Article:

Billings, S.A. and Wei, H.L. (2005) The wavelet-NARMAX representation : a hybrid model structure combining polynomial models with multiresolution wavelet decompositions. International Journal of Systems Science, 36 (3). pp. 137-152. ISSN 1464-5319

https://doi.org/10.1080/00207720512331338120

Prepublication draft of the paper published in
International Journal of Systems Science, Vol. 36, No. 3, 20 February 2005, 137–152

The Wavelet-NARMAX Representation: A Hybrid Model Structure Combining Polynomial Models with Multiresolution Wavelet Decompositions

S.A. Billings and H.L. Wei
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK

s.billings@shef.ac.uk , w.hualiang@shef.ac.uk

Abstract: A new hybrid model structure combing polynomial models with multiresolution wavelet decompositions is introduced for nonlinear system identification. Polynomial models play an important role in approximation theory, and have been extensively used in linear and nonlinear system identification. Wavelet decompositions, in which the basis functions have the property of localization in both time and frequency, outperform many other approximation schemes and offer a flexible solution for approximating arbitrary functions. Although wavelet representations can approximate even severe nonlinearities in a given signal very well, the advantage of these representations can be lost when wavelets are used to capture linear or low-order nonlinear behaviour in a signal. In order to sufficiently utilise the global property of polynomials and the local property of wavelet representations simultaneously, in this study polynomial models and wavelet decompositions are combined together in a parallel structure to represent nonlinear input-output systems. As a special form of the NARMAX model, this hybrid model structure will be referred to as the WAvelet-NARMAX model, or simply WANARMAX. Generally, such a WANARMAX representation for an input-output system might involve a large number of basis functions and therefore a great number of model terms. Experience reveals that only a small number of these model terms are significant to the system output. A new fast orthogonal least squares algorithm, called the matching pursuit orthogonal least squares (MPOLS) algorithm, is also introduced in this study to determine which terms should be included in the final model.

Keywords: Nonlinear system identification; NARMAX models; wavelets; orthogonal least squares.

1. Introduction

Modelling and identification of nonlinear systems have been extensively studied in recent years, and several model structures and modelling approaches have been developed. These include the polynomial NARMAX (N*onlinear* A*utoRegressive* M*oving* A*verage with* e*X*o*genous* inputs) model (Billings and Leontaritis 1982, Leontaritis and Billings 1985), neural networks (Chen et al. 1990b, Chen and Billings 1992, Billings and Chen 1998, Yamada and Yabuta 1993, Delgado et al. 1995), radial basis function networks (Chen et al 1990a, 1992), wavelet networks (Zhang and Benveniste 1992, Zhang 1997) , fuzzy logic based models (Wang 1992), neuro-fuzzy networks (Brown and Harris 1994), wavelet multiresolution decompositions (Billings and Coca 1999, Coca and Billings 2001), support vector machines and kernel methods(Campbell 2002, Lee and Billings 2002), and other basis function expansion based models. In input-output observational data based modelling, the main task is to determine a suitable model structure, which involves the smallest number of input variables (the lagged inputs and outputs for dynamical systems) and adjustable parameters. In practice, however, model parsimony and accuracy are difficult to achieve simultaneously. Therefore, the trade-offs between model parsimony, accuracy and validity have to be considered. Another property often considered while modelling a dynamical system is the prediction (forecasting) capability of the model.

Among existing model structures, polynomial based model structures play a very important role in linear and nonlinear system modelling and identification. The well-established linear and nonlinear models such as AR(X), ARMA(X) (Ljung 1987) and bilinear models, which have been widely used in linear and nonlinear system modelling, all belong to the polynomial model class and can be viewed as special cases of the polynomial NARMAX model (Billings and Leontaritis 1982, Leontaritis and Billings 1985, Pearson 1995, 1999). Polynomials are globally smooth functions. It has been proved that any given continuous function on an infinite interval can be uniformly approximated using a polynomial (Schumaker 1981). Experience shows that even a simple polynomial model can track the linear trend of a dynamical system very well. However, a polynomial model of a low degree possesses a poor ability to track severe nonlinear behaviour, such as jumps and discontinuities.

Local function expansion based model structures including the wavelet decomposition techniques provide a powerful tool for representing nonlinear signals, even severely nonlinear signals with discontinuities. Among almost all the basis functions used for the purpose of approximation, few have had such an impact and spurred so much interest as *wavelets*. Wavelet decompositions outperform many other approximation schemes and offer a flexible capability for approximating arbitrary functions. Wavelet basis functions have the property of localization in both time and frequency. Due to this inherent property, wavelet approximations provide the foundation for representing arbitrary functions economically, using just a small number of basis functions. Wavelet algorithms (Coca and Billing 2001) process data at different scales or resolutions, and this makes wavelet representations more adaptive compared with other basis functions. Although wavelet decompositions can represent nonlinear signals very well, the advantage of these decompositions might be lost when a signal displays linear or low-order nonlinear trends.

In order to sufficiently utilise the global property of polynomial models and the local property of wavelet representations simultaneously, polynomial models and wavelet decompositions will be combined together in a parallel way to represent a nonlinear input-output system in the present study. As a special form of the NARMAX model, this hybrid model structure will be referred to as the WANARMAX model.

One of the common problems in nonlinear system modelling is the curse of dimensionality. Theoretically, an *n*-dimensional system should be represented using an *n*-variate function. However, for large *n*, it is almost always true that the observational data only forms a sparse distribution in the space R^n. Consequently, the identification problem, which can be converted into a regression problem in most cases and for most model structures, is often ill-posed and various methods have been employed to resolve this problem. One way of representing a continuous function of several variables is to decompose a multivariate function into a superposition of a number of continuous functions with fewer variables and this is the essence of Hilbert's 13th problem, which was resolved by Kolmogorov. Several applicable approaches have been proposed to realize the idea of representing multivariate functions using a superposition of a number of functions with fewer variables. The projection pursuit regression algorithm (Friedman 1981), radial basis function networks (Chen et al 1990b, 1992a), and multi-layer perceptron (MPL) architecture (Haykin 1994) are among the representations that have been studied for multivariate functions. The existing strategies that attempt to approximate general functions in high dimensions are based on suppositions of additive functional submodels including the polynomial

NARMAX representation introduced by Billings and Leontaritis (1982, 1985), the multivariate adaptive regression spline (MARS) method introduced by Friedman (1991), and the adaptive spline modelling of observational data (ASMOD) introduced by Kavli (1993).

Although experience shows that most systems in practice can be expressed as a supposition of a number of low-dimensional submodels if the system variables are appropriately selected, a large number of potential model terms might still be involved when expanding each functional component. Practice and experience show that often many of the model terms are redundant and inclusion of redundant terms can result in a complex model structure and the model may become oversensitive to the training data and is likely to exhibit poor generalisation properties. It is therefore important to determine which terms should be included in the model. A new fast orthogonal least squares algorithm, called the matching pursuit orthogonal least squares (MPOLS) algorithm, is introduced in the present paper as one solution to the model term selection problem.

This paper is organised as follows. In Section 2, the wavelet transform and wavelet decompositions are briefly reviewed. In Section 3, the Wavelet-NARMAX model structure, or simply WANARMAX, is introduced. The model term selection problem is discussed in Section 4, where a new matching pursuit orthogonal least squares (MPOLS) algorithm is proposed. Section 5 discusses the implementation of the WANARMAX model. In section 6, two examples are provided to illustrate the applicability of the new modelling framework. Conclusions are given in Section 7.

2. Multiresolution wavelet decompositions

Assume that the wavelet φ and the corresponding scaling function ϕ constitute an orthogonal wavelet system. From wavelet theory (Mallat 1989, Chui 1992, Daubechies 1992), any function $f \in L^2(R)$ can be expressed as the following *multiresolution wavelet decomposition*

$$f(x) = \sum_k \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \ge j_0}\sum_k \beta_{j,k}\varphi_{j,k}(x) \tag{1}$$

where $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$, $\alpha_{j_0,k}$ and $\beta_{j,k}$ are the wavelet decomposition coefficients, $j,k \in Z$ (Z is a set consisting of whole integers), j_0 is an arbitrary integer representing the coarsest resolution or scaling level. Note that from Eq. (1) and the property of wavelet multiresolution analysis, any function $f \in L^2(R)$ can be arbitrarily closely approximated with some sufficiently large integer *J*, that is, for any ε>0, there exists a sufficiently large integer *J*, such that

$$\left\| f(x) - \sum_k \beta_{J,k}\phi_{J,k}(x) \right\| < \varepsilon \tag{2}$$

Therefore,

$$f(x) \approx \sum_k \beta_{J,k}\phi_{J,k}(x) \tag{3}$$

This means that the multiresolutin wavelet series decomposition (1) can be replaced by wavelet series (3) with respect to the orthogonal scaling functions $\phi_{J,k}(x) = 2^{J/2}\phi(2^J x - k)$, where J is a sufficiently large scale number.

Using the concept of *tensor products*, the multiresolution decomposition (1) 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 scaling and wavelet functions (Mallat 1989). The one-dimensional wavelet decomposition (3) can also be extended to d-dimensional (d >1) case by a tensor product approach as below

$$\begin{aligned} & f_{12\cdots d}(x_1(t), x_2(t), \cdots, x_d(t)) \\ & \qquad = \sum_{k_1\cdots k_d} \alpha_{J;k_1,k_2,\cdots k_d} B_d(2^J x_1(t) - k_1, 2^J x_2(t) - k_2, \cdots, 2^J x_d(t) - k_d) \end{aligned} \tag{4}$$

where $k = [k_1, k_2 \cdots, k_d]^T \in Z^d$ is an d-dimensional index, $B_d(\cdot)$ is an d-dimensional scaling function and can be decomposed as the direct product of d one-dimensional functions

$$B_d(x) = B_d(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} \phi(x_i) \tag{5}$$

where $\phi(\cdot)$ is a scalar scaling function.

