibshirani, 1993), is then used to train the MSRBF network, and a parsimonious model, which consists of a relatively small number of regressors, is then used to predict the *Dst* index.

2. The linear-in-the-parameters representation

Consider the identification problem of a single-input and single-output (SISO) nonlinear dynamical system, for which N pairs of input-output observations, $\{u(t), y(t)\}_{t=1}^{N}$, are available. Under some mild conditions a discrete-time nonlinear system can be described by the following NARX model (Leontaritis and Billings, 1985)

$$y(t) = f(y(t-1),\cdots, y(t-n_y), u(t-1),\cdots, u(t-n_u)) + e(t) \tag{1}$$

where $u(t)$, $y(t)$ and $e(t)$ are the system input, output and noise variables; n_u and n_y are the maximum lags in the input and output, respectively; and f is a nonlinear mapping that is in general unknown and needs to be identified from the available observations. It is generally assumed that $e(t)$ is an independent identical distributed noise sequence.

The central task of system identification is to find an efficient approximator $\hat{f}$ for the nonlinear function *f* from the observational data. Several model types can be used to approximate the nonlinear function *f* and different model types often involve totally different training/learning strategies. One of the most commonly used methods is to approximate the nonlinear function *f* using a series of specified basis functions, whose local and global properties are known. One advantage of the basis function approximation is that the expression can easily be converted into a linear-in-the-parameters form, which is an important class of representations for nonlinear function approximation and signal processing. Compared to nonlinear-in-the-parameters models that usually involve complex and time-consuming nonlinear optimisation, linear-in-the-parameters models are simpler to analyse and quicker to compute and estimate.

Let $d = n_y + n_u$ and $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]$ with

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y + 1 \le k \le n_y + n_u \end{cases} \tag{2}$$

A general form of the linear-in-the-parameters regression model is given as

$$y(t) = \hat{f}(\mathbf{x}(t)) + e(t) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) + e(t) = \boldsymbol{\varphi}^T(t)\boldsymbol{\theta} + e(t) \tag{3}$$

where *M* is the total number of candidate regressors, $\phi_m(\mathbf{x}(t))$ (*m*=1,2, …, *M*) are the model regressors and θ_m are the model parameters, and $\boldsymbol{\varphi}(t) = [\phi_1(\mathbf{x}(t)), \cdots, \phi_M(\mathbf{x}(t))]^T$ and $\boldsymbol{\theta}$ are the associated regressor vector and parameter vector respectively.

In the present study, a new multiscale RBF (MSRBF) network model with Gaussian kernels will be used to construct the approximator $\hat{f}$, and this is discussed in the next section.

3. Multiscale RBF networks

The multiscale RBF (MSRBF) network aims to accommodate both the local and the global properties of the basis functions by including both small and large scales (kernel widths) in the network in a hierarchical multiscale way. In the multiscale modelling framework, a set of scale parameters (multiple kernel widths) will be assigned to each basis function.

3.1 The Network Structure

Taking the case of single-input and single-output nonlinear dynamical systems as an example, the MSRBF network possesses the following structure

$$\hat{f}(\mathbf{x}(t)) = \sum_{k=1}^{d} \theta_k^{(\text{linear})} x_k(t) + \sum_{i=0}^{I} \sum_{j=0}^{J} \sum_{m=1}^{N_c} \theta_{i,j,m}^{(\text{RBF})} \varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)}) \tag{4}$$

where $\theta_k^{(\text{linear})}$ and $\theta_{i,j,m}^{(\text{RBF})}$ are constants (unknown parameters), $\varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)})$ is the *m*th Gaussian basis function of the form

$$\varphi_{i,j,m}(\mathbf{x}(t); \mathbf{c}_m, \mathbf{s}_m^{(i,j)}) = \exp\left[-\sum_{k=1}^{n_y} \left(\frac{x_k(t) - c_{m,k}}{s_{y,m}^{(i)}} \right)^2 - \sum_{k=n_y+1}^{n_y+n_u} \left(\frac{x_k(t) - c_{m,k}}{s_{u,m}^{(j)}} \right)^2 \right] \tag{5}$$

where $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]$, defined by (2), is the network input vector, $\mathbf{c}_m(t) = [c_{m,1}, \cdots, c_{m,d}]$ is the centre of the *m*th basis function, and the scale vector $\mathbf{s}_m^{(i,j)}$ for the *m*th basis function in the network is defined as

$$\mathbf{s}_m^{(i,j)} = [\underbrace{s_{y,m}^{(i)}, \cdots, s_{y,m}^{(i)}}_{1:n_y}, \underbrace{s_{u,m}^{(j)}, \cdots, s_{u,m}^{(j)}}_{1:n_u}] \tag{6}$$

The number of the basis functions (or the number of centres) in the network is N_c, the number of scales for the output and input variables in the *m*th basis function is (*I*+1) and (*J*+1) respectively. Thus, for a single-input and single-output system, the network involves a total of $M = (I+1)(J+1)N_c$ basis functions.

