Training Basis Function Networks Including RBF and Multiresolution Wavelet Models with an Adaptive Orthogonal Least Squares Algorithm

Hua-Liang Wei and Stephen A. Billings

Research Report No. 892

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

May 2005

Training Basis Function Networks Including RBF and Multiresolution Wavelet Models with an Adaptive Orthogonal Least Squares Algorithm

Hua-Liang Wei and Stephen A. Billings
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK
S.Billings@Sheffield.ac.uk, W.Hualiang@Sheffield.ac.uk

May 9, 2005

Abstract: This paper concerns the construction and training of basis function networks for the identification of nonlinear dynamical systems. A new adaptive orthogonal least squares (AOLS) algorithm, which integrates basis function (model term) selection, model size determination and model parameter estimation, is developed for basis function network training. The construction and training of a new multiresolution wavelet network, together with a comparison with RBF models, is discussed in detail. A practical three-phase modelling framework to ensure unbiased nonlinear models is obtained using the basis function networks. Several examples are presented to demonstrate the application potential of the new identification techniques.

Index Terms: information criteria, model term selection, model structure detection, neural network, nonlinear system identification, orthogonal least squares, radial basis function, wavelet.

1. Introduction

Basis function networks and their variants, where the approximator (predictor) is expressed as a set of known basis functions which are combined and organized in a prescribed way, have been extensively studied in the literature and have been widely applied in function learning and dynamical modelling for a long period. A variety of prototype functions, both global and local, have been adapted and employed as basis functions to construct approximators for given nonlinear problems. Polynomials [1][2], cerebellar model articulation controller (CMAC) [3], radial basis functions (RBFs)[4], B-splines[5], kernels [6][7] and wavelets [8][9] are among the popular subclasses of basis function network systems. The present study will focus on the construction and training of a new class of multiresolution wavelet models for nonlinear dynamical system identification, coupled with a comparison with RBF models. More useful information on basis function networks can be found in [10].

Radial basis functions were originally proposed as an interpolation method in the late 1980s [4], and were soon connected to neural networks [11]-[13]. These kind of networks were then popularized in the community of nonlinear dynamical modelling, identification and control based on many subsequent results including [14]-[21]. Compared with standard feedforward neural networks, most RBF networks possess in some sense a local property and permit local tuning to track signal variation in given processes [12]. In addition, RBF networks are easier to train due to the network structure [22]. While a wide class of nonlinear functions can be approximated using RBF networks, a large range of severely nonlinear systems, for instance, systems with fast or sharp variations and/or discontinuities, may not be well approximated by this kind of network, which lacks good time-frequency properties compared with multiresolution wavelets.

It is generally recognized that the basis functions should offer some flexibility in adapting to the complexity of the model structure so that the model can match, as closely as possible, the underlying nonlinearity of dynamic systems. With excellent approximation properties associated with multiresolution decompositions [8][9], wavelets outperform many other approximation schemes and are well-suited for approximating arbitrary nonlinear systems, even those with chirps and discontinuities. It was the attractive features possessed by wavelets, especially by multiresolution wavelet decompositions, that motivated the introduction of wavelets as basis functions to form nonlinear models for complex dynamical systems[23]-[28]. It follows that the intrinsic nonlinear dynamics related to real nonlinear systems can easily be captured by an appropriately fitted wavelet model consisting of a small number of wavelet basis functions, and this makes wavelet representations more adaptive compared with many other basis functions.

In a nonlinear model, the relationship between the model outputs and inputs is nonlinear by definition. However, the relationship between the model outputs and the free adjustable model parameters may be either linear or nonlinear. Identification schemes can therefore be classified into two categories, linear-in-the-parameters and nonlinear-in-the-parameters. This study investigates basis function networks, which can be expressed as a linear-in-the-parameters model. More thorough treatments on both linear and nonlinear-in-the-parameters models can be found in [29]. An initial basis function network may involve a great number of candidate model terms whatever basis functions are employed to approximate an unknown nonlinear function, especially for high dimensional multivariable problems. Experience shows that in most cases only a small number of significant model terms are necessary in the final model to represent given observational data. Model subset selection, or model structure detection, is a key step in any identification procedure and consists of detecting and selecting significant model terms from a redundant candidate model term set to determine a parsimonious final model. A new adaptive orthogonal least squares (AOLS) scheme that can be used to select not only the significant model terms but also the optimal number of model terms to arrive at a good balance for the bias-variance trade-off, is introduced.

Motivated by the successful applications of wavelet decompositions, this study aims to develop a new class of basis function networks, accompanied by a new model term selection algorithm, for nonlinear dynamical system identification. The remainder of the paper is organized as follows. In Section 2, the NARMAX representation is briefly summarized. In Section 3, a new AOLS algorithm, which can be used to train general basis function networks of the linear-in-the-parameters form, is developed and discussed in detail. In Section 4, a new class of basis function networks, the multiresolution wavelet networks, are presented. In Section 5, a three-phase modelling framework is proposed to implement the general NARMAX model using basis function networks. Several examples are provided in Section 6. Some suggestions and comments on RBF and multiresolution wavelet models are given in Section 7, and finally the work is concluded in Section 8.

2. Problem Representation

A wide class of input-output nonlinear dynamical systems can be represented by the general nonlinear difference equation model, known as the NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous inputs) model[30][31]. Under some mild conditions a discrete-time multivariable nonlinear system with m outputs and r inputs can be described by the following NARMAX model

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

where $\mathbf{u}(t)=[u_1(t),u_2(t),\cdots,u_r(t)]^T$, $\mathbf{y}(t)=[y_1(t),y_2(t),\cdots,y_m(t)]^T$ and $\mathbf{e}(t)=[e_1(t),e_2(t),\cdots,e_m(t)]^T$ are the system input, output and noise vectors, n_u, n_y and n_e are the maximum lags in input, output and noise, respectively, and $\mathbf{f}$ is some vector-valued and in general unknown nonlinear mapping. In practice it is usually assumed that $\mathbf{e}(t)$ is an independent noise sequence. Model (1) relates the inputs and outputs and takes into account the combined effects of measurement noise, modelling errors and unmeasured disturbances represented by the noise variable $\mathbf{e}(t)$ .One of the reasons that the moving average terms are included in the NARMAX model (1) is to ensure unbiased estimates.

Model (1) will be used for representing both SISO and MIMO nonlinear systems in the present study. Decomposing Eq. (1) into component form gives the formulation of the ith output

$$\begin{aligned} y_i(t)=f_i(&y_1(t-1),\cdots,y_1(t-n_{y1}^{(i)}),\cdots,y_m(t-1),\cdots,y_m(t-n_{ym}^{(i)}),\\ &u_1(t-1),\cdots,u_1(t-n_{u1}^{(i)}),\cdots,u_r(t-1),\cdots,u_r(t-n_{ur}^{(i)}),\\ &e_1(t-1),\cdots,e_1(t-n_{e1}^{(i)}),\cdots,e_m(t-1),\cdots,e_m(t-n_{em}^{(i)}))+e_i(t) \end{aligned} \tag{2}$$

The nonlinear functions $f_i(\cdot)$ (i=1,2, .., m) are generally unknown. In order to simplify the notation it is sometimes assumed that the maximum lags for the different elements of the output $\mathbf{y}(t)$ are the same, that is, $n_{yk}^{(i)}=n_{yk}$ for i=1,2, ..., m and k=1,2, ..., m. Similarly, $n_{uk}^{(i)}=n_{uk}$ for i=1,2, ..., r and k=1,2, ..., m and $n_{ek}^{(i)}=n_{ek}$ for i=1,2, ..., m and k=1,2, ..., m.

One of the most popular representations for the NARMAX model (1) is the polynomial model which takes the function $\mathbf{f}(\cdot)$ as a polynomial with respect to the lagged input, output and noise sequences. An important property of the polynomial NARMAX model is that these models are linear-in-the-parameters so that the model structure and the parameters can be detected and estimated using standard structure detection schemes[32]-[34]. Moreover, as a natural extension of the ARMAX model, the polynomial NARMAX models can be physically interpreted under certain conditions in both the time and the frequency domain.

Taking the SISO case as an example, the power-form polynomial NARMAX model of the degree ℓ can be described as

$$\begin{aligned} y(t)=\theta_0+\sum_{i_1=1}^{d}f_{i_1}(x_{i_1}(t))+\sum_{i_1=1}^{d}\sum_{i_2=i_1}^{d}f_{i_1i_2}(x_{i_1}(t),x_{i_2}(t))+\cdots\\ +\sum_{i_1=1}^{d}\cdots\sum_{i_\ell=i_{\ell-1}}^{d}f_{i_1i_2\cdots i_\ell}(x_{i_1}(t),x_{i_2}(t),\cdots,x_{i_\ell}(t))+e(t) \end{aligned} \tag{3}$$

where $\theta_{i_1i_2\cdots i_m}$ are parameters, $d=n_y+n_u+n_e$ and

$$f_{i_1i_2\cdots i_m}(x_{i_1}(t),x_{i_2}(t),\cdots,x_{i_m}(t))=\theta_{i_1i_2\cdots i_m}\prod_{k=1}^{m}x_{i_k}(t),\ 1\le m\le\ell, \tag{4}$$

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y+1 \le k \le n_y+n_u \\ e(t-(k-n_y-n_u)) & n_y+n_u+1 \le k \le n_y+n_u+n_e \end{cases} \tag{5}$$

The degree of a multivariate polynomial is defined as the highest order among all the terms. For example, the degree of the polynomial $h(x_1,x_2,x_3)=a_1x_1^4+a_2x_2x_3+a_3x_1^2x_2x_3^2$ is $\ell=2+1+2=5$. Similarly, a polynomial NARMAX model with nonlinear degree ℓ means that the order of each term in the model is not higher than ℓ. It can easily be proved that the number of potential model terms in the polynomial NARMAX model (3) of degree ℓ is $M=(n+\ell)!/[n!\ell!]$, where $n=n_y+n_u+n_e$. Clearly, the NARMAX model (1) can describe a large range of nonlinear systems and includes several existing representations including the Volterra series, AR(X), ARMA(X), NAR(X), NARMA and bilinear models as special cases [35].

In practice, the unknown nonlinear function f in model (1) often consists of two parts: the deterministic (noise independent) and the stochastic (noise correlated) submodels shown as below

$$\mathbf{y}(t)=\mathbf{f}_{yu}(\mathbf{y}^{[t-1,n_y]},\mathbf{u}^{[t-1,n_u]})+\mathbf{f}_{yue}(\mathbf{y}^{[t-1,n_y]},\mathbf{u}^{[t-1,n_u]},\mathbf{e}^{[t-1,n_e]})+\mathbf{e}(t) \tag{6}$$

where the vector $\mathbf{z}^{[t-1,n]}$ is defined as $\mathbf{z}^{[t-1,n]}=[\mathbf{z}^T(t-1),\cdots,\mathbf{z}^T(t-n)]^T$. Note that each term of the submodel $\mathbf{f}_{yue}$ is dependent on noise sequence $\mathbf{e}(t-1),\mathbf{e}(t-2),\cdots,\mathbf{e}(t-n_e)$. For a linear-in-the-parameters basis function network, model (6) can be expressed as

$$\mathbf{y}(t)=\Phi_{yu}(t)\Theta_{yu}+\Phi_{yue}(t)\Theta_{yue}+\mathbf{e}(t) \tag{7}$$

where $\Phi_{yu}(t)$ and $\Phi_{yue}(t)$ are regression matrices, and Θ_{yu} and Θ_{yue} are unknown parameter vectors.

The objective of this study is to implement the NARMAX model (1) using basis function networks including RBF and the newly introduced multiresolution wavelet models. A new orthogonal least squares learning algorithm is developed for basis function network training.

3. Basis Function Networks and Training

3.1 Basis function networks

A single hidden-layer feedforward basis function network with d independent variables can be expressed by

$$g(\mathbf{x})=\sum_{i\in I}\theta_i\varphi_i(\mathbf{x};\mathbf{a}_i,\mathbf{b}_i) \tag{8}$$

where $\langle I\rangle$ indicates the number of total elements in the set I, $\mathbf{x}\in\mathbf{R}^d$, $\mathbf{b}_i\in\mathbf{R}^d$ and $\mathbf{a}_i\in\mathbf{R}^{+d}$ (d-dimensional positive value vector). The $\langle I\rangle$ neurons (basis functions) in the sum are linearly connected with $\langle I\rangle$ weights $\theta_1,\cdots,\theta_{\langle I\rangle}$. Each neuron maps a d-variable input $\mathbf{x}$ into a scalar value via a nonlinear mapping φ_i, which is dependent on both the scale (or dilation, kernel width) parameters $\mathbf{a}_i$ and the location (or position, translation, kernel centre) parameters $\mathbf{b}_i$ .The nonlinear functions φ_i are called the basis functions (or traditionally the activation functions). In most cases, the basis functions φ_i for i=1,2, …, $\langle I\rangle$ are chosen as the same mother basis function φ. Radial basis functions are a popular choice to construct networks, and a typical choice for basis functions in the network are the radial basis kernels, for example the Gaussian type kernels $\varphi_i : \mathbf{R}^d \mapsto \mathbf{R}$, $\varphi_i(\mathbf{x};\mathbf{a}_i,\mathbf{b}_i) = \varphi_i(\mathbf{a}_i^T \circ (\mathbf{x}-\mathbf{b}_i))$, where the operator ' $\circ$ ' between two vectors $\mathbf{u} = [u_1,\cdots,u_d]^T$ and $\mathbf{v} = [v_1,\cdots,v_d]^T$ indicates some specified operation, say $\mathbf{u} \circ \mathbf{v} = -(1/2)\mathbf{v}^T \Sigma_{\mathbf{u}}^{-1}\mathbf{v}$ with $\Sigma_u = diag[u_1,\cdots,u_d]$. Another popular choice for the basis functions is wavelets including multiresolution wavelets, which involve a mother wavelet and a corresponding scale function.