3. The WANARMAX model

The WANARMAX model is formed by combining a polynomial model with wavelet decompositions. In this study, polynomial NARMAX models and semi-orthogonal multiresolution wavelet decompositions will be considered and combined in a parallel way.

3.1 The NARMAX representations for nonlinear input-output systems

In the past few decades, modelling and identification techniques for nonlinear systems have been extensively studied with many applications in approximation, prediction and control. Several nonlinear models have been proposed in the literature including the NARMAX model representation which was initially proposed by Billings and Leontaritis (Billings and Leontaritis 1982, Leontaritis and Billings 1985). The NARMAX model takes the form of the following nonlinear difference equation:

$$y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u), e(t-1), \cdots, e(t-n_e)) + e(t) \tag{6}$$

where f is an unknown nonlinear mapping, $u(t)$ and $y(t)$ are the sampled input and output sequences, n_u and n_y are the maximum input and output lags, respectively. The noise variable $e(t)$ with maximum lag n_e, is unobservable but is assumed to be bounded and uncorrelated with the inputs and the past outputs. The model (6) 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 $e(t)$.

One of the popular representations for the NARMAX model (6) is the polynomial representation which takes the function $f(\cdot)$ as a polynomial of degree ℓ and gives the form as

$$
\begin{aligned}
y(t) = \theta_0 + \sum_{i_i=1}^{n} f_{i_1}(x_{i_1}(t)) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}(x_{i_1}(t), x_{i_2}(t)) + \cdots \\
+ \sum_{i_1=1}^{n} \cdots \sum_{i_\ell = i_{\ell-1}}^{n} f_{i_1 i_2 \cdots i_l}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_\ell}(t)) + e(t)
\end{aligned}
\tag{7}
$$

where $\theta_{i_1 i_2 \cdots i_m}$ are parameters, $n = n_y + n_u + n_e$ and

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

$$
x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y + 1 \le k \le n_y + n_u \\ e(t-(k-n_y-n_u)) & n_y + n_u + 1 \le k \le n_y + n_u + n_e \end{cases} \tag{8}
$$

The degree of a multivariate polynomial is defined as the highest order among all terms. For example, the degree of the polynomial $h(x_1, x_2, x_3) = a_1 x_1^4 + a_2 x_2 x_3 + a_3 x_1^2 x_2 x_3^2$ is $\ell =$ 2+1+2=5, which is determined by the last term, $a_3 x_1^2 x_2 x_3^2$. Similarly, a NARMAX model with polynomial degree ℓ means that the order of each term in the model is not higher than ℓ.

The NARX model is a special case of the NARMAX model and takes the form

$$
y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)) + e(t) \tag{9}
$$

In this case the variable $x_k(t)$ defined in (8) reduces to

$$
x_k(t) = \begin{cases} y(t-k), & 1 \le k \le n_y \\ u(t-k+n_y), & n_y + 1 \le k \le n = n_y + n_u \end{cases} \tag{10}
$$

3.2 The wavelet-based ANOVA expansion

Generally, a multivariate nonlinear function can often be decomposed into a superposition of a number of functional components via the well known functional analysis of variance (ANOVA) expansions (Friedman 1991, Chen 1993) as below

$$
\begin{aligned}
y(t) &= f(x_1(t), x_2(t), \cdots, x_n(t)) \\
&= f_0 + \sum_{i=1}^{n} f_i(x_i(t)) + \sum_{1 \le i < j \le n} f_{ij}(x_i(t), x_j(t)) + \sum_{1 \le i < j < k \le n} f_{ijk}(x_i, x_j, x_k) + \cdots \\
&\quad + \sum_{1 \le i_1 < \cdots < i_m \le n} f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)) + \cdots + f_{12\cdots n}(x_1(t), x_2(t), \cdots, x_n(t)) + e(t)
\end{aligned}
\tag{11}
$$

where the first functional component f_0 is a constant to indicate the intrinsic varying trend; f_i, $f_{ij}, \cdots$, are univariate, bivariate, etc., functional components. The univariate functional components $f_i(x_i)$ represent the independent contribution to the system output that arises from the action of the *i*th variable x_i alone; the bivariate functional components $f_{ij}(x_i, x_j)$ represent the interacting contribution to the system output from the input variables x_i and x_j, etc. Let $x_k(t)$ (k=1,2,…,n) be defined as (8) or (10), the ANOVA expansion (11) can then be viewed as a special form of the NARMAX or NARX models for dynamic input and output systems. Although the ANOVA decomposition of the NARMAX model (6) involves up to 2^n different functional components, experience shows that a truncated representation containing the components up to the bivariate or tri-variate functional terms often provides a satisfactory description of $y(t)$ for many high dimensional problems providing that the input variables are properly selected. It is obvious that adopting a truncated ANOVA expansion containing only low-dimensional function components does not mean such an approach will always be appropriate. An exhaustive search for all the possible submodel structures of (11) is demanding and can be prohibitive because of the curse-of-dimensionality. A truncated representation is advantageous and practical if the higher order terms can be ignored. In practice, the constant term f_0 can often be omitted since it can be combined into other functional components.

It will generally be true that, whatever the data set and whatever the modelling approach, the structure of the final model will be unknown in advance. It is therefore not possible to know up to how many order functional components in a truncated ANOVA expansion will be sufficient for a given nonlinear system. This is why model validation methods, which are independent of the model fitting procedure and the model type, are an important part of the NARMAX modelling methodology (Billings and Chen 1998). If the model is adequate to represent the system the residuals should be unpredictable from all linear and nonlinear combinations of past inputs and outputs. This means that the identified model has captured all the predictable information in the data and is therefore the best that can be achieved by any model. It is therefore perfectly acceptable to fit a model that includes just up to one, two or three-dimensional functional terms initially. The model validity tests should then be applied to test if the model that is obtained has captured all the predictable information in the data. If the model fails the model validity tests higher order terms should be included in the initial search set and the procedure should be repeated. It is therefore not necessary to prove that it is always possible to proceed based on just up to certain order submodels. The identification proceeds a stage at a time and uses model validation as the decision making process. This is the NARMAX methodology (Billings and Chen 1998), which is implemented here, and which mimics the traditional approach to analytical modelling. In the latter case the most important model terms are included in the model initially then the less significant terms are added until the model is considered to be adequate. This is exactly what the OLS algorithm and the ERR does but based on the data. The most significant model terms are added first, step by step, a term at a time. The ERR cut-off value is used as a stopping mechanism but the model should never be accepted without applying model validity tests. If these tests fail go back and either reduce the ERR cut-off, or allow more complex model terms in the initial model library, or both and continue until the model validity tests are satisfied.

In practice, many types of functions, such as kernel functions, splines, polynomials and other basis functions can be chosen to express the functional components in model (11). In the present study, however, mutiresolution wavelet decompositions will be chosen to describe the functional components. For example, the functional components $f_p(x_p(t))$ (p=1,2,…,n) and $f_{pq}(x_p(t), x_q(t))$ ($1 \le p < q \le n$) can be expressed using the multiresolution wavelet decompositions as

$$f_p(x_p(t)) = \sum_k \alpha_{j_1,k}^{(p)} \phi_{j_1,k}(x_p(t)) + \sum_{j \geq j_1} \sum_k \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)),\ p = 1,2,\cdots,n, \tag{12}$$

$$\begin{aligned} f_{pq}(x_p(t), x_q(t)) = &\sum_{k_1} \sum_{k_2} \alpha_{j_2;k_1,k_2}^{(pq)(1)} \phi_{j_2,k_1}(x_p(t)) \phi_{j_2,k_2}(x_q(t)) \\ &+ \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(1)} \phi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)) \\ &+ \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(2)} \varphi_{j,k_1}(x_p(t)) \phi_{j,k_2}(x_q(t)) \\ &+ \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(3)} \varphi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)),\ 1 \leq p < q \leq n. \end{aligned} \tag{13}$$

3.3 The WANARMAX model

The wavelet-NARMAX model, or simply WANARMAX, which incorporates a polynomial NARMAX model and a multiresolution wavelet decomposition in a parallel way, can be defined as

$$y(t) = f(x(t)) = f^P(x(t)) + f^W(x(t)) + f^E(\xi(t)) + e(t) \tag{14}$$

where $x(t) = [x_1(t), x_2(t), \cdots, x_n(t)]^T$ and $x_k(t)$ (k=1,2,…,n) are defined as in (10), $f^P(x(t))$ is a polynomial model; $f^W(x(t))$ is a wavelet decomposition model; and $f^E(\xi(t))$ is a polynomial model with respect to the noise variable $e(t)$ and $\xi(t) = [e(t-1), e(t-2), \cdots, e(t-n_e)]^T$ .The submodels $f^P(x(t))$, $f^W(x(t))$ and $f^E(\xi(t))$ can be combined into the WANARMAX model (14) in various forms and the following are some examples

$$f^P(x(t)) = a_0 + \sum_{p=1}^{n} a_p x_p(t) + \sum_{p=1}^{n} \sum_{q=p}^{n} b_{pq} x_p(t) x_q(t) \tag{15}$$

$$f^W(x(t)) = \sum_{p=1}^{n} f_p(x_p(t)) + \sum_{p=1}^{n} \sum_{q=p}^{n} f_{pq}(x_p(t), x_q(t)) \tag{16}$$

$$f^E(\xi(t)) = \sum_{p=1}^{n_e} c_p e(t-p) + \sum_{p=1}^{n_e} \sum_{q=p}^{n_e} c_{pq} e(t-p) e(t-q) \tag{17}$$

where the functional components $f_p(x_p(t))$ (p=1,2,…,n) and $f_{pq}(x_p(t), x_q(t))$ ($1 \leq p < q \leq n$) in (16) can be expressed using the multiresolution wavelet decompositions.