All given observations can be considered as candidate kernel centres providing that the observational data set is not very long. For a long data set, some unsupervised learning algorithms can be used to locate the centres of the basis functions in only those regimens of the input space where significant data are present, and supervised learning approaches can then be used to train the network further. The details for the determination of the centres $\mathbf{c}_m$ and the kernel widths $\mathbf{s}_m^{(i,j)}$ are given below.

3.2 Determine the centres

If the observed data set is not very long, all given observations can be considered as candidate kernel centres $\mathbf{c}_m$. If, however, a long data set is involved and all the observations are still considered as candidate kernel centres, the initial MSRBF network will then include a great number of model terms and the training of the network will be time consuming. To overcome this problem, the well-known k-means clustering algorithm (Duda et al., 2001), coupled with the sum-of-squares criterion proposed by Krzanowski and Lai (1988), can be used to significantly reduce the number of candidate centres of the basis functions in the network. The sum-of-squares clustering algorithm is briefly described as below.

Assume that the data are given in the form of a matrix $\mathbf{X}$ of size $N \times p$, with the *i*th row given by the vector $\mathbf{z}_i = [z_{i1}, \cdots, z_{ip}]$ representing the observation vector of the *i*th object. The given *N* observations can be partitioned into *k* groups (clusters), denoted by $G_1, G_2, \cdots, G_k$, where *k* is an arbitrary integer between 1 and *N*. Let N_j be the number of objects that fall into the *j*th group G_j, and I_j the indices of the N_j observations in G_j. Define $W_k = \sum_{j=1}^{k} d_j$, with

$$d_j = \frac{1}{N_j} \sum_{i_1, i_2 \in I_j} (\mathbf{z}_{i_1} - \mathbf{z}_{i_2})(\mathbf{z}_{i_1} - \mathbf{z}_{i_2})^T \tag{7}$$

To choose an appropriate value for *k*, to determine the number of clusters, Krzanowski and Lai (1988) suggested the following criterion

$$\mathrm{DIFF}(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k \tag{8}$$

and the optimal value of *k* is the value that maximise the statistic below

$$\mathrm{KL}(k) = \left| \frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)} \right| \tag{9}$$

The above sum-of-squares clustering algorithm can be used to select the number of centres for the MSRBF network. For a given training data set of length *N*, let $N_c = \arg\max_k \{\mathrm{KL}(k)\}$. The MSRBF network will thus involve at least N_c (generally $N_c << N$) candidate centres, which can be determined using any *k*-means clustering algorithms.

3.3 Determine the scales

For given *N* pairs of input-output observations, $\{u(t), y(t)\}_{t=1}^N$, let σ_u and σ_y be the standard derivation of $\{u(t)\}_{t=1}^N$ and $\{y(t)\}_{t=1}^N$, respectively. The scale vector (6) can be chosen as

$$s_{y,m}^{(i)} = \beta\alpha^{-i}\sigma_y,\ i=0,\ 1,\ldots,\ I, \tag{10}$$

$$s_{u,m}^{(j)} = \beta\alpha^{-j}\sigma_u,\ j=0,\ 1,\ \ldots,\ J, \tag{11}$$

where m=1,2, …, N_c, and $\alpha > 1$ and $\beta > 1$ are two constants. From our experience, a good choice for the constants α and β is to set $\alpha = 2$ and $1 \le \beta \le 3$. Let

$$\mathcal{D}_3 = \{\varphi_{i,j,m}(\cdot;\mathbf{c}_m,\mathbf{s}_m^{(i,j)}) : i = 0,\cdots,I;\ j = 0,\cdots,J;\ m = 1,\cdots,N_c\} \tag{12}$$

The triple-indexed set D_3 is referred to as the dictionary associated with the new MSRBF networks. For the sake of convenience in the descriptions, rearrange the elements of $\mathcal{D}_3$ so that the triple index (i, j, m) can be indicated by a single index m=1,2,…, M, where $M = (I+1)(J+1)N_c$, to form a single indexed dictionary $\mathcal{D}_1 = \{\phi_m(\cdot) : \phi_m \in \mathcal{D}_3, m = 1,\cdots,M\}$. In this study, the two types of dictionaries $\mathcal{D}_1$ and $\mathcal{D}_3$ will not be distinguished, and a uniform symbol $\mathcal{D}$ will be used to indicate both of the two dictionaries. The network (4) can then be expressed as

$$\hat{f}(\mathbf{x}(t)) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) \tag{13}$$

The derivations given in this section can easily be extended to multiple-input and multiple-output (MIMO) situations, including the two-input and single-output case described in section 5.

