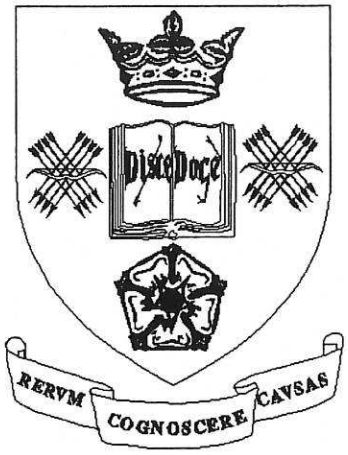
The Wavelet-NARMAX Representation: A Hybrid Model Structure Combining Polynomial Models with Multiresolution Wavelet Decompositions

S. A. Billings, H. L. Wei

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



Research Report No. 841

June 2003

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

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

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 (*Nonlinear AutoRegressive Moving Average with eXogenous* inputs) model (Billings and Leontaritis 1982, Leontaritis and Billings 1985), neural networks (Chen et al. 1990a, Billings et al. 1992, Chen and Billings 1992b, Yamada and Yabuta 1993, Delgado et al. 1995), radial basis function networks (Chen et al 1990b, 1992a), 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 13^{th} 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. Wavelet transform and wavelet decompositions

Wavelet analysis is based on a wavelet prototype function, called the *analysing wavelet*, *mother wavelet*, or simply *wavelet*. Temporal analysis is performed using a contracted, high-frequency version of the same function. Because the signal to be studied can be represented in terms of wavelet decompositions, data operations can also be performed using the corresponding wavelet coefficients.

2.1 The continuous wavelet transform

From wavelet theory, the continuous wavelet transform (CWT) of a given function $f \in L^2(R)$ with respect to the *mother wavelet* φ is defined as (Chui 1992, Daubechies 1992).

$$(W_\varphi f)(a,b) = \int_{-\infty}^{\infty} f(x)\varphi^*_{(a,b)}(x)dx \tag{1}$$

where $\varphi^*_{(a,b)}(x)$ indicates the complex conjugate of the function $\varphi_{(a,b)}(x)$, which is obtained by dilating and translating the mother wavelet $\varphi(x)$ as follows

$$\varphi_{(a,b)}(x) = a^{-\frac{1}{2}}\varphi\left(\frac{x-b}{a}\right), \quad a \in R^+, b \in R \tag{2}$$

The CWT (12) is invertible subject to a mild restriction imposed on the wavelet φ

$$C_\varphi = \int_0^\infty \frac{|\hat{\varphi}(\omega)|^2}{\omega} d\omega < \infty \tag{3}$$

in the sense that

$$f(x) = \frac{1}{C_\varphi}\int_0^\infty \frac{da}{a^2}\int_{-\infty}^{\infty}[(W_\varphi f)(a,b)]\varphi_{(a,b)}(x)db \tag{4}$$

where $\hat{\varphi}$ is the Fourier transform of the function φ. Equation (1) states that the continuous wavelet transform $(W_\varphi f)(a,b)$ is the correlation of $f(x)$ with a scaling (dilation) a and a shift (translation) b. The inverse transform (4) guarantees that the function $f(x)$ can be reconstructed from the CWT and it can be interpreted in at least two different ways. On the one hand, this shows how to reconstruct the function f from the wavelet transform and, on the other hand, the inverse transform gives a recipe showing how to write any arbitrary f as a superposition of the wavelet functions $\varphi_{(a,b)}(x)$.

2.2 Wavelet series

In practical applications the CWT is often discretised in both the scaling and dilation parameters for computational efficiency. Based on this discretization, wavelet decompositions can be obtained to provide an alternative basis function representation. Take the univariate wavelet as an example. The most popular approach to discretise the CWT is to restrict the dilation and translation parameters to a dyadic lattice as $a = 2^{-j}$ and $b = k2^{-j}$ with $j,k \in Z$. Other non-dyadic ways of discretization are also available.

Let $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$ be a wavelet family with respect to $j,k \in Z$. It is can be proved that under some mild assumptions for $\varphi_{j,k}(x)$, any $f \in L^2(R)$ can be uniquely described as (Chui 1992)

$$f(x) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{j,k}\varphi_{j,k}(x) \tag{5}$$

where the convergence of the series in (24) is in $L^2(R)$, namely

$$\lim_{J_1,J_2,K_1,K_2\to\infty}\left\| f(x) - \sum_{j=-J_1}^{J_2}\sum_{k=-K_1}^{K_2} c_{j,k}\varphi_{j,k}(x)\right\| = 0 \tag{6}$$

Eq. (5) is called a *wavelet series*. In comparison with the CWT, the wavelet series is more computationally efficient. But this is obtained at the expense of increased restrictions on the choice of the basic wavelet φ. The wavelet series (5) can be extended to the *d*-dimensional case by taking tensor products of one-dimensional wavelets or by choosing the radial types of wavelets.

2.3 Mulitresolution wavelet decompositions

It is known that for identification problems based on the regression representation it is useful to have a basis of orthogonal (semi-orthogonal or bi-orthogonal) functions whose support can be made as small as required and which provides a universal approximation to any $L^2(R)$ function with arbitrary desired accuracy. One of the original objectives of wavelet theory was to construct orthogonal (semi-orthogonal) basis in $L^2(R)$. The principles for constructing orthogonal wavelets are as follows:

(*i*) The family $\{\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k), j,k \in Z\}$ constitutes an orthogonal basis for the space $L^2(R)$;

(*ii*) There exists a function ϕ, called a *scaling function* related to the mother wavelet φ, such that the elements

of the family $\{\phi(t-k)\}_{k\in Z}$ are mutually orthogonal;

(*iii*) For $\forall j \in Z$, the family $\{\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), k \in Z\}$ constitute an orthogonal basis for $L^2(R)$;

(*iv*) The basic function φ and the scaling function ϕ are related by some deterministic equations.