For a selected wavelet $\varphi(\cdot)$ and the scaling function $\phi(\cdot)$, once the maximum lags n_y, n_u and n_e are given, and the initial(coarsest) and highest(finest) resolution scales in the multiresolution decomposition are determined, the WANARMAX model can be rearranged and converted into a linear-in-the-parameters regression model of the form

$$y(t) = \sum_{i=1}^{M_1} \theta_i^P p_i^P(t) + \sum_{j=1}^{M_2} \theta_j^W p_j^W(t) + \sum_{k=1}^{M_3} \theta_k^E p_k^E(t) + e(t) \tag{18}$$

where the regressors $p_i^P(t)$, $p_j^W(t)$ and $p_k^E(t)$ ($i = 1,2,\cdots,M_1; j = 1,2,\cdots,M_2; k = 1,2,\cdots,M_3$) are related to the autoregressive model $f^P(x(t))$, the wavelet decomposition model $f^W(x(t))$ and moving average model $f^{E}(\xi(t))$, respectively. θ_i^P ,θ_j^W and θ_k^E ($i=1,2,\cdots,M_1$; $j=1,2,\cdots,M_2$; $k=1,2,\cdots,M_3$) are parameters to be estimated. $M_1=(n_y+n_u+1)(n_y+n_u+2)/2$, $M_3=n_e$ and M_2 depends on not only the wavelet type used but also the initial and the highest resolution scales.

A special case for the WANARMAX model (18) is the Wavelet-NARX, or simply WANARX model

$$y(t)=\sum_{i=1}^{M_1}\theta_i^P p_i^P(t)+\sum_{j=1}^{M_2}\theta_j^W p_j^W(t)+e(t) \tag{19}$$

Although many functions can be chosen as scaling and/or wavelet functions, most of these are not suitable in system identification applications, especially in the case of multidimensional and multiresolution expansions. An implementation, which has been tested with very good results, involves B-spline and B-wavelet functions in multiresolution wavelet decompositions (Billings and Coca 1999, Coca and Billings 2001, Wei and Billings 2002). B-spline wavelets were originally introduced by Chui and Wang (1992) to define a class of semi-orthogonal wavelets.

For large n_y and n_u , the model (18) might involve a great number of model terms or regressors. Experience shows that often many of the model terms are redundant and therefore are insignificant to the system output and can be removed from the model. An efficient algorithm is required to determine which terms should be included in the model. The significant model term selection problem is discussed in the next section.

4. Model term selection

The selection of which terms should be included in the WANARMAX model (18) is vital if a parsimonious representation of the system is to be identified. For a selected basic wavelet and associated scaling function, once the initial resolution scale level is given, simply increasing the orders n_y and n_u of the dynamic terms and the highest resolutions in the multiresolution wavelet model will in general result in an excessively over parameterised complex model. Fortunately, experience has shown that only a small number of subsets of these model terms are significant and the remainder can be discarded with little deterioration in prediction accuracy. Several possible ways can be used to determine which terms are significant and should be included in the model, including the well-known orthogonal least squares (OLS) algorithm. In this section, the forward orthogonal least squares (OLS) algorithm is briefly summarised and then a new matching pursuit orthogonal least squares (MPOLS) algorithm is introduced.

The WANARMAX model (18) can be expressed as a linear-in-the-parameters equation of the form

$$y(t)=\sum_{m=1}^{M}\theta_m p_m(t)+e(t) \tag{20}$$

where $p_m(t)=p_m^P(t)$ for $m=1,2,\cdots,M_1$, $p_m(t)=p_m^W(t)$ for $M_1+1\le m\le M_1+M_2$, and $p_m(t)=p_m^E(t)$ for $M_1+M_2+1\le m\le M=M_1+M_2+M_3$. θ_m ($m=1,2,\cdots,M$) are parameters to be estimated. Define

$$P^{(m)}=\{p_{i_k}:1\le i_k\le M;\ \ k=1,2,\cdots,m\},\ \ m=1,2,\ldots,M, \tag{21}$$

The model term selection procedure is in fact an iterative process which searches through a nested term set in the sense that

$$P^{(1)} \subset P^{(2)} \subset \cdots \subset P^{(m)} \subset \cdots \tag{22}$$

This makes both the complexity and the accuracy of the representation based on these term sets increase until a suitable term set is found, that is, there exists an integer M_0 (generally $M_0 << M$), such that the model

$$y(t) = \sum_{k=1}^{M_0} \theta_{i_k} p_{i_k}(t) + e(t) \tag{23}$$

provides a satisfactory representation over the range considered for the measured input-output data.

4.1 The forward orthogonal least squares (OLS) algorithm

A fast and efficient model structure determination approach can be implemented using 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 nonlinear models (Billings et al. 1988, 1989, Korenberg et al. 1988, Chen et al. 1989). This approach has been extensively studied and widely applied in nonlinear system identification (see, for example, Chen et al. 1991, Wang and Mendel 1992, Zhu and Billings 1996, Zhang 1997, Hong and Harris 2001). The forward OLS algorithm involves a stepwise orthogonalization of the regressors and a forward selection of the relevant terms in (20) based on the error reduction ratio (ERR) (Billings et al. 1988, 1989). The procedure can be briefly summarised as follows:

Consider the linear-in-the-parameters model (20), where the regression matrix $P = [p_1, p_2, \cdots, p_M]$ with $p_i = [p_i(1), p_i(2), \cdots, p_i(N)]^T$, N is the length of the observational data set. With the assumption that P is full rank in columns, then P can be orthogonally decomposed as

$$P = WA \tag{24}$$

where A is an $M \times M$ unit upper triangular matrix and W is an $N \times M$ matrix with orthogonal columns $w_1, w_2, \cdots, w_M$ in the sense that $W^T W = D = diag[d_1, d_2, \cdots, d_M]$ with $d_m = w_m^T w_m$. Model (20) can then be expressed as

$$Y = (PA^{-1})(A\Theta) + \Xi = WG + \Xi \tag{25}$$

where $Y = [y(1), y(2), \cdots, y(N)]^T$ are the observations of the system output, $\Theta = [\theta_1, \theta_2, \cdots, \theta_M]^T$ is the parameter vector, $\Xi = [\varepsilon(1), \varepsilon(2), \cdots, \varepsilon(N)]^T$ is the vector of the noise signal, and $G = [g_1, g_2, \cdots, g_M]^T$ is an auxiliary parameter vector, which can be calculated directly from Y and W by means of the property of orthogonality as

$$g_i = \frac{Y^T w_i}{w_i^T w_i}, \qquad i = 1, 2, \cdots, M \tag{26}$$

The parameter vector Θ , which is related to G by the equation $A\Theta = G$, can easily be calculated by solving this equation using a substitution scheme.

The number M of all the candidate terms in model (20) is often very large. Some of these terms may be redundant and should be removed to give a parsimonious model with only M_0 terms ($M_0 << M$). Detection of the significant model terms can be achieved using the OLS procedures described below.

Assume that the residual signal $\varepsilon(t)$ in the model (20) is uncorrelated with the past outputs of the system, then the output variance can be expressed as

$$\frac{1}{N}Y^T Y = \frac{1}{N}\sum_{i=1}^{M} g_i^2 w_i^T w_i + \frac{1}{N}\Xi^T \Xi \tag{27}$$

Note that the output variance consists of two parts, the desired output $(1/N)\sum_{i=1}^{M} g_i^2 w_i^T w_i$ which can be explained by the regressors, and the part $(1/N)\Xi^T \Xi$ which represents the unexplained variance. Thus $(1/N)\sum_{i=1}^{M} g_i^2 w_i^T w_i$ is the increment to the explained desired output variance brought by p_i, and the i th error reduction ratio, ERR_i, introduced by p_i, can be defined as

$$ERR_i = \frac{g_i^2 (w_i^T w_i)}{Y^T Y} \times 100\% = \frac{(Y^T w_i)^2}{(Y^T Y)(w_i^T w_i)} \times 100\%, \quad i = 1, 2, \cdots, M, \tag{28}$$

This ratio provides a simple but effective means for seeking a subset of significant regressors. The significant terms can be selected in a forward-regression manner according to the value of ERR_i step by step. The significant terms can be selected in a forward-regression manner according to the value of ERR_i. Several orthogonalization procedures, such as Gram-Schmidt, modified Gram-Schmidt and Householder transformation (Chen et al. 1989) can be applied to implement the orthogonal decomposition. The improved version of this algorithm (Zhu and Billings 1996) provides a significant reduction in the computations and is advantageous compared to standard Gram-Schmidt algorithm when dealing with high order MIMO systems. Other recent studies by Hong and Harris (2001) have proposed other improvements to this procedure.

***Remark* 1:** The forward orthogonal least squares algorithm for model term selection is described and expounded in a matrix form here for convenience of introducing and explaining the concept of error reduction ratio (ERR). In practical identification, however, this algorithm is often implemented in a forward stepwise way (Wei and Billings 2004). The most significant model terms are added first, step by step, a term at a time. The ERR cut-off value is used as a stopping mechanism but the model should never be accepted without applying model validity tests.