4. Model term selection and the forward orthogonal regression (FOR) algorithm

The MSRBF network (13) may involve a great number of candidate model terms (regressors) when the parameters *I*, *J*, and N_c are large. Many of these candidate model terms, however, may be redundant. The inclusion of redundant model terms often makes the model become oversensitive to the training data and is likely to exhibit poor generalisation properties. It is thus important to determine which terms should be included in the model. In the present study, a forward orthogonal regression (FOR) algorithm (Billings et al., 1989; Chen et al. 1989; Wei et al., 2006), regularised by a Bayesian information criterion (BIC) (Schwarz, 1978; Efron and Tibshirani, 1993), is used to solve the model structure detection problem for the MSRBF network models. Following Billings et al. (1989) and Chen et al. (1989), a squared correlation coefficient will be used to measure the dependency between two associated random vectors. The squared correlation coefficient between two given vectors **x** and **y** of size *N* is defined as

$$C(\mathbf{x},\mathbf{y}) = \frac{(\mathbf{x}^T\mathbf{y})^2}{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})} = \frac{(\sum_{i=1}^{N} x_i y_i)^2}{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2} \tag{14}$$

It has been shown in Wei et al. (2004b) that the above squared correlation coefficient is closely related to the error reduction ratio (ERR) criterion (a very useful index to indicate the significance of model terms), defined in the standard orthogonal least squares (OLS) algorithm for model structure selection (Billings et al.1989, Chen et al. 1989).

4.1 The Forward Orthogonal Regression (FOR) Algorithm

Let $\mathbf{y} = [y(1), \cdots, y(N)]^T$ be a vector of measured outputs at N time instants, and $\boldsymbol{\varphi}_m = [\phi_m(1), \cdots, \phi_m(N)]^T$ be a vector formed by the mth candidate model term, where m=1,2, …, M. Let $\mathcal{D} = \{\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_M\}$ be a dictionary composed of the M candidate bases. From the viewpoint of practical modelling and identification, the finite dimensional set $\mathcal{D}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $\mathcal{D}_n = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\} = \{\boldsymbol{\varphi}_{i_1}, \cdots, \boldsymbol{\varphi}_{i_n}\}$ of n ($n \le M$) bases, from the library $\mathcal{D}$, where $\boldsymbol{\alpha}_k = \boldsymbol{\varphi}_{i_k}$, $i_k \in \{1,2,\cdots,M\}$ and k=1,2, …, n, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$ as below

$$\mathbf{y} = \theta_1\boldsymbol{\alpha}_1 + \cdots + \theta_n\boldsymbol{\alpha}_n + \mathbf{e} \tag{15}$$

or in a compact matrix form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \tag{16}$$

where the matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n]$ is assumed to be of full column rank, $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is the approximation error.

The model structure selection procedure starts from equation (13). Let $\mathbf{r}_0 = \mathbf{y}$, and

$$\ell_1 = \arg\max_{1 \le j \le M} \{C(\mathbf{y}, \boldsymbol{\varphi}_j)\} \tag{17}$$

where the function $C(\cdot,\cdot)$ is the correlation coefficient defined by (14). The first significant basis can thus be selected as $\boldsymbol{\alpha}_1 = \boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal basis can be chosen as $\mathbf{q}_1 = \boldsymbol{\varphi}_{\ell_1}$. Set

$$\mathbf{r}_1 = \mathbf{r}_0 - \frac{\mathbf{y}^T\mathbf{q}_1}{\mathbf{q}_1^T\mathbf{q}_1}\mathbf{q}_1 \tag{18}$$

At the second step, let $\mathbf{q}_j^{(2)} = \boldsymbol{\varphi}_j - [(\boldsymbol{\varphi}_j^T\mathbf{q}_1)/(\mathbf{q}_1^T\mathbf{q}_1)]\mathbf{q}_1$, where $\boldsymbol{\varphi}_j \in \mathcal{D}$ and $j \ne \ell_1$. Define

$$\ell_2 = \arg\max_{j \ne \ell_1} \{C(\mathbf{y}, \mathbf{q}_j^{(2)})\} \tag{19}$$

The second significant basis can thus be chosen as $\boldsymbol{\alpha}_2 = \boldsymbol{\varphi}_{\ell_2}$, and the second associated orthogonal basis can be chosen as $\mathbf{q}_2 = \mathbf{q}_{\ell_2}^{(2)}$. Set

$$\mathbf{r}_2 = \mathbf{r}_1 - \frac{\mathbf{y}^T\mathbf{q}_2}{\mathbf{q}_2^T\mathbf{q}_2}\mathbf{q}_2 \tag{20}$$

In general, the *m*th significant model term can be chosen as follows. Assume that at the (*m*-1)th step, a subset $\mathcal{D}_{m-1}$, consisting of (*m*-1) significant bases, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{m-1}$, has been determined, and the (*m*-1) selected bases have been transformed into a new group of orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ via some orthogonal transformation. Let