In a linear-in-the-parameters basis function network, the scale and location parameters $\mathbf{a}_i$ and $\mathbf{b}_i$ can often be pre-determined to simplify the training procedure. Let $\Gamma = \{(\mathbf{a}_i,\mathbf{b}_i) : i \in I\}$ and $\varphi_{(\mathbf{a}_i,\mathbf{b}_i)}(\mathbf{x}) = \varphi_i(\mathbf{x};\mathbf{a}_i,\mathbf{b}_i)$ for $(\mathbf{a}_i,\mathbf{b}_i) \in \Gamma$. A dictionary used for the network training for a given identification problem can then be defined as $D = \{\varphi_{(\mathbf{a}_i,\mathbf{b}_i)} : (\mathbf{a}_i,\mathbf{b}_i) \in \Gamma, i \in I\}$. Clearly, the dictionary D contains a total of $M = \langle I \rangle$ elements.

In practice, the number M of the total elements in the dictionary D may be very large, and most candidate model terms are either redundant or make very little contribution to the system output and can therefore be removed from the model. Thus, for a given identification problem, where the observed training data is of the form $\{(\mathbf{x}(t),\mathbf{y}(t)) : \mathbf{x} \in \mathbf{R}^d, \mathbf{y} \in \mathbf{R}^m, t = 1,2,\cdots,N\}$, the objective is to select a subset of M_0 ($M_0 \le M$) model terms to fit the given observations by training the network, to lead to a parsimonious approximator

$$\mathbf{y}(t) = \hat{g}(\mathbf{x}(t)) = \sum_{m=1}^{M_0} \theta_{i_m} \varphi_{i_m}(\mathbf{x}(t);\mathbf{a}_{i_m},\mathbf{b}_{i_m}) \tag{9}$$

An efficient model structure determination approach has been developed based on the forward orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) criterion, which was originally introduced to determine which terms should be included in a model [32]-[34]. The OLS-ERR algorithm has been extensively studied and widely applied in nonlinear system identification [19],[36]-[39]. The OLS-ERR algorithm provides a powerful tool to effectively select significant model terms step by step, one at a time, by orthogonalizing the associated regressors in a forward stepwise way based on the ERR criterion, an index indicating the significance of each model term. Most existing OLS algorithms, however, do not provide information on how many significant model terms should be selected and included in the final model. An additional separate procedure is therefore often needed to aid the determination of the optimal number of significant model terms. This study, however, will provide an adaptive OLS algorithm that incorporates model term selection and model size determination in one procedure.

3.2 Model term selection and the orthogonal transformation

For convenience of description, consider the case that involves only one output. Let $\mathbf{y} = [y(1),\cdots,y(N)]^T$ be a vector of measured outputs at N time instants, and $\alpha_m = [\pi_m(1),\cdots,\pi_m(N)]^T$ be a vector associated with the mth candidate model term, where $\pi_m \in D$ for m=1,2, ..., M, and D is a dictionary produced by lagged outputs, inputs and noise terms. From the viewpoint of practical modelling and identification, the finite dimensional set $S = \{\alpha_1,\cdots,\alpha_M\}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $S_n = \{\beta_1, \cdots, \beta_n\} = \{\alpha_{i_1}, \cdots, \alpha_{i_n}\} \subseteq S$, where $\beta_k = \alpha_{i_k}$, $i_m \in \{1,2,\cdots,M\}$ and m=1,2, ..., n, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\beta_1, \cdots, \beta_n$ as below

$$\mathbf{y} = \theta_1\beta_1 + \cdots + \theta_n\beta_n + \mathrm{e} \tag{10}$$

or in a compact matrix form

$$\mathbf{y} = P\theta + \mathrm{e} \tag{11}$$

where the matrix $P = [\beta_1, \cdots, \beta_n]$ is of full column rank, $\theta = [\theta_1, \cdots, \theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is an approximation error. From matrix theory, the full rank matrix P can be orthogonally decomposed as

$$P = QR \tag{12}$$

where R is an $n \times n$ unit upper triangular matrix and Q is an $n \times n$ matrix with orthogonal columns $q_1, q_2, \cdots, q_n$. Substituting (12) into (11), yields

$$\mathbf{y} = (PR^{-1})(R\theta) + \mathbf{e} = Qg + \mathbf{e} \tag{13}$$

where $g = [g_1, \cdots, g_n]^T = R\theta$ is an auxiliary parameter vector. Using the orthogonal property of Q, g_i can be directly calculated from $\mathbf{y}$ and Q as $g_i = (\mathbf{y}^T q_i)/(q_i^T q_i)$ for i=1,2, ..., n. The unknown parameter vector θ can then be easily calculated from g and R by substitution using the special structure of R.

Assume that the error $\mathbf{e}$ in model (13) is uncorrelated with vectors β_j for j=1,2, ..., n, the total sum of squares of the output from the origin can then be expressed as

$$\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{n} g_i^2 q_i^T q_i + \mathbf{e}^T\mathbf{e} \tag{14}$$

Note that the total sum of squares $\mathbf{y}^T\mathbf{y}$ consists of two parts, the desired output $\sum_{i=1}^{n} g_i^2 q_i^T q_i$, which can be explained by the selected regressors (model terms), and the part $\mathbf{e}^T\mathbf{e}$, which represents the residual sum of squares. Thus, $g_i^2 q_i^T q_i$ is the increment to the desired total sum of squares of the output brought by q_i. The ith error reduction ratio (ERR) introduced by q_i (or equally by including β_i), is defined as

$$\mathrm{ERR}[i] = \frac{g_i^2(q_i^T q_i)}{\mathbf{y}^T\mathbf{y}} \times 100\% = \frac{(\mathbf{y}^T q_i)^2}{(\mathbf{y}^T\mathbf{y})(q_i^T q_i)} \times 100\%, \quad i=1,2, \ldots, n, \tag{15}$$

This ratio provides a simple but an effective index to indicate the significance of adding the ith term into the model. The orthogonalization procedure for model term selection is usually implemented in a stepwise way, one term at a time. The *sum of error reduction ratio* (SERR) and the *error-to-signal ratio* (ESR) due to $q_1, \cdots, q_j$ (or equally due to $\beta_1, \cdots, \beta_j$) are defined as

$$\mathrm{SERR}[j] = \sum_{i=1}^{j} \mathrm{ERR}[i] \tag{16}$$

$$\text{ESR}[j] = \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{y}} = 1 - \sum_{i=1}^{j} \frac{g_i^2 q_i^T q_i}{\mathbf{y}^T\mathbf{y}} = 1 - \sum_{i=1}^{j} \text{ERR}[i] = 1 - \text{SERR}[j] \tag{17}$$

The selection procedure will be terminated when ESR of an identified model satisfies some specified conditions. Several orthogonal transforms including Gram-Schmidt, modified Gram-Schmidt and Householder transformations can be applied to implement the orthogonal decomposition [34],[35] and a detailed algorithm will be given in the next two sections.

3.3 Model size determination

The determination of the optimal number of model terms is critical in dynamical modelling. Neither an over-fitting nor an under-fitting model is desirable in practical identification. In practice, however, the true number of terms is generally unknown and needs to be estimated during model identification. Several approaches have been developed for model order and variable selection in the literature including the AIC, BIC, MDL [40]-[43] and many variants [44] (Chap. 6). In this study, a R^2-like statistic, the adjustable prediction error sum of squares (R^2-PRESS) proposed by Allen [45], [46], is modified and will be used to solve the term selection problem.

The commonly used adjustable R^2-statistic is defined as

$$R_a^2 = 1 - \frac{N-1}{N-p}\text{NMSE} \tag{18}$$

where N is the data length, p is the number of model terms included in the identified model, NMSE is the normalised-mean-squared-error defined as

$$\text{NMSE} = \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{N}[y(i) - \hat{y}(i)]^2}{\sum_{i=1}^{N}[y(i) - \bar{y}]^2} \tag{19}$$

where $\text{SST} = \sum_{i=1}^{N}[y(i) - \bar{y}]^2$ denotes the total sum of squared deviations in $\mathbf{y}$ from the mean $\bar{\mathbf{y}}$, $\text{SSE} = \sum_{i=1}^{N}[y(i) - \hat{y}(i)]^2$ denotes the sum of the squared errors (residuals), $\{\hat{y}(i)\}_{i=1}^{N}$ is the one-step-ahead prediction sequence from the identified model with p terms.

The prediction error sum of squares (PRESS) proposed in [45], [46] provides a useful residual scaling, which can be used as a form of cross validation by leaving one point out at a time [47]. The prediction error sum of squares is defined as

$$\text{PRESS} = \sum_{i=1}^{N}[y(i) - \hat{y}_{-i}(i)]^2 = \sum_{i=1}^{N}[\varepsilon_{-i}(i)]^2 \tag{20}$$

where $\hat{y}_{-i}(i)$ are one-step-ahead predicted values from a model fitted using a data set consisting of N-1 observational data point pairs, which are obtained by leaving the ith data point pair out, $\varepsilon_{-i}(i)$ are the PRESS predicted residuals evaluated at the ith point. Let $\varepsilon(i)$ be the normally defined residuals of a model fitted using the total N data points, it can be shown that the relationship between $\varepsilon_{-i}(i)$ and $\varepsilon(i)$ is

$$\varepsilon_{-i}(i) = \frac{y(i)-\hat{y}(i)}{1-\beta_i^T(P^TP)^{-1}\beta_i} = \frac{\varepsilon(i)}{1-h(i,i)} \tag{21}$$

where β_i and P are defined as in (10). Thus PRESS can be reduced to

$$\text{PRESS} = \sum_{i=1}^{N}\left(\frac{\varepsilon(i)}{1-h(i,i)}\right)^2 \tag{22}$$

This shows that the PRESS statistic can be calculated by fitting only one model using the total N data points, but N "leave-one-out" matrices are still required. It can be proved [44] that if $N >> p$, PRESS can be approximated as

$$\text{PRESS} \approx \left(\frac{N}{N-p}\right)^2 \text{SSE} \tag{23}$$

Statistic (23) gives some indication of the predictive capability of the regression model. This will be used to define the adjustable R^2-PRESS statistic given below

$$R^2_{\text{press}} = 1 - \frac{\text{PRESS}}{\text{SST}} = 1 - \left(\frac{N}{N-p}\right)^2 \frac{\text{SSE}}{\text{SST}} \tag{24}$$

Note, however, that sometimes the data length N may be long, say $N \geq 2000$. In this case, the effect of n in the denominator of (24) is minimal due to the fact that $(N/(N-p))^2 \approx 1 + 2p/N \approx 1$ for $p << N/2$. One way to avoid the tendency that small p's are mitigated by a large N is to replace the number p by λp, where λ is an adjustable coefficient. Experience shows that a typical choice for λ is to set $\lambda = \max\{1,\ \rho N\}$ with $0.002 \leq \rho \leq 0.01$. The adjustable R^2-PRESS can then be defined as

$$R^2_{\text{apress}} = 1 - \left(\frac{N}{N-\lambda p}\right)^2 \text{NMSE} \tag{25}$$

Note that the R^2-APRESS statistic (25) is in formulation similar to the adjustable R^2-statistic given by (18). In the next section, the R^2-APRESS statistic will be combined with the criterion ESR (error-to-signal ratio) and will then be incorporated into the orthogonal least squares algorithm.