To satisfy the above aims, an orthogonal wavelet system can be constructed using *multiresolution analysis* (MRA)(Mallat 1989, Chui 1992). Let W_j ($j \in Z$) denote some wavelet subspaces, which are defined as the closure of the linear span of the wavelet functions $\{\varphi_{j,k}\}_{k\in Z}$, namely

$$W_j = \overline{span}\{\varphi_{j,k}, k \in Z\} \tag{7}$$

which satisfy

$$W_i \cap W_j = \{\varnothing\}, \text{ for any } i \neq j \tag{8}$$

where the over-bar denotes closure. It follows that $L^2(R)$ can be decomposed as a direct sum of the spaces W_j:

$$L^2(R) = \cdots \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus \cdots \tag{9}$$

in the sense that every function $f \in L^2(R)$ has a unique decomposition

$$f(x) = \cdots + g_{-1}(x) + g_0(x) + g_1(x) + \cdots = \sum_{j\in Z} g_j(x) \tag{10}$$

The circles around the plus signs in (9) indicate "orthogonal sums". The decomposition of (9) is usually called an *orthogonal decomposition* of $L^2(R)$.

For each $j \in Z$, consider the closed subspaces of $L^2(R)$

$$V_j = \cdots \oplus W_{j-2} \oplus W_{j-1}, \ j \in Z \tag{11}$$

which have the following properties:

(*i*) $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$,

(*ii*) $\overline{(\bigcup_{j\in Z} V_j)} = L^2(R)$ (the over-bar here indicates closure),

(*iii*) $\bigcap_{j\in Z} V_j = \{\varnothing\}$,

(*iv*) $V_{j+1} = V_j \oplus W_j, \ \forall j \in Z$,

(*v*) $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}, \ \forall j \in Z$,

(*vi*) $f(x) \in V_j \Leftrightarrow f(x - 2^j k) \in V_j, \ \forall j,k \in Z$,

(*vii*) $\{\phi(t-k)\}_{k\in Z}$ is an orthogonal basis for V_0.

It is clear that every function $f \in L^2(R)$ can be approximated as closely as desirable by the projections $P_j f$ in V_j. Another important intrinsic property of these spaces is that more and more variations of $P_j f$ are

removed as $j \to -\infty$. In fact, these variations are peeled off, level by level in decreasing order of the rate of variations (frequency bands) and stored in the complementary W_j, shown in property (*iv*).

Assume that the wavelet φ and the corresponding scaling function ϕ constitute an orthogonal wavelet system, then any function $f \in L^2(R)$ can be expressed as the following *multiresolution wavelet decomposition*

$$f(x) = \sum_k \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \geq j_0}\sum_k \beta_{j,k}\varphi_{j,k}(x) \tag{12}$$

where the wavelet coefficients $\alpha_{j_0,k}$ and $\beta_{j,k}$ can be calculated in theory by the inner products:

$$\alpha_{j_0,k} = < f, \phi_{j_0,k} > = \int f(x)\varphi^*_{j_0,k}(x)dx \tag{13}$$

$$\beta_{j,k} = < f, \varphi_{j,k} > = \int f(x)\varphi^*_{j,k}(x)dx \tag{14}$$

and j_0 is an arbitrary integer representing the lowest resolution or scaling level. Notice from (9) that if $j_0 \to -\infty$, the approximation representation (12) becomes the wavelet decomposition (5). In addition, based on (11) and the properties of MRA, any function $f \in L^2(R)$ can be arbitrarily closely approximated in V_J for some sufficiently large integer *J*. That is, for any $\varepsilon > 0$, there exists a sufficiently large integer *J*, such that

$$\left\| f(x) - \sum_k < f, \phi_{J,k} > \phi_{J,k}(x) \right\| < \varepsilon \tag{15}$$

This means that in a wavelet series representation, the wavelet bases can be replaced by orthogonal scaling functions with a large resolution scale.

Using the concept of *tensor products*, the multiresolution decomposition (12) 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). Let $f \in L^2(R^d)$, then $f(x)$ can be represented by the *multiresolution wavelet decomposition* as

$$f(x_1, \cdots, x_d) = \sum_k \alpha_{j_0,k}\Phi_{j_0,k}(x_1, \cdots, x_d) + \sum_{j \geq j_0}\sum_k \sum_{l=1}^{2^d-1} \beta^{(l)}_{j,k}\Psi^{(l)}_{j,k}(x_1, \cdots, x_d) \tag{16}$$

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

$$\Phi_{j_0,k}(x_1, \cdots, x_d) = 2^{j_0 d/2}\prod_{i=1}^{d}\phi(2^{j_0}x_i - k_i) \tag{17}$$

$$\Psi^{(l)}_{j,k}(x_1, \cdots, x_d) = 2^{jd/2}\prod_{i=1}^{d}\eta^{(i)}(2^{j}x_i - k_i) \tag{18}$$

with $\eta^{(i)} = \phi$ or φ (scalar scaling function or the mother wavelet) but at least one $\eta^{(i)} = \varphi$. In the two-dimensional case, the multiresolution approximation can be generated, for example, in terms of the dilation and translation of the two-dimensional scaling and wavelet functions

$$\begin{cases} \Phi_{j,k_1,k_2}(x_1,x_2)=\phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2) \\ \Psi^{(1)}_{j,k_1,k_2}(x_1,x_2)=\phi_{j,k_1}(x_1)\varphi_{j,k_2}(x_2) \\ \Psi^{(2)}_{j,k_1,k_2}(x_1,x_2)=\varphi_{j,k_1}(x_1)\phi_{j,k_2}(x_2) \\ \Psi^{(3)}_{j,k_1,k_2}(x_1,x_2)=\varphi_{j,k_1}(x_1)\varphi_{j,k_2}(x_2) \end{cases} \tag{19}$$

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:

$$y(t)=f(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),e(t-1),\cdots,e(t-n_e))+e(t) \tag{20}$$