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{q}_k}\mathbf{q}_k \tag{21}$$

$$\ell_m = \arg \max_{j \neq \ell_k, 1 \leq k \leq m-1} \{C(\mathbf{y}, \mathbf{q}_j^{(m)})\} \tag{22}$$

where $\boldsymbol{\varphi}_j \in \mathcal{D} - \mathcal{D}_{m-1}$, and $\mathbf{r}_{m-1}$ is the residual vector obtained in the (*m*-1)th step. The *m*th significant basis can then be chosen as $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ and the *m*th associated orthogonal basis can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$. The residual vector $\mathbf{r}_m$ at the *m*th step is given by

$$\mathbf{r}_m = \mathbf{r}_{m-1} - \frac{\mathbf{y}^T\mathbf{q}_m}{\mathbf{q}_m^T\mathbf{q}_m}\mathbf{q}_m \tag{23}$$

Subsequent significant bases can be selected in the same way step by step. From (23), the vectors $\mathbf{r}_m$ and $\mathbf{q}_m$ are orthogonal, thus

$$\| \mathbf{r}_m \|^2 = \| \mathbf{r}_{m-1} \|^2 - \frac{(\mathbf{y}^T\mathbf{q}_m)^2}{\mathbf{q}_m^T\mathbf{q}_m} \tag{24}$$

By respectively summing (23) and (24) for *m* from 1 to *n*, yields

$$\mathbf{y} = \sum_{m=1}^{n} \frac{\mathbf{y}^T\mathbf{q}_m}{\mathbf{q}_m^T\mathbf{q}_m}\mathbf{q}_m + \mathbf{r}_n \tag{25}$$

$$\| \mathbf{r}_n \|^2 = \| \mathbf{y} \|^2 - \sum_{m=1}^{n} \frac{(\mathbf{y}^T\mathbf{q}_m)^2}{\mathbf{q}_m^T\mathbf{q}_m} \tag{26}$$

The residual sum of squares, $\| \mathbf{r}_n \|^2$, which is also known as the sum-squared-error, or its variants, can be used to form criteria for model selection. Note that the quantity $\mathrm{ERR}_m = C(\mathbf{y}, \mathbf{q}_m)$ is just equal to the error reduction ratio (Billings et al., 1989; Chen et al., 1989), brought by including the *m*th basis vector $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ into the model, and that $\sum_{m=1}^{n} C(\mathbf{y}, \mathbf{q}_m)$ is the increment or total percentage that the desired output variance can be explained by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$.

The model term selection procedure can be terminated when some specified termination conditions are met. In the present study, the following Bayesian information criterion (BIC) (Schwarz, 1978; Efron and Tibshirani, 1993) is used to determine the model size

$$\mathrm{BIC}(n) = \frac{N + n[\ln(N) - 1]}{N - n}\mathrm{MSE}(n) = \left[1 + \frac{n\ln(N)}{N - n}\right]\frac{\| \mathbf{r}_n \|^2}{N} \tag{27}$$

The mean-squared-error (MSE) in (27) is defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{t=1}^{N}[y(t) - \hat{y}(t)]^2 \tag{28}$$

where $\hat{y}(t)$ is the model prediction (one-step ahead) produced from the associated n term model. The selection procedure will be terminated at the step where the index function BIC(n) is minimized.

4.2 Parameter estimation

It is easy to verify that the relationship between the selected original bases $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$, and the associated orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n$, is given by

$$\mathbf{A}_n = \mathbf{Q}_n \mathbf{R}_n \tag{29}$$

where $\mathbf{A}_n = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n]$, $\mathbf{Q}_n$ is an $N \times n$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n$, and $\mathbf{R}_n$ is an $n \times n$ unit upper triangular matrix whose entries $u_{ij}(1 \le i \le j \le n)$ are calculated during the orthogonalization procedure. The unknown parameter vector, denoted by $\boldsymbol{\theta}_n = [\theta_1, \theta_2, \cdots, \theta_n]^T$, for the model with respect the original bases, can be calculated from the triangular equation $\mathbf{R}_n \boldsymbol{\theta}_n = \mathbf{g}_n$ with $\mathbf{g}_n = [g_1, g_2, \cdots, g_n]^T$, where $g_k = (\mathbf{y}^T \mathbf{q}_k)/(\mathbf{q}_k^T \mathbf{q}_k)$.

5. *Dst* index modelling and forecasting

In this study, the magnetosphere system was considered to be a structure-unknown (black-box) dynamical system. The objective was to identify a mathematical model that can be used to characterise and predict the activity of the *Dst* index. Previous studies have shown that the *Dst* index is mainly affected by two factors: the solar wind parameter, *VBs*, and the solar wind dynamical pressure, *P*. In the modelling procedure, the magnetosphere system was thus treated to be a two-input and single output system, where the *Dst* index was the system output, and the solar wind parameter *VBs* and the solar wind dynamical pressure *P* were the system inputs. Figure 1 shows 1000 data points of measurements of the *Dst* index (output, in unit of 'nT'), the solar wind parameter VB_s (input, in unit of 'mV/m'), and the solar wind dynamical pressure *P* (input, in unit of 'nPa'), with a sampling interval *T*=1hour. This data set was used for model estimation and another separate data set with 600 data points, measured in another different period, was used for model performance test.