3.4 The adaptive orthogonal least squares (AOLS) learning algorithm

At first sight, the calculation of the R^2-APRESS statistic defined by (25) requires an initial calculation of the value of NMSE, which involves the calculation of the one-step-ahead prediction, $\hat{\mathbf{y}}$. From the definition of ESR in (17), however, the calculation of NMSE is not necessary. In fact, from (17) and (25), the R^2-APRESS for an identified model with p model terms can be calculated as

$$R^2_{\text{apress}}[p] = 1 - \left(\frac{N}{N-\lambda p}\right)^2 \text{NMSE}[p]$$

$$=1-\left(\frac{N}{N-\lambda p}\right)^2\left(\frac{\mathbf{e}^T\mathbf{e}}{\mathrm{SST}}\right)_{[p]}$$

$$=1-\left(\frac{\mathrm{SST}_0}{\mathrm{SST}}\right)\left(\frac{N}{N-\lambda p}\right)^2\left(\frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{\mathrm{SST}_0}\right)_{[p]}$$

$$=1-\left(\frac{\mathrm{SST}_0}{\mathrm{SST}}\right)\left(\frac{N}{N-\lambda p}\right)^2 \mathrm{ESR}[p] \qquad (26)$$

where $\mathrm{SST}_0=\mathbf{y}^T\mathbf{y}=\sum_{i=1}^{N} y^2(i)$ is the total sum of squared deviations in $\mathbf{y}$ from the origin, and SST is defined in (14), and the index or subscript $[p]$ indicates that the associated items are calculated from an identified model with p terms. Note that $\mathrm{ESR}[p]$ (p=1,2, ...) in (26) are available as a by-product of the orthogonalization procedure.

Assume that there exists a number p_0, at which the function $R^2_{\mathrm{apress}}[p]$ with respect to p is a maximum. At the maximum of $R^2_{\mathrm{apress}}[p]$, the following relationships hold

$$R^2_{\mathrm{apress}}[p_0] > R^2_{\mathrm{apress}}[p_0-1] \qquad (27a)$$

$$R^2_{\mathrm{apress}}[p_0] \geq R^2_{\mathrm{apress}}[p_0+1] \qquad (27b)$$

A little rearrangement of (27a) and (27b) gives

$$\frac{\mathrm{ESR}(p_0)}{\mathrm{ESR}(p_0-1)} < \left(\frac{N-\lambda p_0}{N-\lambda(p_0-1)}\right)^2 \qquad (28a)$$

$$\frac{\mathrm{ESR}(p_0+1)}{\mathrm{ESR}(p_0)} \leq \left(\frac{N-\lambda(p_0+1)}{N-\lambda p_0}\right)^2 \qquad (28b)$$

Define two functions

$$\chi_1(p)=\frac{\mathrm{ESR}(p+1)}{\mathrm{ESR}(p)} \qquad (29)$$

$$\chi_2(p)=\left(\frac{N-\lambda(p+1)}{N-\lambda p}\right)^2 \qquad (30)$$

From (28a) and (28b), χ_1 and χ_2 have the following property: $\chi_1(p)<\chi_2(p)$ for $p<p_0$, and $\chi_1(p)\geq\chi_2(p)$ for $p=p_0$. The two functions defined by (29) and (30) will be used as an indicator to find the optimal model term number p_0, where the two indicating functions intersect. In fact, the optimal number p_0 can be chosen as the point where χ_1 enters into a small confidence interval of χ_2 for the first time, say the interval $\chi_1 \pm \delta$, where δ is a small positive number.

The new adaptive orthogonal least squares algorithm (AOLS) can now be described below, where $\alpha_1,\cdots,\alpha_M$ are the vectors associated with the M candidate model terms.

The ALOS algorithm:

```
Step 1: Set $I_1 = \{1,2,\cdots,M\}$; $s_0 = \mathbf{y}^T\mathbf{y}$; $s_1 = (\mathbf{y}-\overline{\mathbf{y}})^T(\mathbf{y}-\overline{\mathbf{y}})$;
    for i=1 to M
        $\beta_i^{(1)} = \alpha_i$;
        $err^{(1)}[i] = \frac{(\mathbf{y}^T\beta_i^{(1)})^2}{s_0(\beta_i^{(1)})^T\beta_i^{(1)}}$; {if $(\beta_i^{(1)})^T\beta_i^{(1)} \approx 0$, set $err^{(1)}[i] = 0$ };
        $a_{ii} = 1$;
    end for
    $\ell_1 = \arg\max_{i\in I_1}\{err^{(1)}[i]\}$; $err[1] = err^{(1)}[\ell_1]$;
    $serr[1] = err[1]$; $esr[1] = 1 - serr[1]$;
    $q_1 = \beta_{\ell_1}^{(1)}$; $g_1 = \frac{\mathbf{y}^T q_1}{q_1^T q_1}$;
Step $j$, $j \ge 2$:
    For j=2 to M
        $I_j = I_{j-1} \setminus \{\ell_{j-1}\}$;
        for $i \in I_j$
            $\beta_i^{(j)} = \beta_i^{(j-1)} - \frac{\alpha_i^T q_{j-1}^T}{q_{j-1}^T q_{j-1}} q_{j-1}^T$;    (31)
            $err^{(j)}[i] = \frac{(\mathbf{y}^T\beta_i^{(j)})^2}{s_0(\beta_i^{(j)})^T\beta_i^{(j)}}$; {if $(\beta_i^{(j)})^T\beta_i^{(j)} < \delta$, set $err^{(j)}[i] = 0$ };    (32)
        end for ( end loop for $i$ )
        $J_j = \{\arg_{i\in I_j}((\beta_i^{(j)})^T\beta_i^{(j)} < \delta)\}$; $I_j = I_j \setminus J_j$;    (33)
        $\ell_j = \arg\max_{i\in I_j}\{err^{(j)}[i]\}$; $err[j] = err^{(j)}[\ell_j]$;
        Calculate: {SERR[j], ESR[j], $R^2_{\text{apress}}[j]$, $\chi_1[j]$, and $\chi_2[j]$};
        $q_j = \beta_{\ell_j}^{(j)}$; $g_j = \frac{\mathbf{y}^T q_j}{q_j^T q_j}$; $a_{jj} = 1$;
        for k=1 to j-1
            $a_{kj} = \frac{\alpha_{\ell_j}^T q_k}{q_k^T q_k}$;
        end for (end loop for $k$ )
    end for (end loop for $j$ )
```

Remark **1:** The AOLS algorithm provides an effective tool for selecting significant model terms in an iterative stepwise way. Terms are selected step by step, one term at a time. Most numerical ill conditioning can be avoided by eliminating the candidate regressors for which $(\beta_i^{(j)})^T\beta_i^{(j)}$ are less than a predetermined threshold δ, say $\delta = 10^{-\tau}$ with $\tau \ge 10$ (see Eqs. (27), (28)). In the case where both the dictionary and the data length are large, other faster OLS algorithms can be adapted into the AOLS to lessen the calculation load and spare computation time. For example, the MPOLS (pursuit matching orthogonal least squares) algorithm proposed in [39] is very fast compared with most existing OLS algorithms, and can be used to handle large scale data with a high SNR, but this is achieved at the expense of producing over-parameterised models compared to OLS.

Remark 2: The assumption that the initial candidate regression vector set $S = \{\alpha_1, \cdots, \alpha_M\}$ is of full dimensionality is unnecessary in the iterative forward AOLS algorithm. In fact, if the M vectors P are linearly dependent, and assuming that the dimension of S is n (<M) , the algorithm will stop at the n-th step.

Remark 3: If required, the selection procedure can be terminated at step M_0 (generally $M_0 << M$), the optimal number of model terms, at which point the function $R^2_{\text{apress}}[m]$ with respect to m will be a maximum that satisfies $\chi_1(M_0) \geq \chi_2(M_0)$. The system output can be expressed as a linear combination of the M_0 selected significant regressrs

$$\mathbf{y} = g_1 q_1 + \cdots + g_{M_0} q_{M_0} + \boldsymbol{\varepsilon} \tag{34}$$

which is equivalent to

$$y(t) = \sum_{i=1}^{M_0} \theta_{\ell_i} \pi_{\ell_i}(t) + \varepsilon(t) \tag{35}$$

where $\pi_{\ell i} \in D$ (the associated dictionary), the parameters $\theta^{(AOLS)} = [\theta_{\ell_1}, \theta_{\ell_2}, \cdots, \theta_{\ell_{M_0}}]^T$ are calculated from the triangular equation $Ag = \theta^{(AOLS)}$ with $g = [g_1, g_2, \cdots, g_{M_0}]^T$ and

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1M_0} \\ 0 & 1 & \cdots & a_{2M_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{M_0-1,M_0} \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The entries a_{ij} $(1 \leq i < j \leq M_0)$ are calculated during the orthogonalization procedure.

4. Multiresolution Wavelet Networks

4.1 Multiresolution wavelet decompositions

From wavelet theory [8],[9], any function $f \in L^2(\mathbf{R})$ can be expressed as the following multiresolution wavelet decomposition

$$f(x) = \sum_k a_{j_0,k} \phi_{j_0,k}(x) + \sum_{j \geq j_0} \sum_k d_{j,k} \psi_{j,k}(x) \tag{36}$$

where $\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k)$, $\phi_{j,k}(x) = 2^{j/2} \phi(2^j x - k)$, and the integer numbers j and k are the scale and location parameters, and j_0 is an arbitrary integer representing the lowest resolution or scale level.

Using the concept of *tensor products*, the multiresolution decomposition (36) can be immediately generalised to the muti-dimensional case, where a multiresolution wavelet decomposition can be defined by taking the tensor product of the one-dimensional scale and wavelet functions [9]. Let $f \in L^2(\mathbf{R}^d)$, then f can be represented by the multiresolution wavelet decomposition as

$$f(x_1,\cdots,x_d)=\sum_k \alpha_{j_0,k}\Phi_{j_0,k}(x_1,\cdots,x_d)+\sum_{j\geq j_0}\sum_k\sum_{l=1}^{2^d-1}\beta_{j,k}^{(l)}\Psi_{j,k}^{(l)}(x_1,\cdots,x_d) \tag{37}$$

where $k=(k_1,k_2,\cdots,k_d)\in \mathbf{Z}^d$ and

$$\Phi_{j_0,k}(x_1,\cdots,x_d)=2^{j_0 d/2}\prod_{i=1}^{d}\phi(2^{j_0}x_i-k_i) \tag{38}$$

$$\Psi_{j,k}^{(l)}(x_1,\cdots,x_d) \;=2^{jd/2}\prod_{i=1}^{d}\eta^{(i)}(2^j x_i-k_i) \tag{39}$$

with $\eta^{(i)}=\phi$ or ψ (scalar scale function or the mother wavelet) but at least one $\eta^{(i)}=\psi$. For some appropriate J, the approximation representation (37) can be approximated using only the scale function ϕ,

$$f(x_1,\cdots,x_d)=\sum_k \alpha_{J,k}\Phi_{J,k}(x_1,\cdots,x_d)=\sum_{k_1,k_2,\cdots,k_d}\alpha_{J;k_1,\cdots k_d}2^{Jd/2}\prod_{i=1}^{d}\phi(2^J x_i-k_i) \tag{40}$$

4.2 B-splines and associated mother wavelets

Although many functions can be chosen as scale and/or wavelet functions, most of these are not suitable for system identification applications, especially in the case of multidimensional and multiresolution expansions. An implementation, which has been tested with very good results, involves B-splines and associated mother wavelet for multiresolution wavelet decompositions[26],[28]. For a comprehensive discussion on B-splines, see Chui [8].

The B-spline function of mth order is defined by the following recursive formula:

$$N_m(x)=\frac{x}{m-1}N_{m-1}(x)+\frac{m-x}{m-1}N_{m-1}(x-1),\; m\geq 2 \tag{41}$$

with

$$N_1(x)=\chi_{[0,1)}(x)=\begin{cases}1 & \textit{if } x\in[0,1)\\ 0 & \textit{otherwise}\end{cases} \tag{42}$$

Define $\phi(x)=N_m(x)$, and

$$\psi(x)=\sum_{k=0}^{3m-2}d_k N_m(2x-k) \tag{44}$$

with the coefficients given by

$$d_k=\frac{(-1)^k}{2^{m-1}}\sum_{j=0}^{m}\binom{m}{j}N_{2m}(k-j+1),\; k=0,1,\cdots,3m-2 \tag{45}$$

TABLE 1
THE B-SPLINES OF ORDER 1 TO 4

	$N_1(x)$	$N_2(x)$	$2N_3(x)$	$6N_4(x)$
$0 \le x < 1$	1	x	x^2	x^3
$1 \le x < 2$	0	$2-x$	$-2x^2+6x-3$	$-3x^3+12x^2-12x+4$
$2 \le x < 3$	0	0	$(x-3)^2$	$3x^3-24x^2+60x-44$
$3 \le x \le 4$	0	0	0	$-x^3+12x^2-48x+64$
elsewhere	0	0	0	0

The functions ϕ and ψ can then generate multiresolution wavelet decompositions [8]. Clearly, the support of the mth order B-spline and the associated wavelet are

$$\begin{cases} \text{supp } \phi = \text{supp} N_m = [0, m] \\ \text{supp } \psi = [0, 2m-1] \end{cases} \tag{46}$$

The most attractive and distinctive properties of B-splines and the associated mother wavelets compared with other wavelets are that they are compactly supported and can be analytically formulated in an explicit form. Most importantly, they form multiresloution wavelet decompositions. To the best of our knowledge, B-splines and the associated mother wavelets are unique because they simultaneously possess the three remarkable properties, namely compactly supported, analytically formulated and multiresolution analysis oriented, among all known wavelets. These splendid properties make B-splines and associated mother wavelets remarkably appropriate for nonlinear dynamical system modelling. The most commonly used B-splines are those of orders 1 to 4, which are shown in Table 1.