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 (20) 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 (20) is the polynomial representation which takes the function $f(\cdot)$ as a polynomial of degree ℓ and gives the form as

$$\begin{aligned} y(t)=\theta_0+\sum_{i_1=1}^{n} f_{i_1}(x_{i_1}(t))+\sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}(x_{i_1}(t),x_{i_2}(t))+\cdots \\ +\sum_{i_1=1}^{n}\cdots\sum_{i_\ell=i_{\ell-1}}^{n} f_{i_1 i_2\cdots i_\ell}(x_{i_1}(t),x_{i_2}(t),\cdots,x_{i_\ell}(t))+e(t) \end{aligned} \tag{21}$$

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

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

$$x_k(t)=\begin{cases} y(t-k) & 1\le k\le n_y \\ u(t-(k-n_y)) & n_y+1\le k\le n_y+n_u \\ e(t-(k-n_y-n_u)) & n_y+n_u+1\le k\le n_y+n_u+n_e \end{cases} \tag{22}$$

The degree of a multivariate polynomial is defined as the highest order among the 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 $\ell = 2+1+2=5$. 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

$$y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)) + e(t) \tag{23}$$

Similar to (21), (23) can be described using a polynomial representation with

$$x_k(t) = \begin{cases} y(t-k), & 1 \le k \le n_y \\ u(t-k+n_y), & n_y + 1 \le k \le n = n_y + n_u \end{cases} \tag{24}$$

3.2 The wavelet-based AVONA 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 as below

$$\begin{aligned} y(t) &= f(x_1(t), x_2(t), \cdots, x_n(t)) \\ &= f_0 + \sum_{i=1}^{n} f_i(x_i(t)) + \sum_{1 \le i < j \le n} f_{ij}(x_i(t), x_j(t)) + \sum_{1 \le i < j < k \le n} f_{ijk}(x_i, x_j, x_k) + \cdots \\ &\quad + \sum_{1 \le i_1 < \cdots < i_m \le n} f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)) + \cdots + f_{12\cdots n}(x_1(t), x_2(t), \cdots, x_n(t)) + e(t) \end{aligned} \tag{25}$$

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 (22) or (24), the ANOVA expansion (25) can then be viewed as a special form of the NARMAX or NARX models for dynamic input and output systems.

The expansion (25) can be referred to as the ANOVA decomposition of the NARAMX or NARX models. Although the ANOVA expansion (25) involves up to 2^n different functional components, experience shows that a truncated representation containing the components up to the bivariate functional terms is often sufficient

$$y(t) = f_0 + \sum_{p=1}^{n} f_p(x_p(t)) + \sum_{p=1}^{n} \sum_{q=p+1}^{n} f_{pq}(x_p(t), x_q(t)) + e(t) \tag{26}$$

This can often provide a satisfactory description of $y(t)$ for many high dimensional problems providing that the input variables are properly selected. The presence of only low order functional components does not necessarily imply that the high order variable interactions are not significant, nor does it mean the nature of the nonlinearity of the system is less severe. An exhaustive search for all the possible submodel structures of (25) 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.

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 (25) and (26). In the present study, however, mutiresolution wavelet decompositions will be chosen to describe the functional components. 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 (12) and (16) as

$$f_p(x_p(t)) = \sum_k \alpha_{j_1,k}^{(p)} \phi_{j_1,k}(x_p(t)) + \sum_{j \ge j_1} \sum_k \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)), \quad p = 1,2,\cdots,n, \tag{27}$$

$$\begin{aligned} f_{pq}(x_p(t), x_q(t)) = & \sum_{k_1} \sum_{k_2} \alpha_{j_2;k_1,k_2}^{(pq)(1)} \phi_{j_2,k_1}(x_p(t)) \phi_{j_2,k_2}(x_q(t)) \\ & + \sum_{j \ge j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(1)} \phi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)) \\ & + \sum_{j \ge j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(2)} \varphi_{j,k_1}(x_p(t)) \phi_{j,k_2}(x_q(t)) \\ & + \sum_{j \ge j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(3)} \varphi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)), \quad 1 \le p < q \le n. \end{aligned} \tag{28}$$

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

$$y(t) = f(x(t)) = f^P(x(t)) + f^W(x(t)) + f^E(\xi(t)) + e(t) \tag{29}$$

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 (24), $f^P(x(t))$ 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 (29) in various forms and the following are some examples

$$f^P(x(t)) = a_0 + \sum_{p=1}^{n} a_p x_p(t) \tag{30a}$$

$$f^P(x(t)) = a_0 + \sum_{p=1}^{n} a_p x_p(t) + \sum_{p=1}^{n} \sum_{q=p}^{n} b_{pq} x_p(t) x_q(t) \tag{30b}$$

$$f^W(x(t)) = \sum_{p=1}^{n} f_p(x_p(t)) \tag{31a}$$

$$f^W(x(t)) = \sum_{p=1}^{n} f_p(x_p(t)) + \sum_{p=1}^{n} \sum_{q=p}^{n} f_{pq}(x_p(t), x_q(t)) \tag{31b}$$

$$f^E(\xi(t)) = \sum_{p=1}^{n_e} c_p e(t-p) \tag{32a}$$

$$f^E(\xi(t)) = \sum_{p=1}^{n_e} c_p e(t-p) + \sum_{p=1}^{n_e}\sum_{q=p}^{n_e} c_{pq} e(t-p)e(t-q) \tag{32b}$$

$$f^E(\xi(t)) = \sum_{p=1}^{n_e} c_p e(t-p) + \sum_{p=1}^{n_e}\sum_{q=p}^{n_e} c_{pq} e(t-p)e(t-q) + \sum_{p=1}^{n}\sum_{q=1}^{n_e} d_{pq} x_p(t)e(t-q) \tag{32c}$$

where the functional components $f_p(x_p(t))$ (p=1,2,...,n) and $f_{pq}(x_p(t), x_q(t))$ ($1 \le p < q \le n$) in (29) can be expressed using the multiresolution wavelet decompositions (27) and (28). Take (30b),(31b) and (32b) as an example, the WANARMAX model (29) can be described as