***Remark* 2:** The candidate terms that are not chosen in the first step are orthogonalized with respect to all previously selected basis functions. Because of the orthogonality the j th term can be selected in the same way as in the first step. w_j is the j th selected orthogonal term and g_j is the corresponding parameter. Any numerical ill conditioning can be avoided by eliminating the candidate basis functions for which $w_i^T w_i$ are less than a predetermined threshold τ, for example, $\tau = 10^{-r}$ and $r \geq 10$.

***Remark* 3:** The assumption that the regression matrix *P* is full rank in columns is unnecessary in the iterative forward OLS algorithm(Wei and Billings 2004). In fact, if the *M* columns of the matrix *P* are linearly dependent, and assuming that the rank in columns of the matrix *P* is *L* (<*M*) , then the algorithm will stop at the *L* -th step.

***Remark 4*:** If required, the procedure can be terminated at the M_0 -th step ($M_0 \leq L$) when $1-\sum_{i=1}^{M_0} ERR_i < \rho$, where ρ is a desired error tolerance called the *cutoff*, which can be learnt during the regression procedure. The final model is the linear combination of the M_0 significant terms selected from the M candidate terms $\{p_i\}_{i=1}^{M}$

$$y(t)=\sum_{i=1}^{M_0} g_i w_i(t)+e(t) \tag{29}$$

which is equivalent to

$$y(t)=\sum_{i=1}^{M_0} \theta_{\ell i} p_{\ell i}(x(t))+e(t) \tag{30}$$

where the parameters $\Theta^{(OLS)}=[\theta_{\ell_1},\theta_{\ell_2},\cdots,\theta_{\ell_{M_0}}]^T$ are calculated from the triangular equation $AG^{(OLS)}=\Theta^{(OLS)}$ with $G^{(OLS)}=[g_1,g_2,\cdots,g_{M_0}]^T$ and

$$A=\begin{bmatrix} 1 & a_{12} & \cdots & a_{1M_0} \\ 0 & 1 & \cdots & a_{2M_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{M_0-1,M_0} \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{31}$$

The entries $a_{ij}\,(1 \leq i < j \leq M_0)$ are given in the above OLS algorithm.

***Remark* 5:** The key point in the forward OLS-ERR algorithm is focused on detecting the most significant model terms from a great number of candidates by introducing an orthogonalization procedure and the concept of error reduction ratio (ERR). The estimation of the model parameters is only a by-product of the model term selection procedure. It requires that the orthogonalization should be explicit so that the significant model terms selected by the algorithm are transparent to the model builders. Many other singular value decomposition methods including the standard SVD, Krylov subspace and Lanczos bidiagonalizaiton methods, which are proved to be more numerically stable compared with the present OLS algorithm, can be used to solve linear least squares problems, where the main task is to estimate the unknown parameters for a given linear equation. These methods, however, cannot provide any information about which model terms are the most significant. In other words, for a given linear-in-the-parameters form (20), these methods cannot tell which model terms or regressors make the most significant contributions to the system output *y*(*t*). These methods could however provide an aid for solving the identification problem by combining the methods with the forward OLS-ERR algorithm. For example, the most significant model terms could then be selected initially using the forward OLS-ERR algorithm, and finally the unknown parameters could be estimated using the more numerically stable and precise methods.

4.2 Matching pursuit orthogonal least squares (MPOLS) algorithm

Note that in the forward OLS algorithm, at each step all the unselected regressors are made to orthogonalize with the previously selected regressors, and most of the computational cost is based on these orthogonalization transforms. An iterated orthogonal projection algorithm, the matching pursuit method, proposed by Mallat and Zhang (1993) is a simple regressor selection algorithm which is relatively computationally efficient. But the matching pursuit algorithm is less efficient than OLS, since the number of regressors selected by the matching pursuit algorithm is almost always larger than that selected by OLS for the same given threshold value of approximation accuracy. A trade-off between the efficiency and the computational cost is considered here by combining the advantages of the forward OLS with the matching pursuit algorithm to create a new algorithm called the matching pursuit orthogonal least squares (MPOLS) algorithm. The algorithm is described below.

For the output vector $Y=[y(1),y(2),\cdots,y(N)]^T$ in (20), find a vector p_{ℓ_1} from the candidate regressor family $\{p_1,p_2,\cdots,p_M\}$, so that p_{ℓ_1} is the "best" matching regressor to *Y*, i.e., p_{ℓ_1} makes the mean squared error of the following linear regression

$$y(t)=c_m p_m(t)+\xi_m(t) \tag{32}$$

achieve a minimum in the sense that

$$\frac{1}{N}\sum_{t=1}^{N}\xi_{\ell_1}^2(t)=\frac{1}{N}\sum_{t=1}^{N}(y(t)-c_{\ell_1}p_{\ell_1}(t))^2=\min_m\left\{\frac{1}{N}\sum_{t=1}^{N}[y(t)-c_m p_m(t)]^2\right\} \tag{33}$$

The "best" matching regressor p_{ℓ_1} can be found using orthogonal projection approach by defining

$$\cos\alpha=\frac{Y^T p_m}{\sqrt{Y^TY}\sqrt{p_m^T p_m}} \tag{34}$$

$$\|p_m^1\|=\|Y\|\cos\alpha=\frac{Y^T p_m}{\sqrt{p_m^T p_m}} \tag{35}$$

Such that

$$\sum_{t=1}^{N}\xi_m^2(t)=\|\xi_m\|^2=\|Y\|^2-\|p_m^1\|^2=Y^TY-\frac{(Y^T p_m)^2}{p_m^T p_m} \tag{36}$$

Thus

$$\ell_1=\arg\max_m\left\{\frac{(Y^T p_m)^2}{p_m^T p_m},1\le m\le M\right\} \tag{37}$$

Set $q_1(t)=p_{\ell_1}(t)$, $w_1(t)=q_1(t)$, $g_1=(Y^T w_1)/(w_1^T w_1)$, $ERR_1=g_1^2(w_1^T w_1)/(Y^TY)$, and $\eta_1(t)=y(t)-g_1 w_1(t)$.

At the second step, find a vector p_{ℓ_2} from the candidate regressor family $\{p_m:1\le m\le M, m\ne\ell_1\}$, so that p_{ℓ_2} is the "best" matching regrssor to η_1. Following the approach in (32) and (33), ℓ_2 should be chosen as

$$\ell_2=\arg\max_m\left\{\frac{(\eta_1^T p_m)^2}{p_m^T p_m},1\le m\le M, m\ne\ell_1\right\} \tag{38}$$

Set $q_2(t) = p_{\ell_2}(t)$. Orthogonalize q_2 with w_1 as below

$$w_2 = q_2 - \frac{w_1^T q_2}{w_1^T w_1} w_1 \tag{39}$$

And set $g_2 = (Y^T w_2)/(w_2^T w_2)$, $ERR_2 = g_2^2 (w_2^T w_2)/(Y^T Y)$, and $\eta_2(t) = \eta_1(t) - g_2 w_2(t)$.

Generally, at step *k*, select

$$\ell_k = \arg\max_m \left\{ \frac{(\eta_{k-1}^T p_m)^2}{p_m^T p_m}, 1 \le m \le M, m \ne \ell_1, m \ne \ell_2, \cdots, m \ne \ell_{k-1} \right\} \tag{40}$$

Set $q_k(t) = p_{\ell_k}(t)$ and orthogonalize q_k with $w_1, w_2, \cdots, w_{k-1}$ as below

$$w_k = q_k - \frac{w_1^T q_k}{w_1^T w_1} w_1 - \frac{w_2^T q_k}{w_2^T w_2} w_2 - \cdots - \frac{w_{k-1}^T q_k}{w_{k-1}^T w_{k-1}} w_{k-1} \tag{41}$$

Calculate $g_k = (Y^T w_k)/(w_k^T w_k)$, $ERR_k = g_k^2 (w_k^T w_k)/(Y^T Y)$, and set $\eta_k(t) = \eta_{k-1}(t) - g_k w_k(t)$.

A similar algorithm has been used for basis selection in wavelet neural networks (Xu 2002). Note that in the MPOLS algorithm, only the most recently selected regressor $q_j = p_{\ell_j}$ at step *j* is made to be orthogonal with the previous selected regressors $q_k = p_{\ell_k}$ (*k*=1,2,…,*j*-1). Therefore, the computational load of the orthogonalization procedure in OLS, which involves making all the unselected regressors orthogonal with the previously selected regressors, is significantly reduced in the new MPOLS algorithm. Therefore, the computational cost of the MPOLS algorithm is much less than that of the OLS algorithm, and the new algorithm is much faster then most existing OLS and fast OLS algorithms.

In the MPOLS algorithm, any numerical ill conditioning can be avoided by eliminating the candidate terms for which $p_i^T p_i$ is less than a predetermined threshold τ , for example, $\tau = 10^{-r}$ and $r \ge 10$. w_j is the j th selected orthogonal term and g_j is the corresponding parameter. If required, the procedure can be terminated at the M_0-th step ($M_0 \le L$) when $1 - \sum_{i=1}^{M_0} ERR_i < \rho$, where ρ is a desired error tolerance, which can be learnt during the regression procedure. The final model is the linear combination of all the selected significant terms in the form of (29) and (30).

Notice that, for the same problem, MPOLS may select different model terms (regressors) and different numbers of model terms compared with OLS even for the same threshold value of termination. It is nearly always true that the MPOLS selects more model terms than that of OLS. However, the first term selected by both algorithms is always the same. The computational efficiency of the MPOLS algorithm compared with OLS can be demonstrated using the CPU time required to perform a bench test example on the same computer. This is illustrated in Table 1.