4.3 Multiresolution wavelet networks

It has been proved in [48] that for a high dimensional problem, the multiresolution decomposition (37) and (40) may involve a large number of wavelet basis functions. A dimensional-reduced wavelet network based on a functional expansion was then proposed [48], to overcome the difficulty of the curse-of-dimensionality. For a given d-dimensional function approximation problem, the functional expansion advocated by [5], [27], [49], [50] is given below

$$\begin{aligned} & f(x_1, x_2, \cdots, x_d) \\ & = \sum_{i=1}^{d} f_i(x_i) + \sum_{1 \le i < j \le d} f_{i,j}(x_i, x_j) + \sum_{1 \le i < j < k \le d} f_{i,j,k}(x_i, x_j, x_k) + \cdots + \sum_{1 \le i_1 < \cdots < i_m \le d} f_{i_1, i_2, \cdots, i_m}(x_{i_1}, x_{i_2}, \cdots, x_{i_m}) \end{aligned} \tag{47}$$

where $m \le d$, $i_m \in \{1, 2, \cdots, d\}$ and the function $f_{1,2,\cdots,j}(\cdot)$ (j=1,2, ...,d) does not contain terms that can be written as functional components with an order smaller than j. For more detailed discussion on the functional expansion (47), see [48]-[50].

In practice, many types of functions have been chosen to express the functional components $f_{1,2,\cdots,j}(\cdot)$ in model (47). In the present study, however, multiresolution wavelet decompositions will be used to approximate each of these functional components in the dimension-reduced functional expansion (47). For example, the decomposition (36) will be applied to express the one-dimensional component f_i for i=1,2, ..., d; the decomposition (40) will be applied to the higher dimensional components $f_{i_1,\cdots,i_m}$ with $m \leq d$ and $i_m \in \{1,2,\cdots,d\}$. B-splines and associated wavelets will be used as the basis functions in these multiresolution decompositions. The resultant model is referred to as the truncated *multiresolution wavelet network*.

4.4 The determination of the scale and location parameters in the wavelet networks

Assume that a function $f \in L^2(\mathbf{R}^d)$ of interest is defined in the hypercube $[a,b]^d$. Without loss of generality, consider the case where a=0 and b=1. For the sake of convenience, the one-dimensional (d=1) case will be considered as an example to illustrate the determination of scale and location parameters in the B-spline and associated wavelet based multiresolution networks.

It is known that both B-splines and associated wavelets are compactly supported and the support for the s-th order B-spline and associated mother wavelet are [0, s] and [0, 2s-1], respectively. At any given scale j in the decomposition (36), only the scale basis functions satisfying $0 < 2^j x - k < s$ are needed, where $0 \leq x \leq 1$. This implies that the location parameter k satisfies $1 - s \leq k \leq 2^j - 1$. Similarly, only the wavelet basis functions satisfying $2 - 2s \leq k \leq 2^j - 1$ are needed at scale j. Therefore, the choice of the location parameter k at scale j in the decomposition is determined by the associated scale parameter j. The scale parameter is thus a key factor in multiresolution wavelet networks. In the following, the scale parameter determination problem will be addressed from two aspects: static function learning and dynamical modelling.

4.4.1 Static function learning

For a static function approximation problem, where the independent variable t is 'time', the initial scale j_0 in the multiresolution wavelet decomposition (36) is often set to zero, and the finest scale $j_{\max}$ can be chosen as $j_{\max}$=int[$\log_2(\beta f_{\max})$], where $f_{\max}$ is the maximum natural frequency of the signal involved, β is a positive number, and int[$\cdot$] denotes taking the integer value of the corresponding number. Results on numerous simulation experiments show that for most signals β can be chosen between 2 and 16. To illustrate the relationship between the scale parameter j and the natural frequency of a signal, consider the two examples given below.

$$s_1(t) = \sin(8\pi t) + \sin(16\pi t) + \sin(64\pi t) \tag{48}$$

$$s_2(t) = \sum_{k=1}^{5} \lambda^{k(\alpha-2)} \sin(c_k \lambda^k t) \tag{49}$$

where $\lambda = 2.5$, $\alpha = 1.1$, $c_k = 5$ (k=1,2,3), $c_4 = 4$, and $c_5 = 2$. The maximum frequency of the signals $s_1(t)$ and $s_2(t)$ is 32Hz and 31.1 Hz, respectively. Both the signals were sampled with a sampling interval 0.002sec over [0,1], and 500 data points were recoded for each of the signals. Based on the recorded data points, the two signals were reconstructed using a multiresolution decomposition below

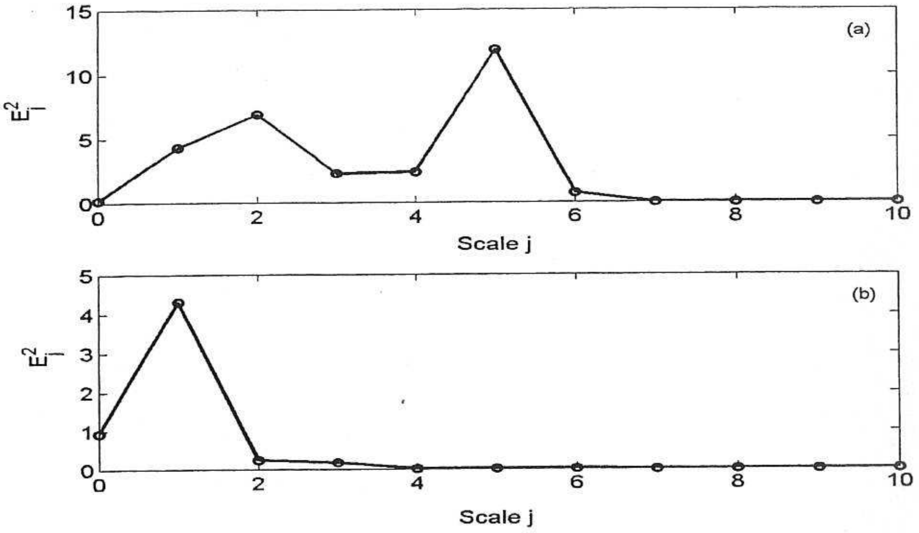

Fig. 1 The wavelet energy for signals $\hat{s}_1(t)$ and $\hat{s}_2(t)$ described by (48) and (49). (a) The energy for the signal recovered for $\hat{s}_1(t)$; (b) the energy for the recovered for $\hat{s}_2(t)$.

$$\hat{s}(t) = \sum_{k\in A} a_{0,k}\phi_{0,k}(t) + \sum_{j=0}^{J}\sum_{k\in B_j} d_{j,k}\psi_{j,k}(t) \qquad (50)$$

where J=10, ϕ and ψ are the 4-th order B-spline and the associated mother wavelet, the index sets A and B_j can easily be determined at any given scale j using the method mentioned above. Define the *wavelet energy* as

$$E_0^2 = \sum_{k\in A}\left|a_{0,k}\right|^2 + \sum_{k\in B_0}\left|d_{0,k}\right|^2 \qquad (51a)$$

$$E_j^2 = \sum_{k\in B_j}\left|d_{j,k}\right|^2,\ j \geq 1, \qquad (51b)$$

The wavelet energy of the signals $\hat{s}_1(t)$ and $\hat{s}_2(t)$ are shown in Fig. 1, which shows that the wavelet energy for $\hat{s}_1(t)$ and $\hat{s}_2(t)$ are distributed at scaling levels 0 to 6, and 0 to 5, respectively. In other words, the signal $\hat{s}_1(t)$ is totally determined by the decomposition with scales 0 to 6, and the signal $\hat{s}_2(t)$ is totally determined by the decomposition with scales 0 to 5. Note that the wavelet energy E_5^2 for signal $\hat{s}_2(t)$ at scale 5 in Fig. 1(b) is 0.0006, which is very small.

4.4.2 Dynamical modelling

In typical discrete time-invariant dynamical modelling, for instance, the NARX modelling, the direct independent input variables are usually the lagged inputs and outputs of the system under study, but not the time 't' as in a static function. Consider a simple example given below.

$$x(t) = -\sqrt{|x(t-1)|} + 0.2[2u(t-1)-1]^3 + \eta(t) \tag{52a}$$

$$y(t) = 1.5 + x(t) + \eta(t) \tag{52b}$$

where $\eta(t)$ was a Gaussian white noise sequence with mean zero and standard deviation 0.01. The input $u(t)$ was assumed to be bounded in the interval [0,1]. The objective here is to find an equivalent representation, using the NARX representation $y(t) = f(y(t-1), u(t-1)) + e(t)$, for the original model (52). Consider the first order functional expansion

$$f(y(t-1), u(t-1)) = f_y(y(t-1)) + f_u(u(t-1)) \tag{53}$$

The unknown nonlinear functions f_1 and f_2 can then be approximated using a multiresolution decomposition below

$$f_x(x(t)) = \sum_{k\in A} a_{0,k}\phi_{0,k}(x(t)) + \sum_{j=0}^{J}\sum_{k\in B_j} d_{j,k}\psi_{j,k}(x(t)) \tag{54}$$

where J =8, $x(t) = y(t-1)$ or $x(t) = u(t-1)$, and the index sets A and B_j are defined as in (50). Setting the input $u(t)$ in the model (52) as an random sequence uniformly distributed in [0,1], 400 input and output data points were collected and were then used to select the wavelet basis and to estimate the unknown parameters. Although a total of 614 wavelet basis functions were involved in the initial network, only 7 basis functions were selected using the AOLS algorithm. It has been shown that the model formed by the 7 selected basis functions, $\phi_{0,-1}(y(t-1))$, $\psi_{0,-3}(y(t-1))$, $\phi_{0,0}(u(t-1))$, $\phi_{0,-1}(u(t-1))$, $\psi_{0,-1}(u(t-1))$, $\psi_{1,-4}(u(t-1))$, and $\psi_{1,-6}(u(t-1))$, provided an excellent approximation for the original model (52) and produced excellent model predicted outputs. Clearly, the scales for these basis functions are concentrated on 0 and 1. Numerous simulation experiments show that when B-spline-wavelet networks are applied to identify nonlinear dynamical systems, the initial scale j_0 in the multiresolution wavelet decomposition (36) can be set to zero, and the finest scale $j_{\max}$ can often be chosen as an integer that is not larger than 5.

5. A Three-Phase Modelling Procedure to Implement the NARMAX Model

As mentioned in Section 2, a typical NARMAX model often consists of two parts: the deterministic (noise independent) and the stochastic (noise correlated) submodels as shown by (6). In the present study, a three-phase modelling approach is proposed to construct basis function networks to implement the NARMAX model. The main idea of the three-phase modelling scheme is as follows:

- Initially, construct a NARX model using basis function networks.
- The effects of correlated noise and unmeasured disturbances must be accommodated using the model residuals (errors) from the identified NARX model. Viewing the modelling error $\varepsilon(t)$ as the output and treating the lagged system outputs $y(t-i)$ and inputs $u(t-j)$, coupled with the lagged error variables $\varepsilon(t-k)$, as the inputs, fit a polynomial model for the error $\varepsilon(t)$.

- Combine the identified error model with the network NARX model, and re-estimate all the model parameters recursively. An unbiased model should then be obtained.