For convenience of description, let $y(t) = Dst(t)$, $u_1(t) = VBs(t)$, and $u_2(t) = P(t)$. Eleven significant variables, $y(t-i)$ (i=1,2,3), $u_1(t-j)$ (j=1,2,3,4), and $u_2(t-k)$ (k=1,2,3,4), were chosen initially using a variable selection procedure (Wei et al., 2004b). The input vector for the MSRBF network model was then chosen to be $\mathbf{x} = [x_1(t), x_2(t), \cdots, x_{11}(t)] = [y(t-1), \cdots, y(t-3), u_1(t-1), \cdots, u_1(t-4), u_2(t-1), \cdots, u_2(t-4)]$.

For the *Dst* index related data, numerical experiments showed that it was difficult to train a standard single scale Gaussian kernel based RBF network using the original measurement data. In fact, many different kernel widths have been tested, trying to construct a standard network model using only a single common kernel width, but all the resulting models failed to provide effective representations for the data. The proposed multiscale modelling framework, however, can be used to describe this data set.

The sum-of-squares clustering algorithm was applied to the training data set and the KL index defined by (9) suggested that the optimal number of clusters for this data set was 34 (see Fig. 2). The 1000 data points were thus partitioned into 34 groups, and the centres of the 34 groups were chosen as the candidate centres for constructing the MSRBF network. The basis functions in the MSRBF network model (4) were of the form

$$\varphi_{i,j,m}(\mathbf{x}(t);\mathbf{c}_m,\mathbf{s}_m^{(p,q,r)}) = \exp\left[-\sum_{k=1}^{3}\left(\frac{x_k(t)-c_{m,k}}{s_{y,m}^{(p)}}\right)^2 - \sum_{k=4}^{7}\left(\frac{x_k(t)-c_{m,k}}{s_{u1,m}^{(q)}}\right)^2 - \sum_{k=8}^{11}\left(\frac{x_k(t)-c_{m,k}}{s_{u2,m}^{(r)}}\right)^2\right] \quad (30)$$

where m=1,2, …, 34, and the mth scale vector $\mathbf{s}_m^{(p,q,r)}$ is given as

$$\mathbf{s}_m^{(p,q,r)} = [\underbrace{s_{y,m}^{(p)},\cdots,s_{y,m}^{(p)}}_{1:3},\underbrace{s_{u1,m}^{(q)},\cdots,s_{u1,m}^{(q)}}_{1:4},\underbrace{s_{u2,m}^{(r)},\cdots,s_{u2,m}^{(r)}}_{1:4}] \quad (31)$$

and the parameters $s_{y,m}^{(p)}$, $s_{u1,m}^{(q)}$ and $s_{u2,m}^{(r)}$ were chosen as follows:

i) $\sigma_y \approx 20$, $\sigma_{u1} \approx 1$, and $\sigma_{u2} \approx 2$.

ii) $s_{y,m}^{(p)} = \beta 2^{-p}\sigma_y$ (p=1,2, 3), $s_{u1,m}^{(q)} = \beta 2^{-q}\sigma_{u1}$ (q=0,1,2), and $s_{u2,m}^{(r)} = \beta 2^{-r}\sigma_{u2}$ (r=0,1,2), with β =2.

Thus, the initial network model involves a total of $11+3^3\times 34 = 929$ candidate model terms (regressors). The forward orthogonal regression (FOR) algorithm was applied to the 929 candidate model terms, over the 1000 training data points, and 13 significant regressors were selected according to the value of BIC, which is shown in Fig. 3. The 13 regressors were used to form the final RBF network model that was used for the *Dst* index prediction.

Figures 4(a) presents one-step-ahead (OSA) predictions, over a typical storm period, and the value of MSE for OSA predictions was calculated to be 9.6768. Figure 4(b) presents the long-term predictions (model predicted outputs, MPO). It is clear from Fig. 4 that the identified MSRBF network model provides very good predictions for the *Dst* index, even during a period of a storm (*Dst* is near to or less than -100nT).

6. Conclusions

Radial basis function networks possess several attractive properties. Motivated by these attractive properties, a novel hybrid multiscale radial basis function (MSRBF) network has been introduced to model and forecast the *Dst* index. Compared with traditional single scale (kernel width) RBF networks, the new multiscale (multi-width) RBF network is more flexible and more powerful to describe complex input-output dynamical systems. While the polynomial submodel in the new network can be used to track the linear trend of the underlying dynamical behaviour, the MSRBF submodel can be used to capture the main underlying nonlinear dynamics by employing multiscale basis functions with different centres and widths. This enhances the capability of both the linear models and the traditional radial basis function networks. With a linear-in-the-parameters form, the new network can easily be trained using the forward orthogonal regression (FOR) algorithm, which combines good effectiveness with high efficiency. The identified network model provides very good short term predictions for the *Dst* index, over the associated data set. Albeit there exists a large discrepancy between long-term predictions and the associated measurements, the model predicts the strong storm very well.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC (UK).