Table 1 The comparison of the computational efficiency between OLS and MPOLS

Cases	Data length (N)	Number of candidate regressors (M)	Number of selected regressors (m)		CPU time (sec)	
			OLS	MPOLS	OLS	MPOLS
Case 1	500	565	12	20	23.23	2.03
Case 2	600	1321	9	15	119.73	10.29
Case 3	1000	705	21	44	226.38	22.15
Case 4	500	1153	110	112	1503.82	21.49
Note: The threshold values to terminate the OLS and MPOLS algorithms were the same.						

5. Implementing a WANARMAX Model

This section summarizes the procedure for implementing a WANARMAX model. The implementation of a WANARMAX model involves several practical issues including observational input-output data pre-processing, significant variable selection (Wei and Billings 2004), resolution level determination in the wavelet decomposition submodels, and model validity tests (Billings and Voon, 1986; Billings and Zhu, 1995).

The iterative identification procedure to implement a WANARMAX model consists of the following steps.

***Step* 1:** *Data pre-processing*

For convenience of implementation, convert the original observational input-output data $u(t)$ and $y(t)$ (t=1,2, …,N) into unit intervals $[0,1]$. The converted input and output are still denoted by $u(t)$ and $y(t)$.

***Step* 2:** *Determining the model initial conditions*

This includes:

(i) Provide values for n_y, n_u, n_e, ρ and ρ_e (where ρ and ρ_e are threshold values for terminating the model term selection procedure, ρ is used in Step 3 and ρ_e in Step 4, notice in general $\rho_e < \rho$).

(ii) Set $e(t)$ =0 for t=1,2,…,N.

(iii) If possible, select the significant variables from all the candidate lagged output and input variables $\{y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u)\}$. This involves the model order determination and variable selection problems.

(iv) Select a polynomial submodel $f^P(x(t))$, a wavelet submodel $f^W(x(t))$, and a noise model $f^E(\xi(t))$ from the representations (15)-(17).

(v) Determine the initial and the highest resolution scales. Generally the initial resolution scales j_1 and j_2 in the wavelet models can be set to $j_1 = j_2$=0, and the highest resolution scales J_1 and J_2 can be chosen in a heuristic way.

***Step* 3:** *Identify the WANARX model*

(i) Calculate the regressors $p_i^P(t)$ and $p_j^W(t)$ ($i=1,2,\cdots,M_1; j=1,2,\cdots,M_2$) which are related to the

the autoregressive models $f^{P}(x(t))$ and the wavelet decomposition model $f^{W}(x(t))$. The regression matrix $P=[P^{P},P^{W}]$ of the WANARX model (19) are formed from these regreesors.

(*ii*) Select the significant terms in the autoregressive models $f^{P}(x(t))$ and the wavelet decomposition model $f^{W}(x(t))$ using the OLS or MPOLS algorithms to obtain parsimonious models of the form (29) and (30).

***Step* 4:** *An iterative loop to identify a WANARMAX model*

(*i*) Set *k*=0 and estimate the initial residuals

$$\begin{aligned}\varepsilon^{(0)}(t) &= y(t)-\hat{y}(t)\\ &= y(t)-\hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),0,\cdots,0)\\ &= y(t)-\sum_{i=1}^{M_0} g_i^{(k)} w_i^{(k)}(t)\end{aligned} \tag{42}$$

where $g_i^{(0)}=g_i$ and $w_i^{(0)}=w_i\,(i=1,2,\cdots,M_0)$ are the orthogonalized regressors and the parameters estimated in Step 3 (*ii*).

(*ii*) Set *k*=:*k*+1. Select significant terms for the moving average model $f^{E}(\xi(t))$, add these terms to the model estimated in Step 3 (*ii*). Re-estimate the parameters for the updated model using the OLS or MPOLS algorithms, and calculate the residuals $\varepsilon^{(k)}(t)$ recursively using

$$\begin{aligned}\varepsilon^{(k)}(t) &= y(t)-\hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),\varepsilon^{(k-1)}(t-1),\cdots,\varepsilon^{(k-1)}(t-n_e))\\ &= y(t)-\sum_{j=1}^{M_0+m_e}\theta_{\ell_j}^{(k)} p_{\ell_j}(t)\end{aligned} \tag{43}$$

or

$$\varepsilon^{(k)}(t)=y(t)-\sum_{j=1}^{M_0+m_e} g_j^{(k)} w_j^{(k)}(t) \tag{44}$$

where m_e is the number of the noise terms selected. The above recursive calculation will be terminated at the *k*th iteration if one of the following the convergence tests is satisfied

$$\sum_{m=1}^{M_0+m_e}\frac{\left|g_m^{(k)}-g_m^{(k-1)}\right|}{\left|g_m^{(k)}\right|}\le\delta_1 \tag{45}$$

and

$$\sum_{t=1}^{N}\left|\varepsilon^{(k)}(t)-\varepsilon^{(k-1)}(t)\right|^2\le\delta_2 \tag{46}$$

where δ_1 and δ_2 are two tolerance values for convergence testing. Numerous tests have shown that less than 10 iterations, typically 3-5 iterations, are sufficient for the algorithm to converge.

***Step* 5:** *Model validity tests*

Apply model validity tests to evaluate the identified model. If the identified model does not satisfy the model validity tests, change some of the initial model conditions in Step 2, especially conditions in (*i*), (*iv*)and (*v*), and repeat Steps 3 to 4.

6. Examples

Two examples, one a simulated system and one based on real data relating to a terrestrial magnetosphere dynamic system, are given to illustrate the effectiveness and applicability of the new modelling framework. The original observational input-output data $u(t)$ and $y(t)$ (t=1,2, …,N) are normalized into the unit interval [0,1] for the convenience of implementation. The modelling can then be performed in [0,1], and the model output can then be recovered to the original system operating domain by taking the inverse transform.

6.1 Simulated example—a nonlinear system

The following nonlinear input-output system

$$y(t)=\frac{y(t-1)y(t-2)+y(t-1)y(t-3)+y(t-2)y(t-3)}{1+y^2(t-1)+y^2(t-2)+y^2(t-3)}$$
$$+2[\sin(y(t-1))][\cos(y(t-2))]+2[\sin(y(t-2))][\cos(y(t-3))]$$
$$+2[\sin(y(t-3))][\cos(y(t-1))]+6u^2(t-1)+u^3(t-2) \tag{47}$$

was simulated using a system input with the form

$$u(t)=2\sin(\pi t/25)+0.5\sin(\pi t/30)+0.02\exp[\sin(\pi t/40)] \tag{48}$$

The estimation set consists of 500 input-output data points which are shown in Figure 1. It was assumed that the real model structure is unknown and setting n_y and n_u to be 5 and 3, respectively, in the initial model, which was assumed to be of the form

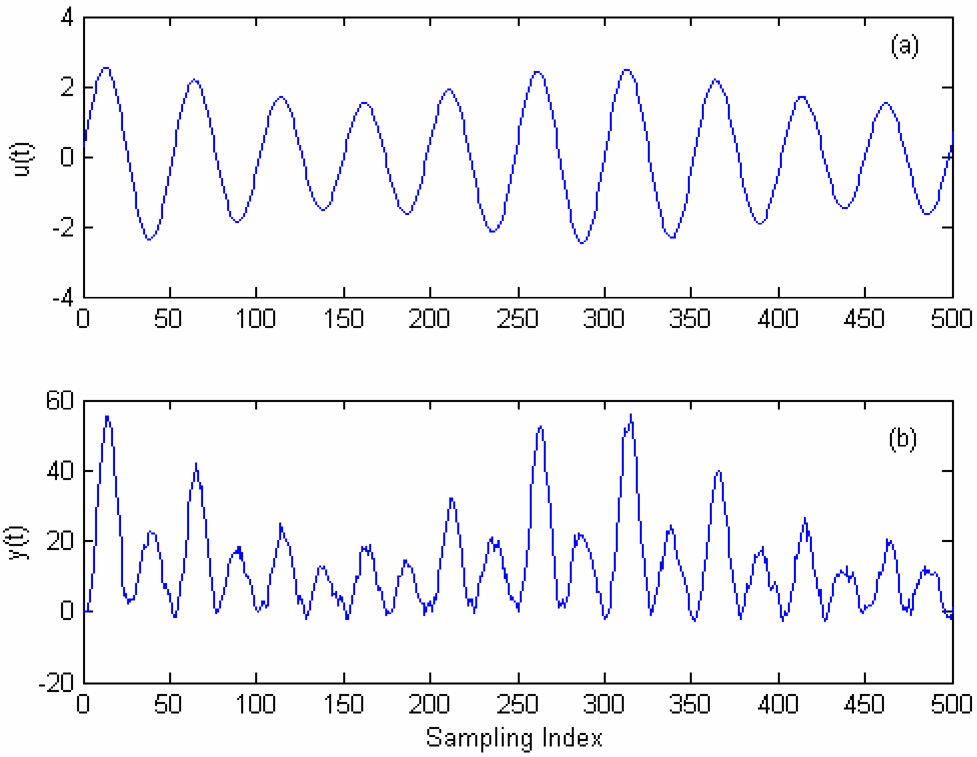

Figure 1 The input and output data of the system described by Eq. (47)

$$
\begin{aligned}
y(t) = {} & f(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-3)) \\
& = a_0 + \sum_{p=1}^{5} a_p y(t-p) + \sum_{p=1}^{3} b_p u(t-p) + \sum_{p=1}^{5}\sum_{q=p}^{5} b_{pq} y(t-p)y(t-q) \\
& \quad + \sum_{p=1}^{3}\sum_{q=p}^{3} c_{pq} u(t-p)u(t-q) + \sum_{p=1}^{5}\sum_{q=1}^{3} d_{pq} y(t-p)u(t-q) \\
& \quad + \sum_{p=1}^{5} f_p(y(t-p)) + \sum_{p=1}^{3} f_{p+5}(u(t-p))
\end{aligned} \tag{49}
$$

where each function $f_p(\cdot)$ can be described using the multiresolution wavelet decomposition (12) as

$$
f_p(x_p(t)) = \sum_{k\in K^0} \alpha_{0,k}^{(p)} \phi_{0,k}(x_p(t)) + \sum_{j=0}^{4}\sum_{k\in K_j} \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)),\ p = 1,2,\cdots,8, \tag{50}
$$

where $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$ and $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ are the 4th order B-spline wavelet and scaling functions, and the finest resolution level was chosen to be *J*=4.. From the definition of the B-spline wavelets (Chui 1992, Chui and Wang 1992), the sets K^0 and K_j can easily be determined as $K^0 = \{-3,-2,-1,0\}$ and $K_j = \{-6,-5,\cdots,-1,\ 0,1,\cdots,2^j-1\}$.