5.1 Implementation of the NARX model using basis function networks

For convenience of description, take the case of SISO dynamical nonlinear system modelling as an example. For a given identification problem, the objective is to build a basis function network to identify the unknown nonlinear mapping f_{yu} in (6). Assuming that N input-output data points, $\{u(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$ have been observed, let $d = n_y + n_u$ and $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]^T$ with

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y + 1 \le k \le n_y + n_u \end{cases} \tag{55}$$

The nonlinear function $f_{yu}(\mathbf{x}(t))$ can be approximated using any basis function networks including RBF and multiresolution wavelet models. A typical choice for the radial basis functions in RBF networks is a set of standard Gaussian kernels, $\varphi_i : \mathbf{R}^d \mapsto \mathbf{R}$ in the sense that

$$\varphi_i(\mathbf{x}(t); \mathbf{a}_i, \mathbf{b}_i) = \varphi_i(\mathbf{a}_i^T \circ (\mathbf{x}(t) - \mathbf{b}_i)) = \exp[-(\mathbf{x}(t) - \mathbf{b}_i)^T \Lambda_i^{-1} (\mathbf{x}(t) - \mathbf{b}_i)/2] \tag{56}$$

where $\mathbf{a}_i = [a_{i,1}, \cdots, a_{i,d}]^T$, $\mathbf{b}_i = [b_{i,1}, \cdots, b_{i,d}]^T$ and $\Lambda_i = diag[a_{i,1}, \cdots, a_{i,d}] = diag[\sigma_{i,1}^2, \cdots, \sigma_{i,d}^2]$. The basis function network (8) can then be written as

$$g(\mathbf{x}(t)) = \sum_{i=1}^{M} \theta_i e^{-\frac{1}{2}(\mathbf{x}(t)-\mathbf{b}_i)^T \Lambda_i^{-1} (\mathbf{x}(t)-\mathbf{b}_i)} \tag{57}$$

Assume that a total of m_{yu} significant basis functions (model terms formed by polynomials, RBFs or wavelets) are selected for the nonlinear function $f_{yu}(\mathbf{x}(t)) = f_{yu}(\mathbf{y}^{[t-1,n_y]}, \mathbf{u}^{[t-1,n_u]})$ in (6), $f_{yu}(\mathbf{x}(t))$ can thus be approximated as

$$\hat{f}_{yu}(\mathbf{x}(t)) = \sum_{m=1}^{m_{yu}} \hat{\theta}_{k_m} \varphi_{k_m}(\mathbf{x}(t); \mathbf{a}_i, \mathbf{b}_i) \tag{58}$$

5.2 Noise modelling

Again, for convenience of description, the SISO case will be taken as an example. In many cases the noise terms in the NARMAX model (1) will form a correlated or coloured noise sequence. This is likely to be the case for most real data sets. In this case the approximation (58) is likely to fail to give a sufficient description due to the bias in the parameter estimates. The effects of correlated noise and unmeasured disturbances must then be characterized by modelling the residuals with respect to the identified NARX model. The NARX modelling error is defined as

$$\varepsilon(t) = y(t) - \hat{f}_{yu}(\mathbf{x}(t)) \tag{59}$$

The residual signal $\varepsilon(t)$ can then be related to the input $u(t)$ and the out $y(t)$ by a nonlinear model. In the present study, the following polynomial model of degree ℓ is applied to model the residual sequence $\varepsilon(t)$

$$\begin{aligned}\varepsilon(t) &= f_{yue}(y^{[t-1,n_y]}, u^{[t-1,n_u]}, \varepsilon^{[t-1,n_e]}) \\ &= \sum_{i_i=1}^{d} \gamma_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{d} \sum_{i_2=i_1}^{d} \gamma_{i_1 i_1} x_{i_1}(t) x_{i_2}(t)) + \cdots \\ &+ \sum_{i_1=1}^{d} \cdots \sum_{i_\ell=i_{\ell-1}}^{d} \gamma_{i_1 i_2 \cdots i_l} x_{i_1}(t) x_{i_2}(t) \cdots x_{i_\ell}(t) + \varepsilon_1(t)\end{aligned} \tag{60}$$

This form of model is used because of the system is nonlinear it is also highly likely that the noise will involve nonlinear cross product terms with both system input and the output. Assume that a total of m_{yue} significant basis functions are selected for the noise model, the nonlinear function $f_{yue}(\widetilde{\mathbf{x}}(t))$ $= f_{yue}(y^{[t-1,n_y]}, u^{[t-1,n_u]}, \varepsilon^{[t-1,n_e]})$ in (6) can then be approximated by

$$\hat{f}_{yue}(\widetilde{\mathbf{x}}(t)) = \sum_{m=1}^{m_{yue}} \hat{\gamma}_{k_m} p_{k_m}(\widetilde{\mathbf{x}}(t)) \tag{61}$$

where the extended regression vector $\widetilde{\mathbf{x}}(t)$ is defined similar to (5) with $e(t-k)$ being replaced by $\varepsilon(t-k)$, $p_m(\cdot)$ are selected model terms of the form $z_1^{i_1}(t)\cdots z_\ell^{i_\ell}(t)$, where $z_j^{i_j}(t) \in \{y(t-1),\cdots,\ y(t-n_y),$ $u(t-1),\cdots,\ u(t-n_u),\ \varepsilon(t-1),\cdots,\ \varepsilon(t-n_e)\}$ for j=1, ..., ℓ , with $0 \le i_j \le \ell$ and $0 \le i_1 + \cdots + i_\ell \le \ell$. Note that at least one $z_j^{i_j}(t)$ is related to $\varepsilon(t-k)$ for $k = 1,2,\cdots,n_e$.

5.3 Parameter re-estimation

In order to obtain an unbiased network model, the identified model $\hat{f}_{yu}$ and $\hat{f}_{yue}$ should be combined as a whole and the model parameters should then be re-calculated. Let $\pi_1(t),\cdots,\pi_{m_{yu}}(t)$ be the m_{yu} selected model terms in (58) with $\pi_m(t) = \pi_m(\mathbf{x}(t))$, and let $\Phi_{yu}(t) = [\pi_1(t),\cdots,\pi_{m_{yu}}(t)]$ and $\Phi_{yue}(t) = [p_1(t),\cdots,p_{m_{yue}}(t)]$. An unbiased model can often be obtained by re-estimating the model parameters in a recursive way below.

(a) Calculate the model parameter estimate $[\hat{\theta}_{yu}^T, \hat{\theta}_{yue}^T]^T$ of the model

$$y(t) = \Phi_{yu}(t)\theta_{yu} + \Phi_{yue}(t)\theta_{yue} \tag{62}$$

and let

$$\varepsilon_1(t) = y(t) - \hat{y}(\widetilde{\mathbf{x}}(t)) = y(t) - [\Phi_{yu}(t)\hat{\theta}_{yu} + \Phi_{yue}(t)\hat{\theta}_{yue}] \tag{63}$$

(b) If $\|\boldsymbol{\varepsilon}_1\| / \|\boldsymbol{\varepsilon}\| \approx 1$, stop the parameter re-estimation procedure; otherwise, go to (c).

(c) Set $\{\varepsilon(t)\}_{t=1}^N = \{\varepsilon_1(t)\}\}_{t=1}^N$, repeat (a).

Note that the residual signal defined by (59) and (63) is in fact the one-step-ahead prediction error, which is different from the often used model prediction error defined as

$$\hat{\varepsilon}(t) = y(t) - \hat{y}_{mpo}(t) \tag{64}$$

where $\hat{y}_{mpo}(t)$ is recursively calculated from an identified estimator $\hat{f}_{yu}$ from some given initial values in the sense that

$$\hat{y}(t) = \hat{f}_{yu}(\hat{y}_{mpo}(t-1), \cdots, \hat{y}_{mpo}(t-n_y), u(t-1), \cdots, u(t-n_u)) \tag{65}$$

A key step following the above three-phase modelling is model validation. A commonly used approach to check the validity of the identified model is to use higher order statistical correlation analysis [51], [52]. An alternative for model validity tests is to check both the short and the long term predictive ability of the model.

6. Numerical Examples and Results

In this section three examples are described to illustrate the applicability and effectiveness of the new AOLS learning algorithm for basis function network training when these are applied to identify nonlinear dynamical systems. In all the three examples, significant model terms were selected and the model size was determined using the new AOLS algorithm. A comparison between RBF and wavelet models is presented.

6.1 Example 1—a chaotic time series

The following piecewise autoregressive model

$$\tilde{y}(t) = \begin{cases} 2\alpha[\tilde{y}(t-1)+1] & -0.5 \le \tilde{y}(t-1) \\ -2\alpha\tilde{y}(t-1) & -0.5 \le \tilde{y}(t-1) < 0.5 \\ 2\alpha[\tilde{y}(t-1)-1] & 0.5 \le \tilde{y}(t-1) \end{cases} \tag{66}$$

where $\alpha = 0.95$, was simulated and 5000 data points were generated with an initial condition $\tilde{y}(0) = 0.5$. A Gaussian white noise $e(t)$ with mean zero and standard derivation 0.02 was then added to the data set to form a noisy data set $y(t) = \tilde{y}(t) + e(t)$. The first 200 noisy data points were used for network training, and the remaining 4800 data points were used for model testing. The first return map produced by the 4800 noisy data points is shown in Fig. 2(a). The input variable was chosen as $x(t) = y(t-1)$ to construct basis function networks as below.

- Start from a full Gaussian RBF model, with mixed basis functions $\varphi_{1i}(t) = \exp\{-\|x(t) - \tilde{c}_i\|^2 / (2\sigma_{1i}^2)\}$ and $\varphi_{2i}(t) = \exp\{-\|x(t) - \tilde{c}_i\|^2 / (2\sigma_{2i}^2)\}$, where all the data points in the estimation data set were chosen as candidate centres, $\tilde{c}_i$, the kernel widths σ_{1i} and σ_{2i} were heuristically chosen as follows: many values for σ_{1i} and σ_{2i} were initially tested and it was found that the values $\sigma_{1i}^2 = r_y^2/8 \approx 0.0625$ and $\sigma_{1i}^2 = r_y^2/4 \approx 0.25$ for i=1, …, 199 were a good choice, since the RBF network model with these two selected kernel widths produced smaller mean-squared-errors. The initial noise model was given by (60), where $n_y = 1$, $n_u = 0$, $n_e = 5$ and ℓ=2. The finally identified noise independent RBF model was

$$y(t)=\sum_{i=1}^{11}\theta_i\varphi_i(y(t-1),c_i,\sigma_i)=\sum_{i=1}^{11}\theta_i\exp\{-[y(t-1)-c_i]^2/(2\sigma_i^2)\} \tag{67}$$

where the estimated model parameters θ_i, the centres c_i and the kernel width σ_i in the associated Gaussian basis functions are as follows: $\{\theta_i\}_{i=1}^{11}$={-0.98899943,-1.63597326,-0.36124466,-0.97257050, -1.11997304, 3.45468596, -3.20488120, -0.34953056, 1.76833004, -0.43156670, 1.30007889}, $\{c_i\}_{i=1}^{11}$ ={0.8211, 0.5113, 0.4585, 0.1487, -0.3294, -0.4983, -0.5337, -0.9827, 0.5113, 0.3418, -0.4670}, and $\sigma_i=0.1768$ for i=1, ..., 8 and $\sigma_i=0.3536$ for i=9, 10, 11.

- Start from a full multiresolution wavelet model of the form (54), where J=5, ϕ and ψ are the second order B-spline and the associated mother wavelet. The initial noise model was given by (60), where $n_y=1$, $n_u=0$, $n_e=5$ and ℓ=2. The finally identified noise independent wavelet model was

$$\begin{aligned}y(t)=&-0.47653994\,\phi_{0,0}(y(t-1))+0.47473611\,\phi_{0,-2}(y(t-1))\\&-0.15802597\,\psi_{0,-1}(y(t-1))+0.15851635\,\psi_{0,-2}(y(t-1))\end{aligned} \tag{68}$$

The time series generated from model (66) is chaotic and is strongly sensitive to the initial conditions. A comparison between the model predicted outputs or many step ahead forecasts, produced from the identified models, and the original data points produced directly from the original model (66) is therefore difficult. As an alternative, the first return maps, produced by the identified models, were therefore used to test the validity of the identified models. Starting from the same initial condition $y(0)=0.5$, the two identified models (67) and (68) were simulated and 5000 data points were collected for each model. The first return maps constructed using the simulated data points (from t=201 to 5000) are shown in Fig. 2(b) and (c). It is clear from Fig. 2 that the identified wavelet model is superior to the RBF model for this unsmooth and discontinuous chaotic time series. In fact, the first return map produced from the wavelet identified model is identical to that produced by the original model (66).

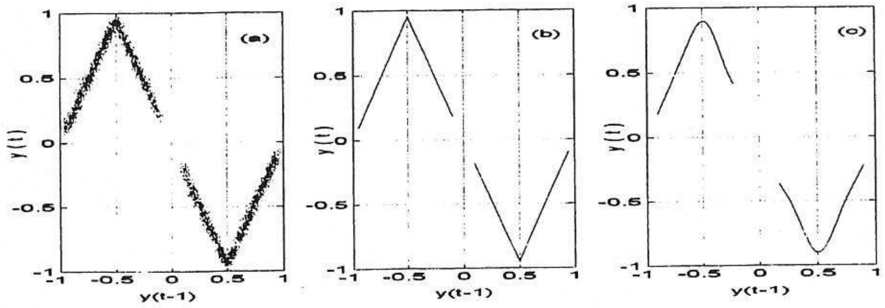

Fig. 2 The first return maps for the time series produced by the original model (66) and by the identified models (67) and (68). (a) from original model (66) with added noise; (b) from the wavelet model (68); (c) from the RBF model (67). All the four maps were formed using 4800 data points.