$$\begin{aligned}
y(t) = a_0 &+ \sum_{p=1}^{n} a_p x_p(t) + \sum_{p=1}^{n}\sum_{q=p+1}^{n} b_{pq} x_p(t) x_q(t) \\
&+ \sum_{p=1}^{n}\sum_{k} \alpha_{j_1,k}^{(p)} \phi_{j_1,k}(x_p(t)) + \sum_{p=1}^{n}\sum_{j\ge j_1}\sum_{k} \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)) \\
&+ \sum_{1\le p<q\le n}\sum_{k_1}\sum_{k_2} \alpha_{j_2;k_1,k_2}^{(pq)(1)} \phi_{j_2,k_1}(x_p(t))\phi_{j_2,k_2}(x_q(t)) \\
&+ \sum_{1\le p<q\le n}\sum_{j\ge j_2}\sum_{k_1}\sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(1)} \varphi_{j,k_1}(x_p(t))\phi_{j,k_2}(x_q(t)) \\
&+ \sum_{1\le p<q\le n}\sum_{j\ge j_2}\sum_{k_1}\sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(2)} \varphi_{j,k_1}(x_p(t))\phi_{j,k_2}(x_q(t)) \\
&+ \sum_{1\le p<q\le n}\sum_{j\ge j_2}\sum_{k_1}\sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(3)} \varphi_{j,k_1}(x_p(t))\varphi_{j,k_2}(x_q(t)) \\
&+ \sum_{p=1}^{n_e} c_p e(t-p) + \sum_{p=1}^{n_e}\sum_{q=p}^{n_e} c_{pq} e(t-p)e(t-q) + e(t)
\end{aligned} \tag{33}$$

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 and highest resolution scales in the multiresolution decomposition are determined, the model (33) can be rearranged and converted into a linear-in-the-parameters regression model of the form

$$y(t) = \sum_{i=1}^{M_1} \theta_i^P p_i^P(t) + \sum_{j=1}^{M_2} \theta_j^W p_j^W(t) + \sum_{k=1}^{M_3} \theta_k^E p_k^E(t) + e(t) \tag{34}$$

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. $M_1 = 1 + (n_y + n_u)(n_y + n_u + 1)/2$, $M_3 = n_e$ 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 (34) is the Wavelet-NARX, or simply WANARX model

$$y(t) = \sum_{i=1}^{M_1} \theta_i^P p_i^P(t) + \sum_{j=1}^{M_2} \theta_j^W p_j^W(t) + e(t) \tag{35}$$

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 (34) 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 (34) 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 subset 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 (34) can be expressed as a linear-in-the-parameters equation of the form

$$y(t) = \sum_{m=1}^{M} \theta_m p_m(t) + e(t) \tag{36}$$

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 $M_1 + M_2 + 1 \le m \le M = M_1 + M_2 + M_3$. $\theta_m (m = 1,2,\cdots,M)$ are parameters to be estimated. Define

$$P^{(m)} = \{p_{i_k} : 1 \le i_k \le M;\ \ k = 1,2,\cdots,m\},\ \ m=1,2,\ldots,M, \tag{37}$$

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

$$P^{(1)} \subset P^{(2)} \subset \cdots \subset P^{(m)} \subset \cdots \tag{38}$$

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

$$y(t) = \sum_{k=1}^{M_0} \theta_{i_k} p_{i_k}(t) + e(t) \tag{39}$$

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 (36) based on the error reduction ratio (ERR) (Billings et al. 1988, 1989). The procedure can be briefly summarised as follows:

A compact matrix form corresponding to (36) is

$$Y = P\Theta + \Xi \tag{40}$$

where $Y=[y(1),y(2),\cdots,y(N)]^T$, $P=[p_1,p_2,\cdots,p_M]$, $p_i=[p_i(1),p_i(2),\cdots,p_i(N)]^T$, $\Theta=[\theta_1,\theta_2,\cdots,\theta_M]^T$, $\Xi=[e(1),e(2),\cdots,e(N)]^T$. Assume that the regression matrix P can be orthogonally decomposed as

$$P = WA \tag{41}$$

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 $W^TW=D=diag[d_1,d_2,\cdots,d_M]$. The space spanned by the orthogonal basis $w_1,w_2,\cdots,w_M$ is the same as that spanned by the basis set $p_1,p_2,\cdots,p_M$, and (40) can be expressed as

$$Y = (PA^{-1})(A\Theta) + \Xi = WG + \Xi \tag{42}$$

where $G=[g_1,g_2,\cdots,g_M]^T$ is an auxiliary parameter vector, which can be calculated directly from Y and W by means of the property of orthogonality as

$$G = D^{-1}W^TY \tag{43}$$

or

$$g_i = \frac{Y^Tw_i}{w_i^Tw_i}, \quad i=1,2,\cdots,M \tag{44}$$

The parameter vector Θ is related to G by the equation $A\Theta=G$, and this can be solved using either a classical or modified Gram-Schmidt algorithm (Chen et al. 1989).

The number M of all the candidate terms in model (36) 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 ($M_0 << M$). Assume that the residual signal $e(t)$ in the model (36) is uncorrelated with the past outputs of the system, then the output variance can be expressed as

$$\frac{1}{N}Y^TY = \frac{1}{N}\sum_{i=1}^{M} g_i^2 w_i^T w_i + \frac{1}{N}\Xi^T\Xi \tag{45}$$

Note that the output variance consists of two parts, one is 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 other part, $(1/N)\Xi^T\Xi$, 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 w_i, and the ith error reduction ratio, ERR_i, introduced by w_i, can be defined as

$$ERR_i = \frac{g_i^2 w_i^T w_i}{Y^T Y} \times 100\% = \frac{(Y^T w_i)^2}{(Y^T Y)(w_i^T w_i)} \times 100\%, \quad i = 1,2,\cdots,M, \tag{46}$$

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. 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 the classical 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 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* 2:** The assumption that the regression matrix P is full rank in columns is unnecessary in the iterative forward OLS algorithm. In fact, if the M columns of the matrix P are linearly dependent, and assuming that the rank in columns is L ($<M$), then the algorithm will stop at the M_1-th step.