The initial model (49) contains 565 model regressors, but most of these are likely to be redundant and should be removed from the initial model. Both the OLS and MPOLS algorithms were used to select the significant regressors, and two validated parsimonious models were obtained

$$
y(t) = \hat{f}^{(OLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-3)) = \sum_{k=1}^{12} \theta_k^{(OLS)} p_k^{(OLS)}(t) \tag{51}
$$

$$
y(t) = \hat{f}^{(MPOLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-3)) = \sum_{k=1}^{20} \theta_k^{(MPOLS)} p_k^{(MPOLS)}(t) \tag{52}
$$

The parameters, regressors and the corresponding error reduction ratios (ERR) of the models (51) and (52) are listed in Table 2 and Table 3, respectively. A comparison of the model predicted outputs and the measurements, are shown in Figure 2. Note that more model terms has been selected by the MPOLS algorithm than that selected by the forward OLS algorithm, but the model predicted outputs of the MPOLS identified model (52) is worse than that from the OLS identified model (51), this behaviour will be investigated in a later paper. The model predicted output (MPO) is defined as

$$
\hat{y}_{mpo}(t) = \hat{f}(\hat{y}_{mpo}(t-1),\cdots,\hat{y}_{mpo}(t-n_y),u(t-1),\cdots,u(t-n_u),0,\cdots,0) \tag{53}
$$

The model predicted outputs are recursively estimated and are used to calculate the model prediction errors

$$
\hat{e}_{mpo}(t) = y(t) - \hat{y}_{mpo}(t) \tag{54}
$$

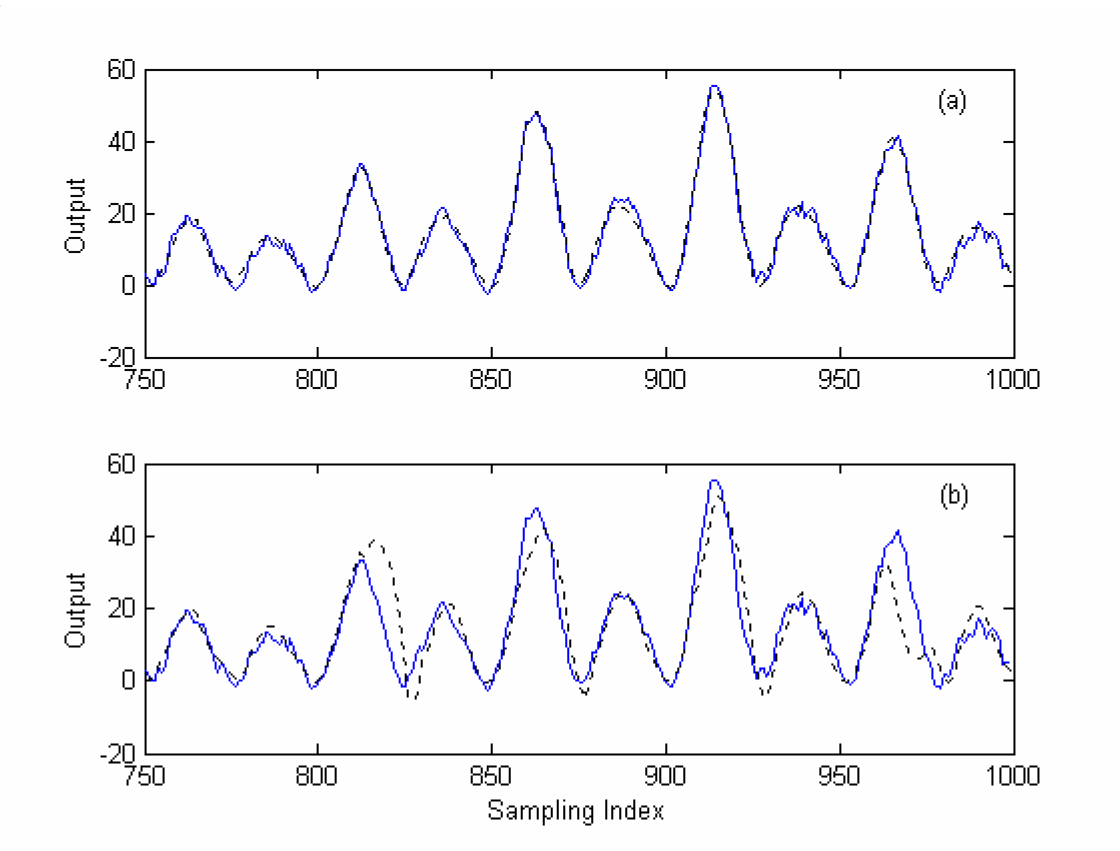

Figure 2 The comparison of the model predicted output (MPO) and the measurements for the system described by Eq (47). (a) The model predicted outputs based on the model (51) ; (b) The model predicted outputs based on the model (52). (The solid line denotes the measurements, and the dashed line denotes the model predicted outputs.)

Table 2 The regressors, parameters and the corresponding ERRs estimated using OLS for the system described by Eq (47)

Number k	Terms $p_k^{(OLS)}(t)$	Parameters $\theta_k^{(OLS)}$	$ERR_k \times 100\%$
1	$y(t-1)$	5.02655e-001	97.52096
2	$y(t-4)$	-9.37588e-002	1.04316
3	$\phi_{0,-1}(u(t-3))$	-6.55070e-001	0.23092
4	$\varphi_{0,-2}(u(t-1))$	7.21870e-001	0.10046
5	$\varphi_{1,-3}(y(t-1))$	7.63680e-002	0.22474
6	$\varphi_{0,-3}(y(t-1))$	1.90501e-002	0.08508
7	$\varphi_{0,0}(u(t-1))$	-2.23549e+001	0.11981
8	$\varphi_{0,-3}(y(t-1))$	-5.04206e-001	0.02497
9	$\varphi_{4,2}(y(t-5))$	3.73955e-003	0.01516
10	$\phi_{0,-1}(u(t-2))$	1.41307e+000	0.01250
11	$y(t-2)u(t-2)$	-2.49814e+000	0.01455
12	$y(t-2)u(t-3)$	2.10633e+000	0.03581
Note: The CPU time spent on selecting these model terms from all the candidate model term set is 23.23*s*.			

Table 3 The regressors, parameters and the corresponding ERRs estimated using MPOLS for the system described by Eq (47)

Number k	Terms $p_k^{(MPOLS)}(t)$	Parameters $\theta_k^{(MPOLS)}$	$ERR_k \times 100\%$
1	$y(t-1)$	1.01732e+000	97.52096
2	$\varphi_{1,-3}(y(t-5))$	1.67365e-001	0.51440
3	$\phi_{0,0}(y(t-5))$	-9.08939e-001	0.51530
4	$\varphi_{0,2}(u(t-1))$	2.26668e-001	0.20425
5	$\varphi_{0,-4}(y(t-1))$	-5.24924e+000	0.11191
6	$\varphi_{1,-1}(u(t-3))$	-8.15303e-002	0.08418
7	$\varphi_{0,0}(y(t-4))$	-1.40831e+000	0.04319
8	$\varphi_{1,-1}(y(t-1))$	-4.91165e-002	0.02270
9	$\varphi_{1,-2}(y(t-5))$	-4.16277e-002	0.03402
10	$\varphi_{3,6}(u(t-1))$	4.07545e-002	0.02683
11	$\varphi_{2,-2}(y(t-1))$	-7.13154e-003	0.02574
12	$\varphi_{2,-4}(y(t-4))$	2.73731e-002	0.02045
13	$\varphi_{4,10}(y(t-4))$	-1.10107e-002	0.01004
14	$\varphi_{2,1}(u(t-1))$	-1.60958e-002	0.01619
15	$\varphi_{0,-3}(y(t-5))$	3.44345e-003	0.00903
16	$\varphi_{4,7}(y(t-2))$	8.70263e-003	0.01084
17	$\varphi_{1,-1}(y(t-4))$	-1.46078e-002	0.00858
18	$\varphi_{2,3}(y(t-3))$	-1.17200e+000	0.00893
19	$\varphi_{4,5}(y(t-1))$	4.28377e-003	0.00737
20	$\varphi_{4,12}(y(t-2))$	-9.62771e-003	0.00821
Note: The CPU time spent on selecting these model terms from all the candidate model term set is 2.03*s*.			