6.2 Example 2—a rational model

Consider a rational system [35] described by the model

$$y(t) = \frac{y(t-1)y(t-2)}{a^2 + y^2(t-1) + y^2(t-2)} + \frac{u(t-1)}{b^2 + c^2u^2(t-1)} + \eta(t) \tag{69}$$

where a=b=1, c=0.1, the input $u(t)$ was assumed to be bounded in [-10, 10], and $\eta(t)$ was a noise determined by

$$\eta(t) = w(t) + 0.95w(t-1) - 0.25w(t-2) - 0.6w(t-3) \tag{70}$$

with $w(t)$ a Gaussian white noise of zero mean and a standard variation $\sigma_w^2 = 0.25$. The model was simulated by setting the input signal $u(t)$ as a random sequence uniformly distributed in [-10, 10] and 1000 input-output data points were collected. The first 500 data points were used for model estimation and the remainder were used for model testing. Note that when identifying the polynomial and RBF models, the original observational data points were used. In the wavelet modelling, however, the original input-output observations were normalized to [0,1]. From the information that $\underline{u} \le u(t) \le \overline{u}$ and $\underline{y} \le y(t) \le \overline{y}$, where $\overline{u} = -\underline{u}$ =10, $\overline{y} = -\underline{y}$ =8, both the observed input and the output sequences were normalized to [0,1] via a simple transformation. The normalized input and output variables will still be denoted as $u(t)$ and $y(t)$, respectively. The final output was recovered to the original range by using the inverse transform.

The input vector was set to $\mathbf{x}(t) = [y(t-1), y(t-2), u(t-1)]^T$ for the NARX modelling, and the extended input vector was set to $\widetilde{\mathbf{x}}(t) = [y(t-1), y(t-2), u(t-1), \varepsilon(t-1), \varepsilon(t-2), \varepsilon(t-3)]^T$ for the noise modelling.

Following the three-phase modelling approach described in Section 5, a multiresolution wavelet model was identified. For a comparison, polynomial and RBF models were also identified.

(A) The multiresolution wavelet model.

Start from a full multiresolution wavelet model of the form

$$\begin{aligned} f_{yu}(\mathbf{x}(t)) &= f_1(x_1(t)) + f_2(x_2(t)) + f_2(x_3(t)) \\ &\quad + f_{1,2}(x_1(t), x_2(t)) + f_{1,3}(x_1(t), x_3(t)) + f_{2,3}(x_2(t), x_3(t)) \\ &\quad + f_{1,2,3}(x_1(t), x_2(t), x_3(t)) \end{aligned} \tag{71}$$

where $f_p(x_p(t))$ for p=1,2,3 were decomposed as (54), the finest scale was set to J=4, and the second and the third function components $f_{p,q}(x_p(t), x_q(t))$ $(1 \le p < q \le 3)$ and $f_{1,2,3}(x_1(t), x_2(t), x_3(t))$ were decomposed as

$$f_{p,q}(x_p(t), x_q(t)) = \sum_{k_1,k_2} \alpha_{0;k_1,k_2} \phi_{0,k_1}(x_p(t))\phi_{0,k_2}(x_q(t)) + \sum_{k_1,k_2} \alpha_{1;k_1,k_2} \phi_{1,k_1}(x_p(t))\phi_{1,k_2}(x_q(t)) \tag{72}$$

$$\begin{aligned} f_{1,2,3}(x_1(t), x_2(t), x_3(t)) &= \sum_{k_1,k_2,k_3} \alpha_{0;k_1,k_2,k_3} \phi_{0,k_1}(x_1(t))\phi_{0,k_2}(x_2(t))\phi_{0,k_3}(x_3(t)) \\ &\quad + \sum_{k_1,k_2,k_3} \alpha_{1;k_1,k_2,k_3} \phi_{1,k_1}(x_1(t))\phi_{1,k_2}(x_2(t))\phi_{1,k_3}(x_3(t)) \end{aligned} \tag{73}$$

In (71)-(73), ϕ and ψ are the 4th-order B-spline and the associated mother wavelet. The initial noise model was given by (60), where $n_y = 2$, $n_u = 1$, $n_e = 3$ and ℓ=2. The final identified wavelet model was

$$\begin{aligned} y(t) = &-9.9765\phi_{0,-3}(y(t-1))\phi_{0,0}(y(t-2)) - 0.1027\phi_{1,0}(y(t-1))\phi_{1,-2}(y(t-2)) \\ &+1.1827\phi_{0,-1}(u(t-1)) - 5.2522\psi_{0,-5}(u(t-1)) + 0.3473\psi_{0,-2}(u(t-1)) - 0.5692\varepsilon(t-2) \\ &+0.6722\varepsilon(t-1)y(t-2) + 0.4127\varepsilon(t-1)u(t-1) + 0.2300u(t-1)\varepsilon(t-3) \end{aligned} \tag{74}$$

Note that all the variables in model (74) were pre-normalized. The output variable $y(t)$ can easily be recovered to its original operating region via a simple inverse transform. The NMSE (defined by (19)) of the model predicted outputs produced by the wavelet model (74) over the test data set, points from 501 to 1000, was 0.0505. Note that no cross product model terms with respect to the input and output variables were selected in (74), this is consistent with the structure of the original model (69).

(B) The polynomial model.

The polynomial modelling starts from a full model of degree 3 with 84 candidate model terms. The initial noise model was chosen as (60) with a nonlinear degree ℓ=2. The finally identified polynomial model was

$$\begin{aligned} y(t) = &-0.0385y(t-1) + 0.9151u(t-1) + 0.0251y(t-1)y(t-2) \\ &+0.0015y(t-1)u(t-1) + 0.0015y^3(t-1) - 0.0008y^3(t-2) \\ &+0.0004y(t-1)u^2(t-1) - 0.0044u^3(t-1) + 0.5825\varepsilon(t-1) \\ &-0.6141\varepsilon(t-2) + 0.1616\varepsilon(t-3) + 0.0304y(t-2)\varepsilon(t-1) \end{aligned} \tag{75}$$

The NMSE of the model predicted outputs produced by the identified polynomial model over the test data set, points from 501 to 1000, was 0.0531.

(C) The RBF model.

The RBF modelling starts from a full Gaussian RBF model, with basis functions

$$\varphi_j(\mathbf{x}(t)) = \exp\left\{-\frac{[x_1(t)-c_{j,1}]^2}{2\sigma_{j,1}^2} - \frac{[x_2(t)-c_{j,2}]^2}{2\sigma_{j,2}^2} - \frac{[x_3(t)-c_{j,3}]^2}{2\sigma_{j,3}^2}\right\} \tag{76}$$

where j=3, ...500, $[x_1(t), x_2(t), x_3(t)] = [y(t-1), y(t-2), u(t-1)]$, $c_{j,k}$ are kernel centres (all the data points in the estimation data set were chosen as candidate centres), the kernel width $\sigma_{j,k}$ were heuristically set to as $\sigma_{j,1} = \sigma_{j,2} = 6/\sqrt{2}$, $\sigma_{j,3} = 10/\sqrt{2}$. The full noise model was chosen as (60) with a nonlinear degree ℓ=2. The final identified RBF model contained 21 process model terms and 5 noise related model terms. The NMSE of the model predicted outputs produced by the identified RBF model over the test data set, points from 501 to 1000, was 0.0517. The selected centres, the estimated parameters for the noise independent model terms are listed in Table 2.

TABLE 2
THE SELECTED CENTERS AND ESTIMATED PARAMETERS FOR THE NOISE INDEPENDENT MODEL TERMS OF THE RBF MODEL FOR EXAMPLE 2

	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$	θ_j
1	-0.5198	-0.0052	9.7587	6.10298572e-01
2	0.0733	-0.1830	-6.8522	-5.14616041e+00
3	6.3857	2.7879	9.6704	4.57447061e+00
4	-2.3420	-2.3592	4.1460	4.74336317e+00
5	-5.0780	5.3060	-9.1272	-5.53278461e+00
6	7.0828	-1.7361	-9.7295	-2.57470429e+00
7	-3.7754	-4.9176	-4.6782	-7.29073581e+00
8	4.8757	6.3564	-8.7013	-3.84545119e+00
9	-3.5588	6.1964	8.5410	3.74115367e+00
10	-1.6217	-4.2519	9.2070	4.58380816e+00
11	-0.0140	-4.3961	-7.8379	-2.94773805e+00
12	5.4803	-5.4308	8.3219	3.39948056e+00
13	-0.5899	-1.6691	8.1653	-4.89696779e+00
14	6.1964	-5.4268	-7.5638	-1.63678934e+00
15	4.8875	3.3054	6.6301	7.07969308e-01
16	-0.4658	1.5105	-8.9139	4.77193075e+00
17	-6.5893	-4.5160	-9.8072	-7.96276493e+00
18	-0.5565	7.5010	-4.9753	-2.64513534e+00
19	-4.1618	-4.8630	-7.4391	1.04446119e+01
20	-6.0931	-4.4730	6.3917	1.88267519e+00
21	1.7748	5.0005	4.1891	1.94600598e+00

To check the model performance further, both the original model (69) and the identified polynomial, RBF and wavelet models were simulated with a same initial condition by setting the noise as zero (noise free models were considered to facilitate the comparison between different model predicted outputs), and by setting the input signal as

$$u(t) = \begin{cases} 5\sin(12\pi t/200) & 0 \le \mathrm{t} \le 60 \\ -5+5\sin(20\pi t/200) & 60 < \mathrm{t} \le 120 \\ 8 & 120 < \mathrm{t} \le 150 \\ -8 & 150 < \mathrm{t} \le 175 \\ 5\sin(6\pi t/200)+5\sin(20\pi t/200) & 175 < \mathrm{t} \end{cases} \qquad (77)$$

The model predicted outputs generated from the identified polynomial, RBF and wavelet models were compared with the noise free output produced by the original model (69). These outputs are shown in Fig. 3, from which it can be seen that all the identified models produce satisfactory model predicted outputs. At the lower frequency and lower amplitude range, the difference between the model predicted outputs are very slight. At large positive amplitudes of the input, the polynomial model produces an over response. At sharply varying points, the response of the RBF model oscillates slightly. The wavelet model performs consistently well across all excitation regimes.

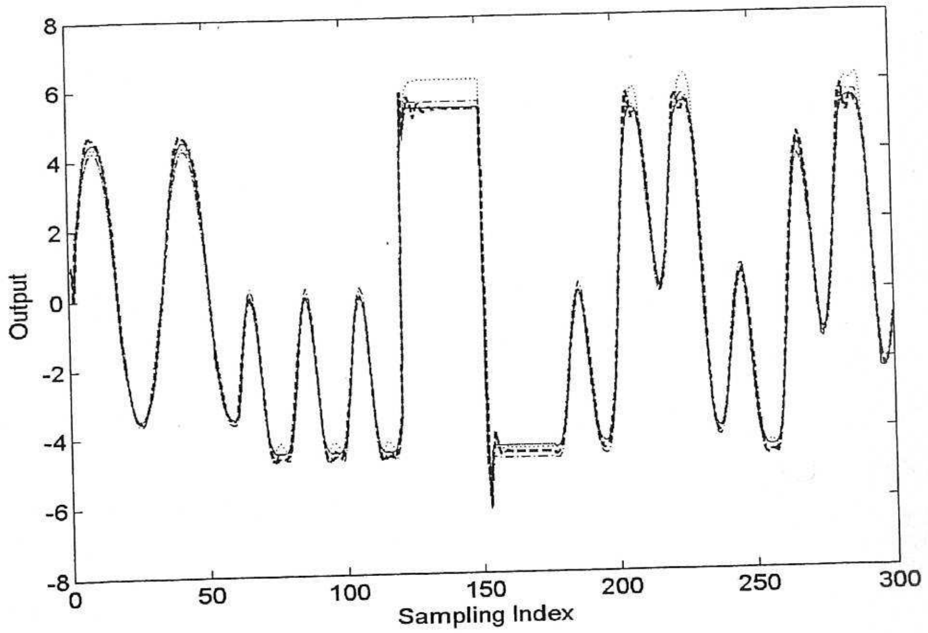

Fig. 3 Model predicted outputs for the system described by (69) under the input given by (77). The thin solid line '–' indicates the noise free output from the original model (69); the bold dashed line '--' indicates the model predicted output from the identified RBF model; the thin dotted line '.' indicates the model predicted output from the polynomial model; the thin dot-dashed line '.-' indicates the model predicted output from the wavelet model.

6.3 Example 3—a terrestrial magnetosphere dynamical system

One of the main problems of solar terrestrial physics is to understand the global dynamics of the terrestrial magnetosphere under the influence of the solar wind. The sun is a source of a continuous flow of charged particles, ions and electrons called the solar wind. The terrestrial magnetic field shields the Earth from the solar wind, and forms a cavity in the solar wind flow that is called the terrestrial magnetosphere. The magnetopause is a boundary of the cavity, and its position on the day side (sunward side) of the magnetosphere can be determined as the surface where there is a balance between the dynamic pressure of the solar wind outside the magnetosphere and the pressure of the terrestrial magnetic field inside. A complex current system exists in the magnetosphere to support the complex structure of the magnetosphere and the magnetopause. Changes in the solar wind velocity, density or magnetic field lead to changes in the shape of the magnetopause and variations in the magnetospheric current system. In addition if the solar wind magnetic field has a component directed towards the south a reconnection between the terrestrial magnetic field and the solar wind magnetic field is initiated. Such a reconnection results in a very drastic modification to the magnetospheric current system and this phenomenon is referred to as magnetic storms. During a magnetic storm, which can last for hours, the magnetic field on the Earth's surface will change as a result of the variations of the magnetospheric current system. Changes in the magnetic field induce considerable currents in long conductors on the terrestrial surface such as power lines and pipe-lines. Unpredicted currents in power lines can lead to blackouts of huge areas, the Ontario Blackout is just one recent example. Other undesirable effects include increased radiation to crew and passengers on long flights, and effects on communications and radio-wave propagation. Forecasting geomagnetic storms is therefore highly desirable and can aid the prevention of such effects.