References

Baker, D. N., Klimas, A. J., McPherron, R. L., and Buchner, J. The evolution from weak to strong geomagnetic activity: an interpretation in terms of deterministic chaos. *Geophys. Res. Lett.*, 17(1), 41-44, 1990.

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

Billings S. A. and Chen S., The determination of multivariable nonlinear models for dynamic systems using neural networks, in : *Neural Network Systems Techniques and Applications*, C.T. Leondes, Eds.. San Diego: Academic Press, 231-278, 1998.

Billings, S. A. and Wei, H. L. The wavelet-NARMAX representation: a hybrid model structure combining the polynomial models and multiresolution wavelet decompositions, *International Journal of Systems Science*, 36(3), 137-152, 2005a.

Billings S. A. and Wei, H. L. A new class of wavelet networks for nonlinear system identification, *IEEE Trans. Neural Networks*, 16(4), 862-874, 2005b.

Boaghe, O.M., Balikhin, M.A., Billings, S.A., and Alleyne, H. Identification of nonlinear processes in the magnetosphere dynamics and forecasting of *Dst* index, *J. Geophys. Res.*, 106(A12), 30047-30066, 2001.

Burton, R.K., McPherron, R.L. and Russell, C.T. An empirical relationship between interplanetary conditions and *Dst*. *J. Geophys. Res.*, 80, 4204-4214, 1975.

Chen, S., Billings, S. A., Cowan, C. F. N., and Grant, P. M. Practical identification of NARMAX models using radial basis functions. *Int. J. Control*, 52(6), 1327-1350, 1990.

Chen, S., Billings, S. A., Neural networks for nonlinear dynamic system modelling and identification, *Int. J. Control*, 56(2), 319-346, 1992.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Recognition*. (2nd ed.). New York: John Wiley & Sons, 2001.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

Goertz, C.K., Shan, L.H., and Smith, R.A. Prediction of geomagnetic activity. *J. Geophys. Res.*, 98(A5), 7673-7684, 1993.

Harris, C. J., Hong, X., and Gan, Q. *Adaptive Modelling, Estimation and Fusion from Data* : *A Neurofuzzy Approach*, Berlin : Springer-Verlag, 2002.

Hartman, E. J., Keeler, J. D., and Kowalski, J. M. Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*, 2(2), 210-215, 1990.

Hernandez, J.V., Tajima, T., and Horton, W. Neural net forecasting for geomagnetic activity. *Geophys. Res. Lett.*, 20(23), 2707-2710, 1993.

Klimas, A.J., Vassiliadis, D., Baker, D.N. *Dst* index prediction using data-derived analogues of the magnetospheric dynamics. *J. Geophys. Res.*, 103(A9), 20435-20447, 1998.

Krzanowski, W. J. and Lai, Y. T. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1), pp. 23-34, 1988.

Leontaritis, I. J. and Billings, S. A. Input-output parametric models for non-linear systems, *Int. J. Control*, 41(2), 303-344, 1985.

Liu, G. P. *Nonlinear Identification and Control : A Neural Network Approach*. London : Springer, 2001.

Lundstedt, H., Gleisner, H., Wintoft, P. Operational forecasts of the geomagnetic *Dst* index. *Geophys. Res. Lett.*, 29(24), Art. No. 2181, 2002.

McPherron, R.L. Predicting the Ap index from past behaviour and solar wind velocity. *Physics and Chemistry of the Earth Part C*, 24(1-3), 45-56, 1999.

O'Brien, T.P. and McPherron, R.L. An empirical phase space analysis of ring current dynamics: solar wind control of injection and decay. *J. Geophys. Res.*,105 (A4), 7707-7719, 2000.

Park, J. and Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2), 246-257, 1991.

Poggio, T. and Girosi, F. M. Networks for approximation and learning. *Proc. IEEE*, 78(9), 1481-1497,