6.2 A terrestrial magnetosphere dynamical system

While the results obtained for the simulated system in section 6.1 demonstrate the applicability of the wavelet-NARMAX model, it does not provide a realistic test for the new hybrid modelling structure. To achieve the latter objective, a data set related to a terrestrial magnetosphere dynamic system was considered.

The sun is a source of a continuous flow of charged particles, ions and electrons called the solar wind. The terrestial 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 D_{st} index is used to measure the disturbance of the geomagnetic field in the magnetic storm. Numerous studies of correlations between the solar wind parameters and magnetospheric disturbances show that the product of the solar wind velocity V and the southward component of the magnetic field, quantified by B_s, represents the input that can be considered as the input to the magnetosphere. Denote the multiplied input by VB_s.

Figure 3 shows 1000 data points of measurements of the solar wind parameter VB_s (input) and the D_{st} index (output) with a sample period T=1hour. The purpose here is to identify a nonlinear model to represent the input-output relationship between VB_s (input) and D_{st}. The effects of other inputs on the system will be neglected in the present study.

The objective here was to construct a hybrid wavelet-NARMAX model of the form (14). The first 500 input-output data points were used for model identification and the remaining 500 data points were used for testing. Ten significant variables {$y(t$-1), …, $y(t$-5),$u(t$-1), …, $u(t$-5)} were initially selected using a variable selection algorithm. The initial model was chosen as below:

$$\begin{aligned} y(t) &= f(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),\cdots,e(t-10)) \\ &= a_0 + \sum_{p=1}^{10} a_p x_p(t) + \sum_{p=1}^{10}\sum_{q=p}^{10} b_{pq} x_p(t) x_q(t) + \sum_{p=1}^{10} f_p(x_p(t)) \\ &\quad + \sum_{p=1}^{10} c_p e(t-p) + e(t) \end{aligned} \tag{55}$$

where $x_p(t) = y(t-p)$ for p=1,..,5 and $x_p(t) = u(t-p+5)$ for p=6,…10, and each function $f_p(\cdot)$ can be expressed as Eq. (50) .

The implementation procedure 5.2 was performed step by step, and both the OLS and MPOLS algorithms were used in the model identification procedure, finally two validated parsimonious models were obtained

$$\begin{aligned} y(t) &= \hat{f}^{(OLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),...,e(t-10)) \\ &= \sum_{k=1}^{14} \theta_k^{(OLS)} p_k^{(OLS)}(t) \end{aligned} \tag{56}$$

$$\begin{aligned} y(t) &= \hat{f}^{(MPOLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),\cdots,e(t-10)) \\ &= \sum_{k=1}^{16} \theta_k^{(MPOLS)} p_k^{(MPOLS)}(t) \end{aligned} \tag{57}$$

The parameters, regressors and the corresponding error reduction ratios (ERR) of the selected regressors in models (56) and (57) are listed in Table 4 and Table 5, respectively. A comparison of the model predicted outputs and the measurements are shown in Figure 4, which clearly indicates that the model predicted outputs provide good long term predictions and give confidence in the identified model. The discrepancy between the model predicted outputs and the measured values of the $\boldsymbol{D_{st}}$ index are believed to be the result of other inputs which affect the system output but were not included in the current model.

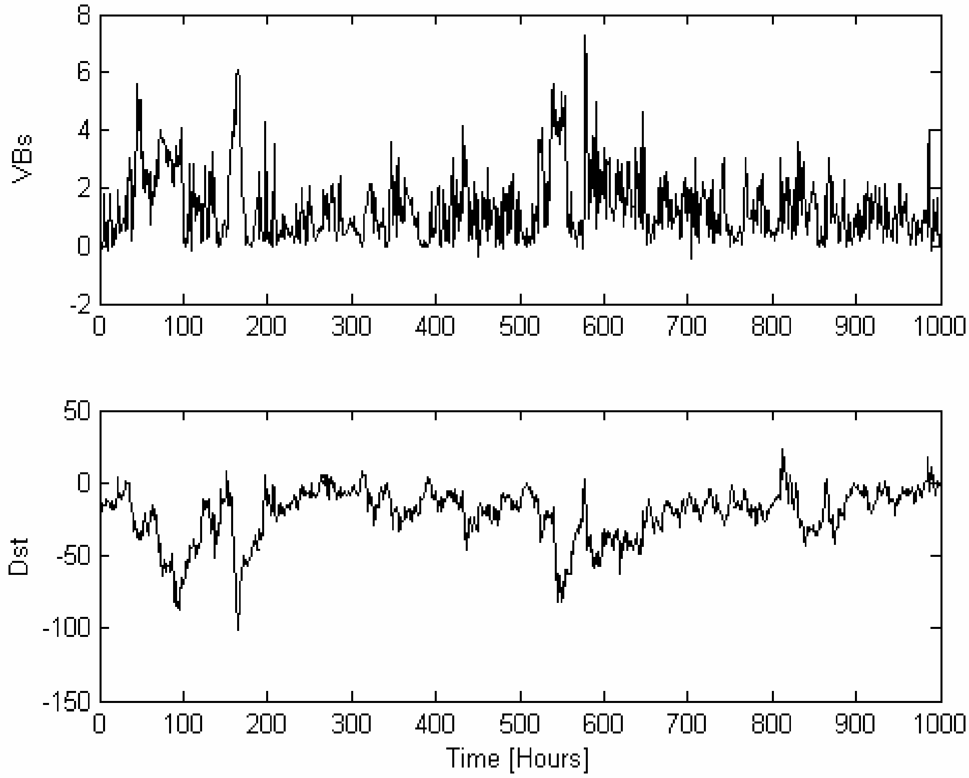

Figure 3   The input (VBs) and output (Dst) data of a terrestrial magnetospheric dynamic system.

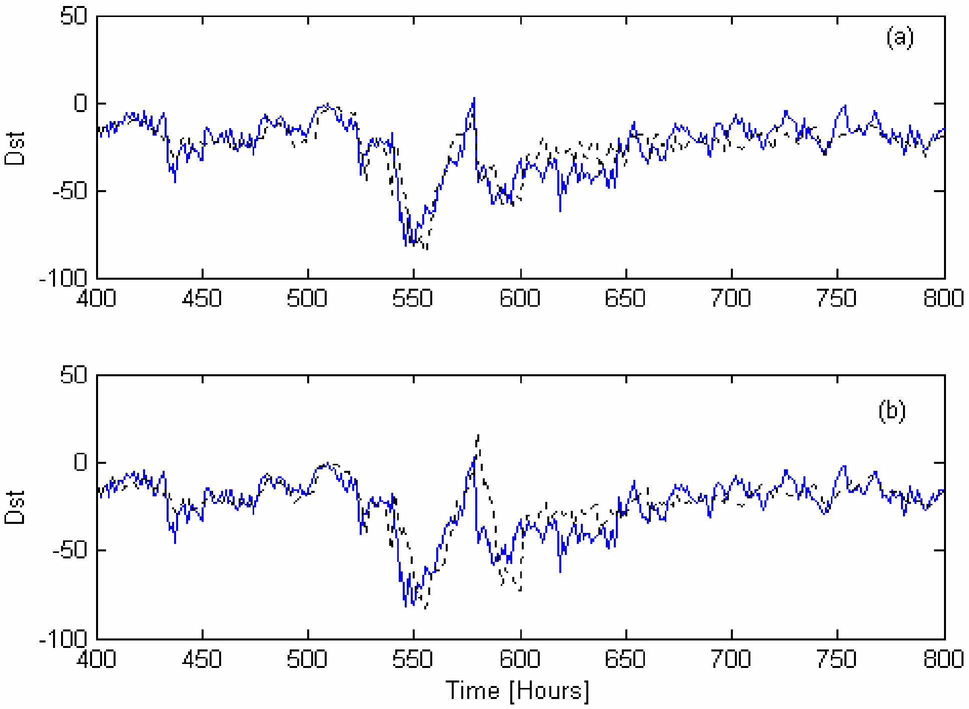

Figure 4   The comparison of the model predicted output (MPO) and the measurements for a terrestrial magnetospheric dynamic system.(a) The model predicted outputs based on the model (56) ; (b) The model predicted outputs based on the model (57).(The solid line denotes the measurements, and the dashed line denotes the model predicted outputs.)

Table 4 The regressors, parameters and ERRs estimated using OLS for a terrestrial magnetospheric dynamic system.

Number k	Terms $p_k^{(OLS)}(t)$	Parameters $\theta_k^{(OLS)}$	$ERR_k \times 100\%$
1	$y(t-1)$	8.86991e-001	95.64488
2	$\phi_{0,-3}(u(t-1))$	7.28895e-001	1.53870
3	$\varphi_{0,-4}(u(t-1))$	2.92761e+000	1.01020
4	$\varphi_{2,1}(u(t-2))$	8.09016e-002	0.71025
5	$\varphi_{2,-1}(y(t-2))$	1.22450e-002	0.70824
6	$\varphi_{4,3}(y(t-1))$	1.04799e-002	0.09612
7	$\varphi_{3,1}(y(t-2))$	9.99869e-003	0.00544
8	$\varphi_{3,2}(y(t-2))$	-5.38155e-003	0.00525
9	$e(t-1)$	1.23283e-002	0.00107
10	$e(t-2)$	-3.47584e-001	0.00093
11	$e(t-3)$	4.00556e-001	0.00045
12	$e(t-5)$	9.64407e-003	0.00042
13	$e(t-7)$	-2.14539e-001	0.00012
14	$e(t-8)$	-5.24350e-002	0.00009
Note: The CPU time spent on selecting the process model terms from all the candidate model term set is 20.59s.			