The *Dst* index is used to measure the disturbance of the geomagnetic field during magnetic storms. Therefore, the *Dst* index provides useful information for studying geomagnetic storms. The forecasting of the *Dst* index is very important in helping to prevent the negative effects of geomagnetic storms. Fig. 4 shows 1000 data points of measurements of the solar wind parameter VB_S (input, measured unit: mV/m) and the *Dst* index (output, measured unit: nT) with a sampling interval T=1hour. Inspection of the Fig. 3 shows that several strong magnetospheric storms(*Dst* < -100 nT) and substrong stroms (*Dst* < -50 nT) took place during the time period under investigation. This data set was separated into the estimation set consisting of 500 input-output data points and the validation set consisting of the remaining data points. The objective was to identify input-output nonlinear basis function network models based on the estimation data set. This model was then used to predict the *Dst* index. Previous studies have shown that the data set shown in Fig. 4 can be adequately fitted by choosing the input vector as $\mathbf{x}(t)=[y(t-1),\cdots,y(t-4),u(t-1),u(t-2)]^T$, where $y(\cdot)$ and $u(\cdot)$ indicate the measurements of the system output (*Dst*)and input (VB_S), respectively.

Starting from a full wavelet model described by (71)-(73), where ϕ and ψ are the second-order B-spline and the associated mother wavelet, a multiresolutioin wavelet network was trained. The final identified wavelet model with the noise related model terms omitted was given as

$$\begin{aligned} y(t)=0.7146\,\phi_{0,0}(y(t-1)) &+0.0575\,\phi_{0,-1}(y(t-1))+0.0096\,\psi_{2,0}(y(t-1)) \\ &-0.0053\,\psi_{2,1}(y(t-1))+0.0738\,\psi_{1,-2}(y(t-2))+0.0802\,\psi_{2,3}(y(t-2)) \\ &+0.3462\,\psi_{2,-2}(y(t-2))+0.1403\,\phi_{0,0}(y(t-3))+0.0308\,\psi_{0,-2}(y(t-4)) \\ &+0.0018\,\psi_{2,1}(y(t-4))-0.0402\,\psi_{0,-2}(u(t-1))+0.0031\,\psi_{1,-1}(u(t-1)) \\ &-0.6165\,\psi_{2,3}(u(t-1))+0.0076\,\psi_{1,-2}(u(t-2)) \end{aligned} \tag{78}$$

Note again that all the variables in model (78) were pre-normalized into [0, 1]. If necessary, the output variable $y(t)$ could be recovered to its original operating region via a simple inverse transform. The NMSE of the model predicted outputs produced by the wavelet model (78) over the test data set, points from 501 to 1000, was 0.1094.

The 2 hour ahead prediction given by model (78) over the range from 750 to 100 hrs is shown in Fig. 5(a).

The RBF modelling starts from a full Gaussian RBF model with basis functions

$$\varphi_j(\mathbf{x}(t))=\exp\left\{-\sum_{k=1}^{6}\frac{[x_k(t)-c_{j,k}]^2}{2\sigma_{j,k}^2}\right\} \tag{79}$$

where j=5, …500, $x_k(t)=y(t-k)$ for k=1,2,3,4 and $x_k(t)=u(t-k+4)$ for k=5,6, $c_{j,k}$ are kernel centres (all the data points in the estimation data set were chosen as candidate centres). For this data set, numerical experimental results show that it was difficult to train a standard Gaussian kernel based RBF network with only a single common kernel width. In the present study, the kernel widths $\sigma_{j,k}$ were therefore set to two values: $\sigma_{j,k}=100/\sqrt{2}$ for the system output (k=1,2,3,4) and $\sigma_{j,k}=6.5/\sqrt{2}$ for the system input (k=5,6). The final identified RBF model contains 24 process model terms. The NMSE of the model predicted outputs produced by the identified polynomial model over the test data set, points from 501 to 1000, was 0.1139. The selected centres, the estimated parameters for the noise independent model terms are listed in Table 3. The 2-hour ahead prediction given by the identified RBF model over the range from 750 to 100 hrs is shown in Fig. 5(b). It is believed that the discrepancy between the model predictions and the measured values of the *Dst* index results from the effects of other inputs which affect the system output but were not included in the current model [53].

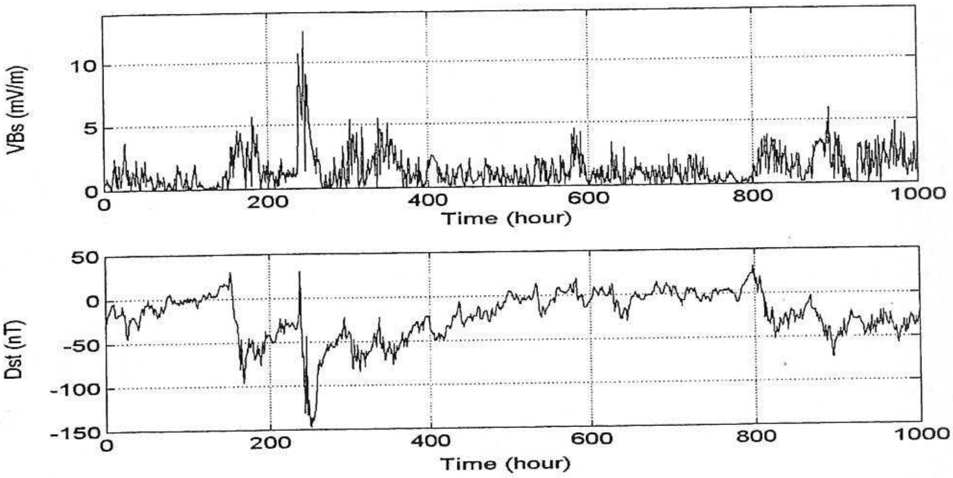

Fig. 4 The measurements for the input (VBs) and output (Dst) of the terrestrial magnetosphere dynamical system described in Example 3.

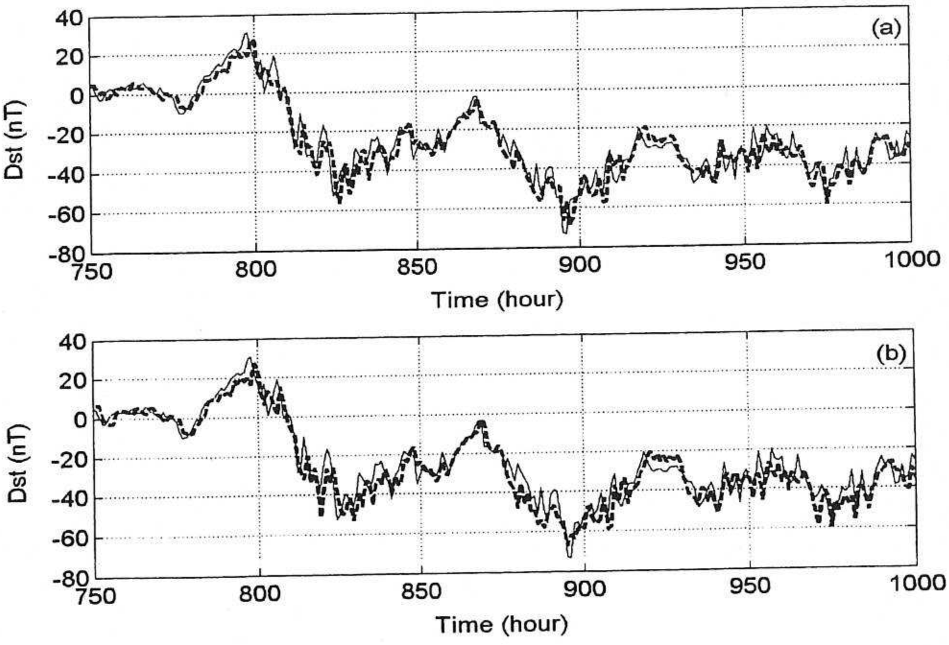

Fig. 5 A comparison between 2 hour ahead predictions and the measurements for the Dst index of the terrestrial magnetosphere dynamical system described in Example 3. (a) 2 hour ahead prediction produced by the identified wavelet model (78); (b) 2 hour ahead prediction produced by the identified RBF model given in Table 3. The thin solid line indicates the measurements; the bold dashed lines indicate 2 hour ahead predictions.

TABLE 3
THE SELECTED CENTERS AND ESTIMATED PARAMETERS FOR THE IDENTIFIED RBF MODEL TERMS FOR THE DST INDEX OF THE TERRESTRIAL MAGNETOSPHERE DYNAMICAL SYSTEM DESCRIBED IN EXAMPLE 3

	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$	$c_{j,4}$	$c_{j,5}$	$c_{j,6}$	θ_j
1	-87.3540	-96.2400	-81.9860	-66.7370	3.4843	3.1662	122.2151
2	-147.6290	-143.5500	-127.6130	-117.8910	5.3171	7.6633	25.3108
3	-61.3570	-67.6950	-2.8700	30.1600	7.2953	8.8522	3.7119
4	-43.3260	-110.6350	-131.3010	-70.7650	12.4846	9.8357	-112.7570
5	-66.0320	-61.0420	-76.3280	-78.7280	0.6540	0.1271	-67.1099
6	30.4300	19.0140	10.2830	14.9830	0.8600	0.0001	219.4544
7	-134.3520	-43.3260	-110.6350	-131.3010	1.3171	12.4846	-87.9296
8	-117.8910	-113.9410	-134.3520	-43.3260	9.0861	1.4930	-133.8170
9	-2.8700	30.1600	-11.9240	-25.6100	10.6937	5.6606	-51.1194
10	-113.9410	-134.3520	-43.3260	-110.6350	1.4930	1.3171	-123.6927
11	-78.5270	-49.6550	-42.7680	-39.4570	5.4342	4.1025	102.4736
12	-110.6350	-131.3010	-70.7650	-61.3570	9.8357	5.4094	110.7024
13	-53.1730	-61.9080	-67.7840	-68.4770	2.4990	2.6919	-35.4674
14	3.5210	6.5580	8.2720	6.9450	0.1978	0.0821	-874.5269
15	1.6320	2.3750	4.4950	2.7110	0.0660	0.2835	834.4480
16	-70.7650	-61.3570	-67.6950	-2.8700	6.1930	7.2953	-135.9411
17	-127.6130	-117.8910	-113.9410	-134.3520	8.7964	9.0861	-87.9641
18	-136.9530	-136.7980	-147.6290	-143.5500	4.3761	4.8883	-164.8797
19	-67.6950	-2.8700	30.1600	-11.9240	8.8522	10.6937	-32.4883
20	-69.8500	-53.4640	-46.4460	-47.8340	5.6776	0.1949	-115.5480
21	-131.3010	-70.7650	-61.3570	-67.6950	5.4094	6.1930	-82.4364
22	-3.7400	0.4570	-1.7990	0.6270	0.0747	0.8420	-146.2663
23	-50.1790	-62.0870	-62.6530	-69.1730	3.2360	0.0000	70.2239
24	-131.7670	-140.0310	-136.9530	-136.7980	3.3190	3.4475	71.4186

7. Discussions

Once the prototype basis functions have been chosen, basis function networks can often be constructed via either radial, tensor product, or ridge approaches to represent multivariate nonlinear mappings. RBF and wavelet models are among the most popular representations. The main advantage of RBF models is that the radial construction often leads to a smaller number of candidate regressors (model terms) compared with multiresolution wavelet models where compactly supported wavelets and tensor products are used. This advantage of RBF models becomes more significance for higher dimensional problems. For example, it was noted in [48] that for identification problems involving a large dimensionality the implementation of high-dimensional multiresolution wavelet networks via a tensor product approach can involve many potential model terms. But many of the problems associated with wavelet models can be mitigated when variable and term selection algorithm are used to determine best model subset. Compared with RBF network models, multiresolution wavelet models are more flexible and can be used to effectively describe not only smoothly varying ordinary nonlinear systems but also sharply varying severely nonlinear systems. Comparing multiresolution wavelet models with RBF models in detail, the following points are worth noting:

i) The compactly supported wavelet basis functions, for example, the B-splines and associated wavelets, define a hierarchical multiresolution structure with fixed and regular dilation-translation parameters. Thus the location and scale of each basis function is known beforehand. Although most radial basis functions are nearly compactly supported, they only vanish rapidly as the independent variables of these functions are far from the centre. Therefore, the scale and location parameters in RBF models have to be defined by means of a separate approach before hand or during the network training. While some efficient clustering algorithms are available for pre-selecting kernel centres to assistant RBF network training [12],[15],[29],[54], effective algorithms for selecting and optimizing the scale parameters are still needed to enhance the flexibility and generalization properties of RBF models.

ii) In a RBF model, every basis function depends on all the process variables. This is not always reasonable since in general it is not necessary that every variable of a process interacts directly with all the other variables. In the compactly supported wavelet multiresolutoin model, it is not required that every basis function (model term) include all the process variables. This allows more flexibility in selecting the correct model structure to capture the underlying nonlinear dynamics.

iii) In compactly supported multiresolution wavelet networks, the basis functions for instance the B-spline and associated wavelet are compactly supported. Thus, at a given resolution scale, the number of basis functions involved is deterministic and thus the total number of candidate model terms is determined by both the coarsest and the finest scale parameters. In RBF models, the number of total candidate model terms is dependent of either the length of the training data set (all the data pionts in the estimation data set are viewed as candidate centres), or the number of pre-selected centres and scales.