1990.
Pulkkinen, T.I. and Baker, D.N. Global substorm cycle: what can the models tell us? *Surveys in Geophysics*, 18, 1-37, 1997.
Schwarz, G. Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464, 1978.
Sharifi, J., Araabi, B. N., Lucas, C. Multi-step prediction of *Dst* index using singular spectrum analysis and locally linear neurofuzzy modelling. *Earth Planets and Space*, 58 (3), 331-341, 2006.
Takalo, J., and Timonen, J. Neural network prediction of AE data. *Geophys. Res. Lett.*, 24(19), 2403-2406, 1993.
Vassiliadis, D., Klimas, A.J., Baker, D.N., and Roberts, D.A. A description of the solar-wind magnetosphere coupling based on nonlinear filters. *J. Geophys. Res.*,100(A3), 3495-3512, 1995.
Vassiliadis, D., Klimas, A.J., Valdivia, J.A., and Baker, D.N. Models of *Dst* geomagnetic activity and of its coupling to solar wind parameters. *Physics and Chemistry of the Earth Part C*, 24(1-3), 107-112, 1995.
Watanabe, S., Sagawa, E., Ohtaka, K., and Shimazu, H. Prediction of the *Dst* index from solar wind parameters by a neural network method. *Earth Planets and Space*, 54 (12), 1263-1275, 2002.
Wei, H.L., Billings, S.A. and Balikhin, M.A. Prediction of the *Dst* index using multiresolution wavelet Models. *J. Geophys. Res.*, 109(A7), A07212, doi:10.1029/2003JA010332, 2004a.
Wei, H. L., Billings, S. A., and Liu, J. Term and variable selection for nonlinear system identification, *Int. J. Control*, 77(1), 86-110, 2004b.
Wei, H. L., Billings, S. A., and Balikhin, M. A. Wavelet based nonparametric NARX models for nonlinear input-output system identification. Accepted by *International Journal of Systems Science*, 2006.
Wu, J.G., and Lundsted, H. Neural network modelling of solar wind magnetosphere interaction. *J. Geophys. Res.*, 102(A7), 14457-14466, 1997.

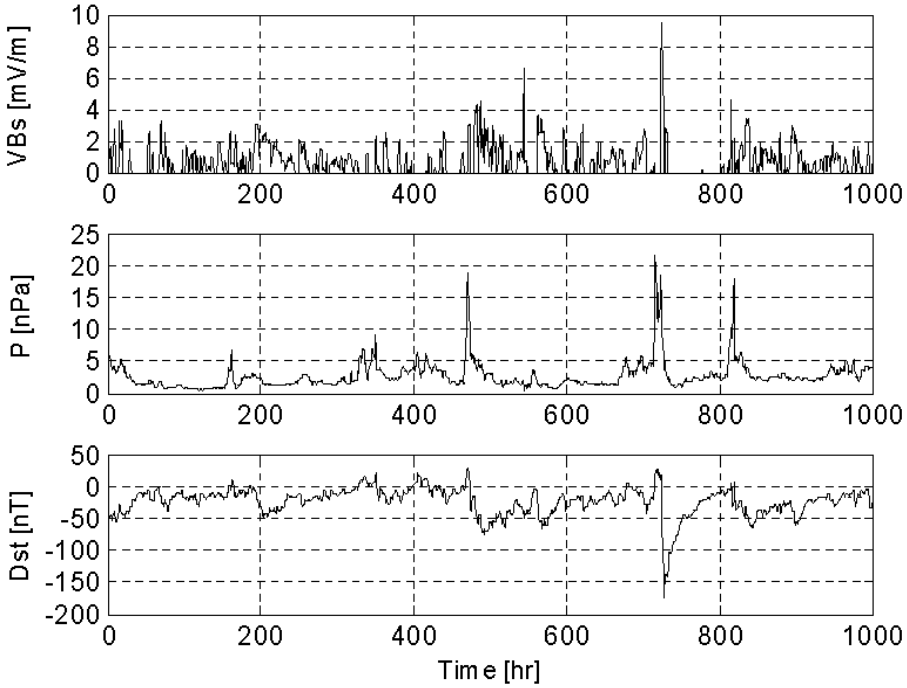

Fig. 1. The two inputs (the solar wind parameter *VBs* and the dynamical pressure, *P*) and the output (the *Dst* index).

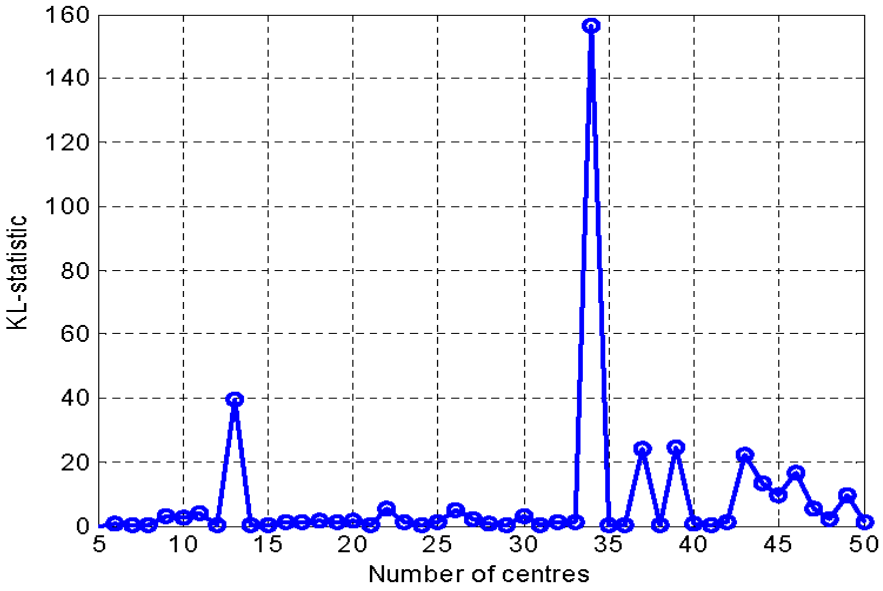

Fig. 2. The KL statistic, defined by (9), versus the number of centres.