***Remark* 3:** 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, 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 $\{p_i\}_{i=1}^{M}$

$$y(t) = \sum_{i=1}^{M_0} g_i w_i(t) + e(t) \tag{47}$$

which is equivalent to

$$y(t) = \sum_{i=1}^{M_0} \theta_{\ell_i} p_{\ell_i}(x(t)) + e(t) \tag{48}$$

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

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1M_0} \\ 0 & 1 & \cdots & a_{2M_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{M_0-1,M_0} \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{49}$$

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

4.2 Matching pursuit orthogonal least squares (MPOLS) algorithm

Note that in the forward Gram-Schmidt 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 (36) or (40), 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

$$y(t)=c_m p_m(t)+\xi_m(t) \tag{50}$$

achieve a minimum in the sense that

$$\frac{1}{N}\sum_{t=1}^{N}\xi_{\ell_1}^2(t)=\frac{1}{N}\sum_{t=1}^{N}(y(t)-c_{\ell_1}p_{\ell_1}(t))^2=\min_m\left\{\frac{1}{N}\sum_{t=1}^{N}[y(t)-c_m p_m(t)]^2\right\} \tag{51}$$

The "best" matching regressor p_{ℓ_1} can be found by means of a geometrical approach, see Figure 1. From Figure 1,

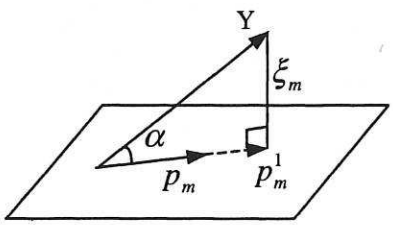


Figure 1 Diagram of least squares algorithm

$$\cos\alpha=\frac{Y^T p_m}{\sqrt{Y^T Y}\sqrt{p_m^T p_m}} \tag{52}$$

$$\left\|p_m^1\right\|=\|Y\|\cos\alpha=\frac{Y^T p_m}{\sqrt{p_m^T p_m}} \tag{53}$$

Thus

$$\sum_{t=1}^{N}\xi_m^2(t)=\|\xi_m\|^2=\|Y\|^2-\left\|p_m^1\right\|^2=Y^T Y-\frac{(Y^T p_m)^2}{p_m^T p_m} \tag{54}$$

Therefore,

$$\ell_1 = \arg\max_m \left\{ \frac{(Y^T p_m)^2}{p_m^T p_m}, 1 \le m \le M \right\} \tag{55}$$

Set $q_1(t) = p_{\ell_1}(t)$, $w_1(t) = q_1(t)$, $g_1 = (Y^T w_1)/(w_1^T w_1)$, $ERR_1 = g_1^2 (w_1^T w_1)/(Y^T Y)$, 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 (51) and (55), ℓ_2 should be chosen as

$$\ell_2 = \arg\max_m \left\{ \frac{(\eta_1^T p_m)^2}{p_m^T p_m}, 1 \le m \le M, m \ne \ell_1 \right\} \tag{56}$$

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

$$w_2 = q_2 - \frac{w_1^T q_2}{w_1^T w_1} w_1 \tag{57}$$

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

$$\ell_k = \arg\max_m \left\{ \frac{(\eta_{k-1}^T p_m)^2}{p_m^T p_m}, 1 \le m \le M, m \ne \ell_1, m \ne \ell_2, \cdots, m \ne \ell_{k-1} \right\} \tag{58}$$

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

$$w_k = q_k - \frac{w_1^T q_k}{w_1^T w_1} w_1 - \frac{w_2^T q_k}{w_2^T w_2} w_2 - \cdots - \frac{w_{k-1}^T q_k}{w_{k-1}^T w_{k-1}} w_{k-1} \tag{59}$$

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 jth 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 (47) and (48).

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 discusses some practical problems in the implementation of a wavelet-NARMAX model and 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, resolution scale determination in the wavelet decomposition submodels, and model validity tests.

5.1 Significant variable selection

The first problem encountered in WANARMAX modelling is how to determine which variables should be included in the model. For a linear regression model, the model terms and the variables are exactly the same, they are the regressors. However, variables and terms are generally distinct in a typical nonlinear model. It is often the case in practice that some of the variables $x_1(t)$, $x_2(t), \cdots, x_n(t)$ in the model (29) are redundant and only a subset of these variables is significant. Inclusion of redundant variables might result in a much more complex model since the number of model terms increases dramatically with the number of variables. Furthermore, including redundant variables might lead to a large number of free parameters in the model, and as a consequence the model may become oversensitive to the training data and is likely to exhibit poor generalisation properties. Therefore, it is important to determine which variables should be included in the model.

The purpose of variable selection is to pre-select a subset consisting of the significant variables or to eliminate redundant variables from all the candidate variables of a system under study prior to model term detection. It is required that the selected significant variables alone should sufficiently represent the system.

5.2 Data pre-processing

The original observational input-output data $u(t)$ and $y(t)$ (t=1,2, …,N) are often normalized into a standard domain, for example the unit interval [0,1], for the convenience of implementation. This is especially true when a compactly supported wavelet and/or a scaling function are used in the wavelet model (31). Taking the univariate Haar wavelet (the first-order B-spline wavelet) as an example, it is much easier to select the starting resolution level and the range of the shift parameters if the sample data has been normalized to [0, 1].

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.

5.3 Determination of the resolution scale and shift parameter

In theory, the multiresolution wavelet decomposition (12) and (16) are infinite expansions. In practice, however, it is impossible to include infinite terms in these wavelet decompositions. Therefore, the infinite decompositions are always truncated at appropriate dilations (resolutions) and translations.