On the basis of the above discussions, some suggestions are given below:

i) *Observe the data as an initial step.* If some chirps, discontinuous points, or sharply varying trends are apparent in the signals, try a multiresolution wavelet model first. With remarkable inherent local properties, multiresolutioin wavelet models with compactly supported basis functions can usually provide more desirable results for severely nonlinear signals compared with other basis functions.

ii) *Dimensionality consideration.* For low dimensional problems, say the dimension is not higher than 5, multiresolution wavelet models are a good choice. For higher dimensional problems, say the dimension is larger than 10, a truncated multiresolution wavelet network given in Section 4.3 should be considered. Otherwise, a RBF network, or a radial wavelet network, where a radial basis function is chosen as the activation function, should provide an alternative.

iii) *Starting from a polynomial model.* The fact that RBF and wavelet models provide excellent representations for nonlinear systems does not mean that they should always be used. A general principle in system identification is that the model should be no more complex than is required to capture the underlying nonlinear dynamics. The well known parsimonious principle is particularly relevant in nonlinear system identification since the size of a nonlinear model constructed using local basis functions may easily become explosively large. It has been noted that a large class of nonlinear dynamics can be well described by polynomial models. Thus try a polynomial model first. If a polynomial model fails to capture the underlying nonlinear dynamics, then try other more complex models.

iv) *Select the significant model terms.* Whatever model form is selected experience shows that only a small subset of the total candidate term set are usually required to produce an excellent fit. Variable and term selection techniques should therefore always be employed.

v) *Hybrid basis function networks.* A hybrid basis function network, where different types of local and global basis functions are combined and integrated in some specified way, may produce an improved result. For example, a hybrid polynomial and wavelet model [55] can exploit the local property of wavelet basis functions and the global property of polynomials simultaneously, and can therefore gives a more parsimonious and flexible representation. Inspired by the hierarchical multiresolution structure of wavelet models with fixed and regular dilation-translation parameters, the capability of RBF networks may be greatly improved by introducing some multi-scale basis functions into the networks [56].

8. Conclusions

The construction and training of basis function networks for nonlinear dynamical modelling have been discussed in detail. A new adaptive orthogonal least squares (AOLS) algorithm has been developed for selecting significant model terms and determining the appropriate model size. Based on the new introduced criterion, the error-to-signal ratio (ESR), the new R^2-like statistic, and the adjustable prediction error sum of squares (R^2-APRESS), the new AOLS algorithm can correctly select the significant model terms and will automatically terminate by maximizing the values of R^2-APRESS.

The construction of multiresolution wavelet networks, where the compactly supported B-slines and associated mother wavelets are chosen as the basis functions, have been discussed in detail. With excellent time-frequency properties, the new wavelet networks can represent ordinary, as well as severely nonlinear dynamics with desirable approximation accuracy. This class of wavelet networks, however, can suffer from the curse-of-dimensionality, as observed for many other basis function networks. A truncated multiresolution wavelet network, where only lower dimensional functional components are employed to approximate an unknown function, is therefore often considered. For large dimensional problems, radial wavelet or RBF networks may be an excellent alternative.

The three-phase modelling approach, where the identification of a process model is followed by noise modelling, should produce an unbiased model by eliminating the effects of the noise on the model parameter estimates. The effectiveness of the new AOLS algorithm has been demonstrated by modelling both simulated and real data sets using RBF and multiresolution wavelet networks.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC(UK). We are grateful to Dr M. Balikhin for providing the magnetosphere data.

References

[1] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems.* New York: John Wiley & Sons,1980.

[2] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Procession.* New York: John Wiley & Sons, 2000.

[3] J. S. Albus, "A new approach to manipulator control: The cerebellar model articulation controller(CMAC)," *Trans. ASME J. Dyn. Sys. Meas. Control*, vol. 63, pp. 220-227, September 1975.

[4] M. J. D. Power, "Radial basis functions for multivariable interpolation: a review," in *Algorithms for approximation*, J. C. Mason and M. G. Cox, Eds., Oxford: Clarendon Press, pp. 143-167, 1987.

[5] J. H. Friedman,"Multivariate adaptive regression splines," *Ann. Stat.*, vol. 19, pp. 1-67, Mar. 1991.

[6] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[7] B. Scholkopf and A. J. Smola, *Learning with Kernel*. Cambridge: MIT Press, 2002.

[8] C. K. Chui, *An Introduction to Wavelets*. Boston: Academic Press, 1992.

[9] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11(7), pp. 674-693, July 1989.

[10] T. A. Johansen and R. Murray-Smith, "The operating regime approach to nonlinear modelling and control," in Multiple Model Approaches to Modelling and Control, R. Murray-Smith and T. A. Johansen, Eds., London: Taylor & Francis, 1997, pp. 3-73.

[11] D. S. Broomhead and D. Lowe, " Multivariable functional interpolation and adaptive networks," *Complex Systems*, vol. 2, pp. 321-355, 1988.

[12] J. Moody and C. Darken, " Fast learning in networks of locally tuned processing units," *Neural Comput.*, vol. 1, pp. 281-294, 1989.

[13] D. Lowe, "Adaptive radial basis function nonlinearities, and the problem of gneralization," in *IEE Int. Conference on Artificial neural Networks*, London, UK, 1989, pp. 171-175.

[14] E. J. Hartman, J. D. Keeler, and J. M. Kowalski, "Layered neural networks with Gaussian hidden units as universal approximations," *Neural Comput.*, vol. 2, pp. 210-215, 1990.

[15] S. Chen, S. A. Billings, C. F. N. Cowan, and P. M. Grant, "Practical identification of NARMAX models using radial basis functions," *Int. J. Control*, vol. 52, pp. 1327-1350, Dec. 1990.

[16] T. Poggio and F. M. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. 78, pp. 1481-1497, Sep. 1990.

[17] T. Poggio and F. M. Girosi, "Regularization algorithms for learning that are equivalent to multiplayer networks," Science, vol. 247, pp. 978-982, Feb. 1990.

[18] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Computation*, vol. 3, pp.246-257, 1991.

[19] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 302-309, March 1991.

[20] S. Chen, S. A. Billings, and P. M. Grant, "Recursive hybrid algorithm for nonlinear system identification using radial basis function networks," *Int. J. Control*, vol. 55, pp. 1051-1070, May 1992.

[21] S. Chen and S. A. Billings, "Neural networks for nonlinear dynamic system modelling and identification," *Int. J. Control*, vol. 56, pp. 319-346, Aug. 1992.

[22] D. Gorinevsky, "On the persistency of excitation in radial basis function network identification of nonlinear-systems," *IEEE Trans. Neural Networks*, vol. 6, pp. 1237-1244, Sep. 1995.

[23] Q. H. Zhang, and A. Benveniste, "Wavelet networks ," *IEEE Trans. Neural Networks*, vol. 3, pp. 889-898, Nov. 1992.

[24] L. Y. Cao, Y. G. Hong, H. P. Fang, and G. W. He, "Predicting chaotic time series with wavelet networks," *Physica* D, vol. 85, pp. 225-238, July 1995.

[25] Q. H. Zhang, "Using wavelet network in nonparametric estimation," *IEEE Trans. Neural Networks*, vol. 8, pp. 227-236, Mar. 1997

[26] S. A. Billings, and D. Coca, "Discrete wavelet models for identification and qualitative analysis of chaotic systems," *Int. J. Bifurcat. Chaos*, vol. 9, pp.1263-1284, July 1999.

[27] H. L. Wei, and S. A. Billings, "A unified wavelet-based modeling framework for nonlinear system identification: the WANARX model structure," *Int. J. Control*, vol. 77, pp.351-366, Mar. 2004.

[28] H. L. Wei, S. A. Billings, and M. A. Balikhin, "Prediction of the Dst index using multiresolution wavelet Models," *J. Geophysical Res.—Space Physics*, vol. 109, A07212-07224, July 2004.

[29] S. A. Billings and S. Chen, "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, 1998, pp. 231-278.

[30] I. J. Leontaritis, and S. A. Billings, "Input-output parametric models for non-linear systems—part I: deterministic non-linear systems," *Int. J. Control*, vol. 41, pp.303-328, 1985a.

[31] I. J. Leontaritis, and S. A. Billings, "Input-output parametric models for non-linear systems—part II: stochastic non-linear systems," *Int. J. Control*, vol. 41, pp.329-344, 1985b

[32] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, 1988, vol. 48, pp. 193-210, July 1988.

[33] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, pp.2157-2189, June 1989.

[34] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, pp.1873-1896, Nov.1989.

[35] R. K. Pearson, *Discrete-Time Dynamic Models.* Oxford: Oxford University Press, 1999

[36] L. X. Wang, and J. M. Mendel, "Fuzzy basis functions, universal approximations, and orthogonal least squares learning," *IEEE Trans. Neural Networks*, vol. 3, pp.807-814, Sep. 1992.

[37] X. Hong, and C. J. Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Trans. Neural Networks*, vol. 12, pp.435-439, Mar. 2001.

[38] X. Hong, P. M. Sharkey, and K. Warwick, "A robust nonlinear identification algorithm using PRESS statistic and forward regression," *IEEE Trans. Neural Networks*, vol. 14, pp.454-458, Mar. 2003.

[39] H. L. Wei, and S. A. Billings, "Identification and reconstruction of chaotic systems using multiresolution wavelet decompositions," *Int. J. Syst. Sci.*, vol. 35, pp. 511-526, July 2004.

[40] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716-723, Dec. 1974.

[41] G. Schwartz, " Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461-464, March 1978.

[42] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.*, vol. 11, pp. 416-431, June 1983.

[43] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, pp. 387-392, 1985.

[44] A. J. Miller, *Subset Selection in Regression.* London: Chapman and Hall, 1990.

[45] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, pp. 469-475, 1971.

[46] D. M. Allen, "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, vol. 16, pp. 125-127, 1974.

[47] R. Myers, *Classical and Modern Regression with Applications* (2nd Ed.). Boston: PWS-KENT Publishing Company, 1990.

[48] S. A. Billings and H. L. Wei, "A new class of wavelet networks for nonlinear system identification," Accepted for publication on *IEEE Neural Networks*, 2005.

[49] Z. H. Chen, "Fitting multivariate regression functions by interaction spline models," *J. Roy. Stat. Soc. B Met*, vol. 55, pp.473-491, 1993.

[50] G. Y. Li, C. Rosenthal, and H. Rabits, "High dimensional model representations," *J. Phys. Chem. A*, vol. 105, pp. 7765-7777, Aug. 2001.

[51] S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, vol. 44, pp.235- 244, July 1986.

[52] S. A. Billings and Q. M. Zhu, "Nonlinear model validation using correlation tests," *Int. J. Control*, vol. 60, pp.1107- 1120, Dec. 1994.

[53] O. M. Boaghe, M. A. Balikhin, S. A. Billings, and H. Alleyne, "Identification of nonlinear processes in the magnetosphere dynamics and forecasting of *Dst* index," *J. Geophysical Res.—Space Physics*, vol. 106, A12, pp. 30047-30066, Dec. 2001.

[54] S. Haykin, Neural Networks—A Comprehensive Foundation (2nd ed.). New Jersey: Prentice Hall, 1999.

[55] S. A. Billings and H. L. Wei, "The wavelet-NARMAX representation: a hybrid model structure combining polynomial models with multiresolution wavelet decompositions," *Int. J. Syst. Sci.*, vol. 36, pp. 137-152, Feb. 2005.

[56] S. Ferrari, M. Maggioni, and N. A. Borghese, "Multiscale approximation with hierarchical radial basis functions networks," *IEEE Trans. Neural Networks*, vol. 15, pp.178-188, Jan. 2004.