Table 5 The regressors, parameters and ERRs estimated using MPOLS for a terrestrial magnetospheric dynamic system.

Number k	Terms $p_k^{(MPOLS)}(t)$	Parameters $\theta_k^{(MPOLS)}$	$ERR_k \times 100\%$
1	$y(t-1)$	9.92291e-001	95.64488
2	$\varphi_{0,-2}(u(t-1))$	1.02467e-001	1.31859
3	$\phi_{0,-3}(y(t-1))$	6.50852e-001	1.22031
4	$\varphi_{4,11}(u(t-1))$	-4.06704e-002	0.81145
5	$\varphi_{2,-1}(y(t-2))$	2.29453e-002	0.60765
6	$\varphi_{2,2}(y(t-2))$	1.10544e-001	0.08649
7	$\varphi_{4,3}(u(t-2))$	3.67041e-001	0.01626
8	$\varphi_{2,1}(u(t-5))$	6.17316e-002	0.00545
9	$\varphi_{4,4}(y(t-4))$	-5.45452e-003	0.00486
10	$e(t-1)$	5.66383e-003	0.00118
11	$e(t-2)$	2.86554e-002	0.00073
12	$e(t-4)$	-7.00413e-002	0.00029
13	$e(t-5)$	-3.90424e-002	0.00013
14	$e(t-7)$	1.19670e-002	0.00020
15	$e(t-8)$	3.28276e-002	0.00008
16	$e(t-9)$	-7.32255e-003	0.00006
Note: The CPU time spent on selecting the process model terms from all the candidate model term set is 1.38s.			

7. Conclusions

A novel hybrid modelling framework, which combines polynomial models with multiresolution wavelet decompositions, has been proposed for nonlinear input-output system identification. In a wavelet-NARMAX model, or simply WANARMAX, a high-dimensional system is initially expressed as a supposition of a number of low-dimensional submodels, and then each submodel is expanded using polynomial models and multiresolution wavelet decompositions. The new WANARMAX model structure not only significantly alleviates the difficulty of the curse-of-dimensionality for high-order and high-dimensional nonlinear system modelling, but also makes it possible to sufficiently utilise the global property of polynomial models and the local property of wavelet representations simultaneously.

A large number of potential model terms are usually involved in a WANARMAX model when each submodel is expanded using multiresolution wavelet decompositions. Most of the model terms are redundant and only a small number of significant model terms need to be included in the final model. Either the widely-used forward OLS algorithm or the new MPOLS algorithm proposed here can be used to select the significant model terms. The computational cost of the MPOLS algorithm is much less than that of the OLS algorithm. However, the MPOLS is less efficient than the forward regression OLS, that is, for the same given problem, it is nearly always true that the MPOLS selects more model terms than that selected by OLS with the same threshold value for termination. The MPOLS routine also tends to produce model predicted outputs that are not as good as those from an OLS identified model.

The WANARMAX model can be used to describe a wide class of nonlinear systems including severely nonlinear systems. The linear or low-order nonlinear trends of the system can be easily tracked by polynomial models and the local nonlinear behaviour can be captured by wavelet decompositions. This enables the WANARMAX model to be more flexible than either a single polynomial model or a wavelet decomposition model.

Acknowledgment

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

References

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

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

Billings, S.A. and Coca, D.(1999), Discrete wavelet models for identification and qualitative analysis of chaotic systems, *International Journal of Bifurcation and Chaos*, **9(7)**, 1263-1284.

Billings, S.A., Korenberg, M. and Chen, S.(1988), Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm, *International Journal of Systems Science*, **19(8)**,1559-1568.

Billings, S.A. and Leontaritis,I.J.(1982), Parameter estimation techniques for nonlinear systems, *The 6th IFAC Symposium on Identification and Systems Parameter Estimation*, Washington, pp 427-432.

Billings,S.A. and Zhu,Q.M.(1995), Model validation tests for multivariable nonlinear models including neural networks, *International Journal of Control*, **62(4)**,749-766.

Billings,S.A., and Voon,W.S.F.(1986), Correlation based model validity tests for nonlinear models, *International Journal of Control*, **44(1)**,235-244.

Brown, M., and Harris, C.J.(1994), *Neuralfuzzy Adatptive Modelling and Control*. Englewood Cliffs, NJ: Prentice-Hall.

Campbell, C.(2002), Kernel methods: a survey of current techniques, *Neuralcomputing*, **48**, 63-84.

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

Chen,S.,Billings, S.A.,Cowan, C.F.N., and Grant, P.W.(1990a), Nonlinear system identification using radial basis functions, *International Journal of Systems Science.*, **21(12)**, 2513-2539.

Chen,S., Billings,S.A., and Grant,P.M.(1990b), Nonlinear system identification using neural networks, *International Journal of Control*, **51(6)**, 1191-1214.

Chen, S, Cowan, C.F.N., Grant, P.M. (1991), Orthogonal least-squares learning algorithm for radial basis function networks, *IEEE Trans Neural Networks*, 2 (2), 302-309.

Chen,S.,Billings, S. A. and Grant, P. W.(1992), Recursive hybrid algorithm for nonlinear system identification using radial basis function network, *International Journal of Control.*, **55(5)**,1051-1070.

Chen,S., Billings,S.A., and Luo,W.(1989), Orthogonal least squares methods and their application to non-linear system identification, *International Journal of Control*, **50(5)**,1873-1896.

Chen, Z.H. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society, Series B (Methodological)***,** **55(2)**, 473-491.

Chui,C. K.(1992), *An Introduction to Wavelets.* Boston; London : Academic Press.

Chui, C. K. and Wang, J. H.(1992), On compactly supported spline wavelets and a duality principle, *Trans. of the American Mathematical Society*, **330(2)**, 903-915.

Coca, D. and Billings, S.A.(2001), Non-linear system identification using wavelet multiresolution models, *International Journal of Control*, **74(18)**,1718-1736.

Daubechies,I.(1992), *Ten lectures on wavelets*. Philaelphia, Pennsylvania : Society for Industrial and Applied Mathematics.

Delgado, A., Kambhamp,A. C., and Warwick, K.(1995), Dynamic recurrent neural-network for system identification and control, *IEE Proceedings-Control Theory and Applications*, **142(4)**, 307-314.

Friedman,J.H. and Stuetzle, W.(1981), Projection pursuit regression, *Journal of the American Statistical Association*, **76(376)**, 817-823.

Friedman,J. H.(1991), Multivariate adaptive regression splines, *The Annals of Statistics*, **19(1)**, 1-67.

Haykin,S.(1994), *Neural networks: a comprehensive foundation*. New York : Macmillan; Oxford : Maxwell Macmillan International.

Hong, X. and Harris, C. J.(2001), Nonlinear model structure detection using optimum experimental design and orthogonal least squares, *IEEE Transactions On Neural Networks*, **12(2)**, 435-439.

Kavli, T. (1993), ASMOD—An algorithm for adaptive spline modelling of observational data, *International Journal of Control*, **58(4)**,947-967.

Korenberg, M., Billings, S.A., Liu, Y. P. and McIlroy P.J.(1988), Orthogonal parameter estimation algorithm for non-linear stochastic systems, *International Journal of Control*, **48(1)**,193-210.

Lee, K.L., and Billings, S.A. (2002), Time series prediction using support vector machines, the orthogonal and the regularized orthogonal least-squares algorithms, *International Journal of Systems Science.*, **33(10)**, 811-821.

Leontaritis,I.J. and Billings, S.A.(1985), Input-output parametric models for non-linear systems, (part I: deterministic non-linear systems; part II: stochastic non-linear systems), *Int. Journal of Control*, **41(2)**,303-344.

Ljung, L. (1987), *System Identification: Theory for the User*. New Jersey: Prentice-Hall.

Mallat,S.G.(1989),A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. On Pattern analysis and machine intelligence*, **11(7)**,674-693.

Mallat, S.G., and Zhang, Z.(1993), Matching pursuits with time-frequency dictionaries, *IEEE Transactions on*

Signal Processing, **41(12)**, 3397-3415.

Pearson, R. K.(1995),Nonlinear input/output modelling, *Journal of Process Control*, **5(4)**, 197-211.

Pearson, R.K.(1999), *Discrete-time dynamic models*, New York; Oxford: Oxford University Press.

Schumaker, L.L.(1981), *Spline Functions: Basic theory*. New York: John Wiley & Sons.

Wang, L.X. and Mendel, J.M.(1992), Fuzzy basis functions, universal approximations, and orthogonal least squares learning, *IEEE Trans Neural Networks*, **3(5)**,807-814.

Wei, H.L., and Billings, S.A.(2002), Identification of time-varying systems using multi-resolution wavelet models, *International Journal of Systems Science*, **33(15)**,1217-1228.

Wei, H.L., Billings, S.A. and Liu J. (2004). Term and variable selection for nonlinear system identification. *International Journal of Control*, **77(1)**, 86-110.

Xu, J., and Ho, D.W.C. (2002), A basis seletion algorithm for wavelet neural networks, *Neurocomputing*, **48**, 681-689.

Yamada, T., and Yabuta, T. (1993), Dynamic system identification using neural networks, *IEEE Transactions on Systems Man and Cybernetics*, **23(1)**, 204-211.

Zhang, Q., and Benveniste,A.(1992), Wavelet networks, *IEEE Trans. Neural Networks*, **3(6)**, 889-898.

Zhang,Q. (1997),Using wavelet network in nonparametric estimation, *IEEE Trans. Neural Networks*, **8(2)**, 227-236.

Zhu, Q.M. and Billings, S.A.(1996), Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks, *International Journal of Control*, **64(5)**,871-886.