Consider the one-dimensional multiresolution wavelet decomposition (12) and assume that the function $f(x)$ is defined in [0, 1] and x is an independent variable which is uniformly distributed in $[0,1]$, that is, x itself can be considered as "time", then the basis functions (dilated and translated versions of the wavelet and scaling function) in the multiresolution wavelet decomposition (12) are mutually orthogonal and the decomposition is unique. Assume also that the Haar wavelet (the first-order B-spline wavelet) and scaling function are used in the decomposition, then a truncated decomposition with the initial resolution scale j_0 and the highest resolution scale j_{max}=J can be expressed as

$$f(x)=\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0,k}\phi_{j_0,k}(x)+\sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1}\beta_{j,k}\varphi_{j,k}(x) \tag{60}$$

Clearly, the higher the upper resolution scale level J, the more accurate the approximation is. A recommended approach for selecting the highest scale J is to utilize the features of the sampled signal, for example, the natural frequency of the signal to be approximated. Assume that the maximum natural frequency of the sampled signals is $f_{\max}$, the highest scale can be empirically chosen as $j_{\max}=[\log_2(Mf_{\max})]$, where M is a positive number, say between 2^4 and 2^6, and $[\cdot]$ denotes taking the integer value of the corresponding number (Wei and Billings 2002).

In practical identification problems, however, the orthogonality of the mutiresolution wavelet decomposition might be lost, since most observational data fail to satisfy the uniform distribution assumption. Also in dynamic systems modelling, the variables, $x_i(t)$ (i=1,2,…,n) in (27) and (28) are usually the lagged outputs $y(t-i)$

($i = 1,2,\cdots,n_y$) or lagged inputs $u(t\text{-}j)$ ($j = 1,2,\cdots,n_u$), which are usually sparse in the normalized interval [0, 1]. The empirical rule $j_{\max} = [\log_2(Mf_{\max})]$ for selecting the highest resolution scale can however still be used.

For a compactly supported wavelet, the shift parameter k is determined by the corresponding resolution scale j. For example, at a given scale j, the shift parameter k in the Haar wavelet multiresolution decomposition (61) is chosen as $k = 0,1,\cdots,2^{j-1}$. Generally, for a compactly supported wavelet $\varphi(x)$ with an integer support $S_\varphi = [0, K_s]$, where K_s is integer, the support for the dilated and translated wavelet $\varphi_{j,k}(x) = 2^{j/2}(2^j x - k)$ is $[2^{-j}k,\ 2^{-j}(K_s + k)]$, therefore, the shift parameter k at a resolution scale j should be taken as $-(K_s - 1) \le k \le 2^{j-1} - 1$. This is also true for a compactly supported scaling function $\phi(x)$.

5.4 Model validity tests

Several methods of model validation have been proposed for nonlinear system identification (Billings and Voon 1986, Billings and Zhu 1995). The noise sequence $e(t)$ in the NARMAX model (20) is assumed to be independent, bounded and uncorrelated with the past inputs and outputs, and no other a priori information is known. Let $\hat{f}(\cdot)$ represent an estimated model for the system $f(\cdot)$, the residuals $\varepsilon(\cdot)$ can be estimated as

$$\begin{aligned}\varepsilon(t) &= y(t) - \hat{y}(t) \\ &= y(t) - \hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),\varepsilon(t-1),\cdots,\varepsilon(t-n_e))\end{aligned} \tag{61}$$

If the model structure and parameter values are correct, $\varepsilon(\cdot)$ will be unpredictable from all linear and nonlinear combinations of past inputs and outputs. For nonlinear SISO systems, this can be tested by computing the following correlation functions (Billings and Voon 1986)

$$\begin{cases} \gamma_{\varepsilon\varepsilon}(\tau) = \delta(\tau), & \forall\,\tau \\ \gamma_{u\varepsilon}(\tau) = 0, & \forall\,\tau \\ \gamma_{\overline{u}^2\varepsilon}(\tau) = 0, & \forall\,\tau \\ \gamma_{\overline{u}^2\varepsilon^2}(\tau) = 0, & \forall\,\tau \\ \gamma_{(u\varepsilon)\varepsilon}(\tau) = 0, & \tau \ge 0 \end{cases} \tag{62}$$

where $\overline{u}^2(t) = u^2(t) - \overline{u^2(t)} = u^2(t) - E[u^2(t)]$, and the correlation function $\gamma_{\xi\zeta}(\cdot)$ can be estimated as

$$\gamma_{\xi\zeta}(\tau) = \frac{\sum_{t=1}^{N-\tau}\xi(t)\zeta(t+\tau)}{\sqrt{\left(\sum_{t=1}^{N-\tau}\xi^2(t)\right)\left(\sum_{t=1}^{N-\tau}\zeta^2(t)\right)}} \tag{63}$$

The first two conditions in (62) form the traditional tests used in linear system identification. The remaining three conditions involve cross correlation tests between the input and residuals, by which all possible omitted nonlinear terms can be detected. In practice, if these correlation functions fall within the confidence intervals at a given significance level $\alpha(0<\alpha<1)$, say α=0.05, which corresponds to the 95% confidence interval, the model is viewed as adequate and acceptable. For large N (the data length), these confidence intervals are approximately $\pm 1.96/\sqrt{N}$.

Although the model validation tests are normally justified on the basis of the calculation of correlations between the input and the residuals, Billings and Zhu (1995) showed that the use of outputs enhances the performance of the tests and allows the number of individual correlation tests to be reduced. When the output is introduced, only two tests are required

$$\begin{cases} \gamma_{\overline{(y\varepsilon)}\overline{\varepsilon}^2}(\tau)=\lambda\delta(\tau) \\ \gamma_{\overline{(y\varepsilon)}\overline{u}^2}(\tau)=0 \end{cases} \quad \text{for } \forall\tau \tag{64}$$

and these can be more efficient in cases of MIMO system identification.