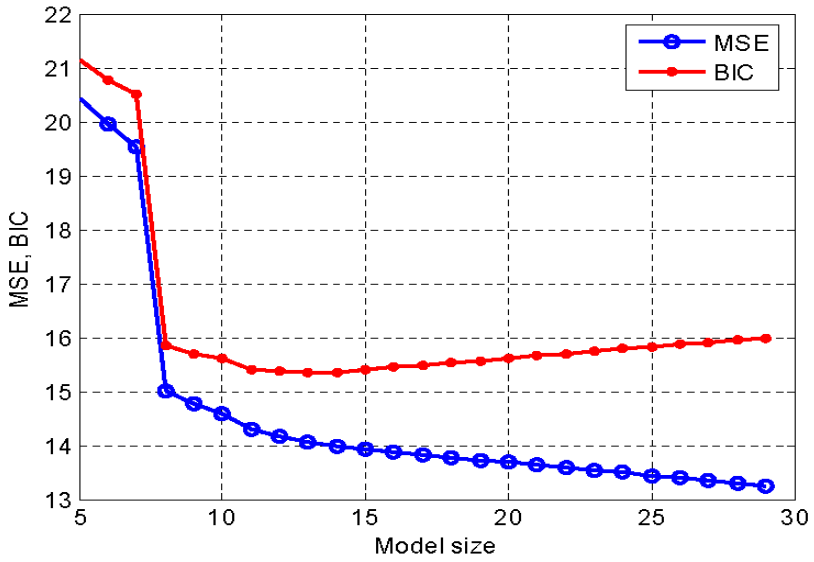

Fig. 3. BIC versus the number of significant regressors selected from the candidate model terms.

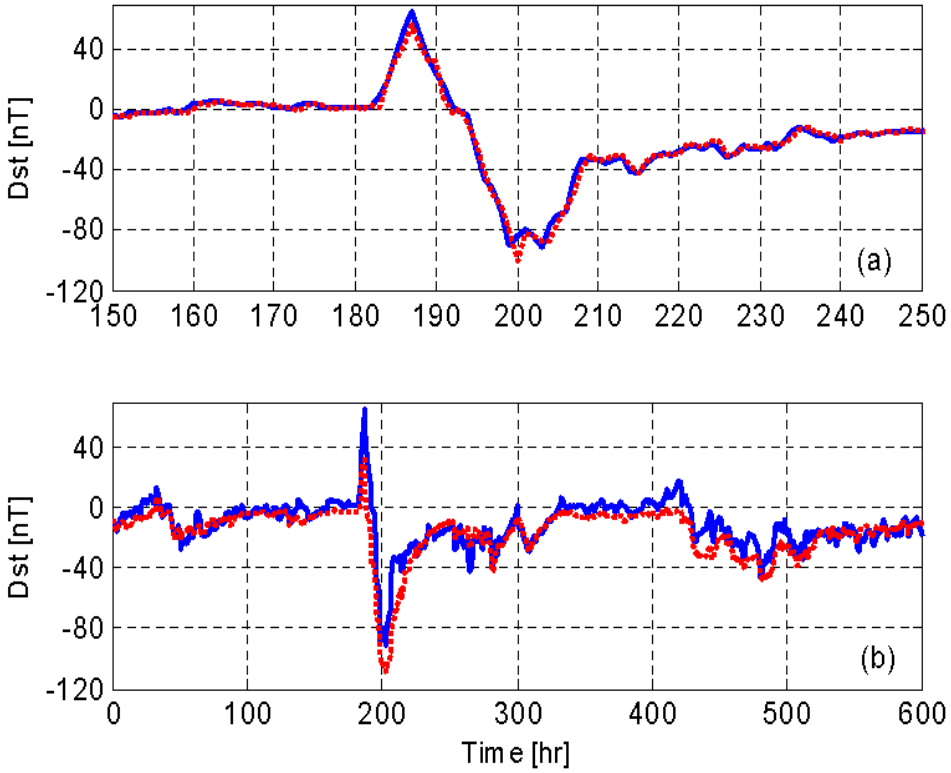

Fig. 4. Prediction performance of the identified RBF network model, over the validation data set. (a) one-step-ahead predictions; (b) long-term predictions. The solid lines indicate the measurements and dashed lines indicate the model predictions.