An alternative approach for validating the model is to check the prediction capability of the fitted model. This can reveal severe model deficiencies which would otherwise go undetected. The measure of the predictive capability of a model is not based on the one-step-ahead prediction errors, but based on the multi-step-ahead (the long-term) prediction errors. Generally, the observational data are split into an estimation set which is used to identify the model, and a testing set (or validation set) which is used to judge the predictive ability of the model. This is often referred to as *cross-validation*, and provides an efficient tool not only for validating the estimated model but also for the estimation of the model. The most powerful approach to validate an estimated model is to check the model behaviour, using the model predicted output (MPO) defined as

$$\hat{y}_{mpo}(t)=\hat{f}(\hat{y}_{mpo}(t-1),\cdots,\hat{y}_{mpo}(t-n_y),u(t-1),\cdots,u(t-n_u),0,\cdots,0) \tag{65}$$

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

$$\hat{e}_{mpo}(t)=y(t)-\hat{y}_{mpo}(t) \tag{66}$$

where $y(t)$ (t=1,2,…,N) are the system measurements.

5.5 An iterative implementation procedure

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 parameters 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 $t=1,2,\ldots,N$.

(*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 (30a)-(32c).

(*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 (35) 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 (47) and (48).

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

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

$$\begin{aligned}\varepsilon^{(0)}(t) &= y(t)-\hat{y}(t)\\ &= y(t)-\hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),0,\cdots,0)\\ &= y(t)-\sum_{i=1}^{M_0} g_i^{(k)} w_i^{(k)}(t)\end{aligned} \tag{67}$$

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

$$\begin{aligned}\varepsilon^{(k)}(t) &= y(t)-\hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),\varepsilon^{(k-1)}(t-1),\cdots,\varepsilon^{(k-1)}(t-n_e))\\ &= y(t)-\sum_{j=1}^{M_0+m_e}\theta_{\ell_j}^{(k)} p_{\ell_j}(t)\end{aligned} \tag{68}$$

or

$$\varepsilon^{(k)}(t) = y(t) - \sum_{j=1}^{M_0+m_e} g_j^{(k)} w_j^{(k)}(t) \tag{69}$$

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

$$\sum_{m=1}^{M_0+m_e} \frac{\left| g_m^{(k)} - g_m^{(k-1)} \right|}{\left| g_m^{(k)} \right|} \le \delta_1 \tag{70}$$

and

$$\sum_{t=1}^{N} \left| \varepsilon^{(k)}(t) - \varepsilon^{(k-1)}(t) \right|^2 \le \delta_2 \tag{71}$$

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 (*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.

6.1 Simulated example—a nonlinear system

The following nonlnear input-output system

$$\begin{aligned} y(t) = & \frac{y(t-1)y(t-2) + y(t-1)y(t-3) + y(t-2)y(t-3)}{1 + y^2(t-1) + y^2(t-2) + y^2(t-3)} \\ & + 2[\sin(y(t-1))][\cos(y(t-2))] + 2[\sin(y(t-2))][\cos(y(t-3))] \\ & + 2[\sin(y(t-3))][\cos(y(t-1))] + 6.0u^2(t-1) + u^3(t-2) \end{aligned} \tag{72}$$

was simulated using a system input with the form

$$u(t) = 2\sin(\pi t/25) + 0.5\sin(\pi t/30) + 0.02\exp[\sin(\pi t/40)] \tag{73}$$

The estimation set consists of 500 input-output data points which are shown in Figure 2. Setting n_y= 5, n_u =3, and initially selecting the model structure for this system to be of the form

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

$$+\sum_{p=1}^{3}\sum_{q=p}^{3}c_{pq}u(t-p)u(t-q)+\sum_{p=1}^{5}\sum_{q=1}^{3}d_{pq}y(t-p)u(t-q)$$
$$+\sum_{p=1}^{5}f_p(y(t-p))+\sum_{p=1}^{3}f_{p+5}(u(t-p)) \tag{74}$$

where each function $f_p(\cdot)$ can be described using the multiresolution wavelet decomposition (27) as

$$f_p(x_p(t))=\sum_{k\in K^0}\alpha_{0,k}^{(p)}\phi_{0,k}(x_p(t))+\sum_{j=0}^{4}\sum_{k\in K_j}\beta_{j,k}^{(p)}\varphi_{j,k}(x_p(t)),\ p=1,2,\cdots,8, \tag{75}$$

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 $K^0=\{-3,-2,-1,0\}$ and $K_j=\{-6,-5,\cdots,-1,\ 0,1,\cdots,2^j-1\}$ for j=0,1,2,3,4.

The initial model (74) contains 565 model regressors, but most of these are likely to be redundant and should be removed form the initial model. Both the OLS and MPOLS algorithms were used to select the significant regressors, and two parsimonious models were obtained

$$y(t)=\hat{f}^{(OLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-3))=\sum_{k=1}^{12}\theta_k^{(OLS)}p_k^{(OLS)}(t) \tag{76}$$

$$y(t)=\hat{f}^{(MPOLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-3))=\sum_{k=1}^{20}\theta_k^{(MPOLS)}p_k^{(MPOLS)}(t) \tag{77}$$

The parameters, regressors and the corresponding error reduction ratios (ERR) of the models (76) and (77) are listed in Table 2 and Table 3, respectively. A comparison of the model predicted outputs and the measurements, are shown in Figure 3. 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 (77) is worse than that from the OLS identified model (76), this behaviour will be investigated in a later paper.

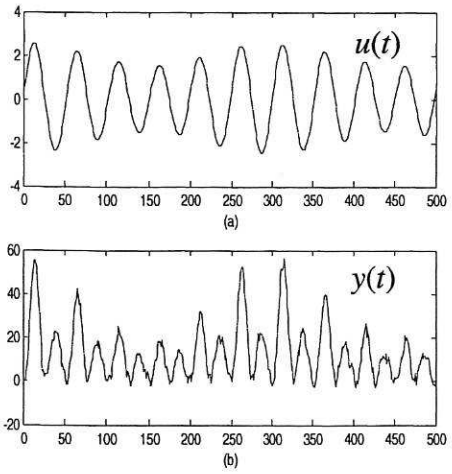


Figure 2 The input and output data of the system described by Eq. (72). (a) Input; (b) Output.

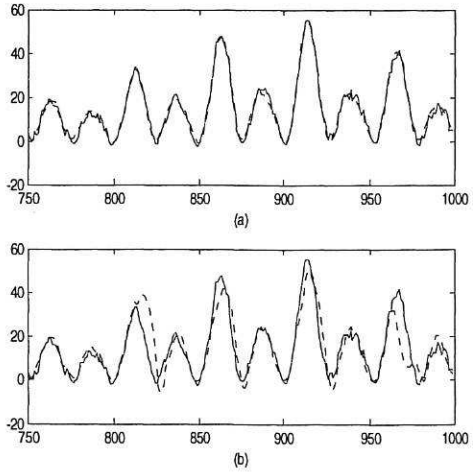


Figure 3 The comparison of the model predicted output (MPO) and the measurements for the system described by Eq (72). (a) The model predicted outputs based on the model (76) ; (b) The model predicted outputs based on the model (77). (The solid line denotes the measurements, and the dashed line denotes the model predicted outputs.)

Table 2 The regressors, parameters and the corresponding ERRs estimated using OLS for the system described by Eq (72)

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 threshold value ρ =0.006, the CPU time spent on selecting these model terms from all the candidate model term set is 23.23s.			

Table 3 The regressors, parameters and the corresponding ERRs estimated using MPOLS for the system described by Eq (72)

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 threshold value ρ =0.008, the CPU time spent on selecting these model terms from all the candidate model term set is 2.03s.			

6.2 A terrestrial magnetosphere dynamic 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 4 shows 1000 data points of measurement 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 (29). 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 $\{y(t\text{-}1), \ldots, y(t\text{-}5), u(t\text{-}1), \ldots, u(t\text{-}5)\}$ were initially selected using a variable selection algorithm. The initial model was chosen as below:

$$\begin{aligned} y(t) &= f(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),\cdots,e(t-10)) \\ &= a_0 + \sum_{p=1}^{10} a_p x_p(t) + \sum_{p=1}^{10}\sum_{q=p}^{10} b_{pq} x_p(t) x_q(t) + \sum_{p=1}^{10} f_p(x_p(t)) \\ &\quad + \sum_{p=1}^{10} c_p e(t-p) + e(t) \end{aligned} \tag{78}$$

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. (75) .

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 parsimonious models were obtained

$$\begin{aligned} y(t) &= \hat{f}^{(OLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),...,e(t-10)) \\ &= \sum_{k=1}^{14} \theta_k^{(OLS)} p_k^{(OLS)}(t) \end{aligned} \tag{79}$$

$$\begin{aligned} y(t) &= \hat{f}^{(MPOLS)}(y(t-1),\cdots,y(t-5),u(t-1),\cdots,u(t-5),e(t-1),\cdots,e(t-10)) \\ &= \sum_{k=1}^{16} \theta_k^{(MPOLS)} p_k^{(MPOLS)}(t) \end{aligned} \tag{80}$$

The parameters, regressors and the corresponding error reduction ratios (ERR) of the selected regressors in models (79) and (80) are listed in Table 4 and Table 5, respectively. A comparison of the model predicted outputs and the measurements are shown in Figure 5, 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 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.

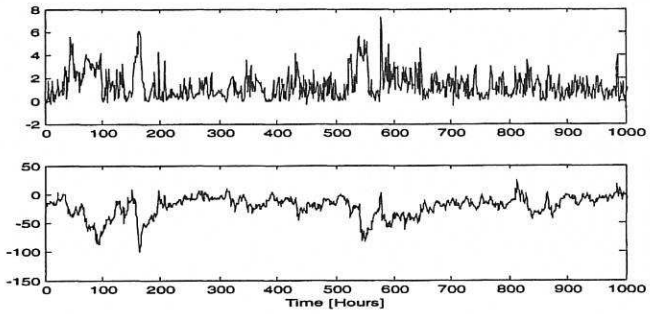


Figure 4 The input and output data of a terrestrial magnetospheric dynamic system. (a) Input; (b) Output.

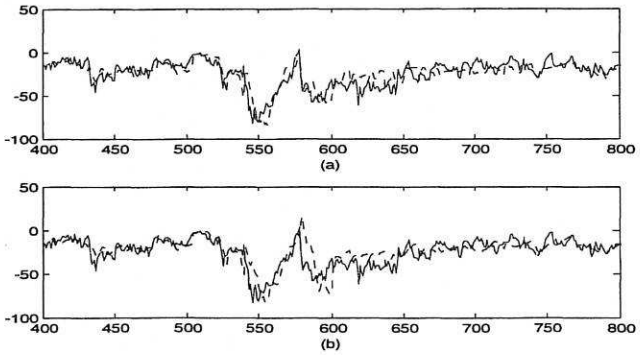


Figure 5 The comparison of the model predicted output (MPO) and the measurements for a terrestrial magnetospheric dynamic system. (a) The model predicted outputs based on the model (79) ; (b) The model predicted outputs based on the model (80). (The solid line denotes the measurements, and the dashed 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. 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 Voon,W.S.F.(1986), Correlation based model validity tests for nonlinear models, *International Journal of Control*, **44(1)**,235-244.

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., 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., Jamaluddin, H.B., and Chen,S.(1992), Properties of neural networks with applications to modelling nonlinear dynamic systems, *International Journal of Control*, **55(1)**,193-224.

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 Coca, D.(1999), Discrete wavelet models for identification and qualitative analysis of chaotic

systems, *International Journal of Bifurcation and Chaos*, **9(7)**, 1263-1284.

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

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

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.(1992a), 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.(1992b), Neural networks for nonlinear system modelling and identification, *International Journal of Control*, **56(2)**, 319-346.

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), *Descrete-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.

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.