Article:
Billings, S.A. and Wei, H.L. (2005) A new class of wavelet networks for nonlinear system identification. IEEE Transactions on Neural Networks, 16 (4). pp. 862-874. ISSN: 1045-9227

https://doi.org/10.1109/TNN.2005.849842

A New Class of Wavelet Networks for Nonlinear System Identification

Stephen A. Billings and Hua-Liang Wei

***Abstract*—A new class of wavelet networks (WNs) is proposed for nonlinear system identification. In the new networks, the model structure for a high-dimensional system is chosen to be a superimposition of a number of functions with fewer variables. By expanding each function using truncated wavelet decompositions, the multivariate nonlinear networks can be converted into linear-in-the-parameter regressions, which can be solved using least-squares type methods. An efficient model term selection approach based upon a forward orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) is applied to solve the linear-in-the-parameters problem in the present study. The main advantage of the new WN is that it exploits the attractive features of multiscale wavelet decompositions and the capability of traditional neural networks. By adopting the analysis of variance (ANOVA) expansion, WNs can now handle nonlinear identification problems in high dimensions.**

***Index Terms*—Nonlinear autoregressive with exogenous inputs (NARX) models, nonlinear system identification, orthogonal least squares (OLS), wavelet networks (WNs).**

I. Introduction

WAVELET theory [1]–[3] has been extensively studied in recent years and has been widely applied in various areas throughout science and engineering. Dynamical system modeling and control using artificial neural networks (ANNs), including radial basis function networks (RBFNs), has also been studied widely and a number of systematic approaches have been proposed [4]–[16]. The idea of combining wavelets with neural networks has led to the development of wavelet networks (WNs), where wavelets were introduced as activation functions of the hidden neurons in traditional feedforward neural networks with a linear output neuron. Although it was considered that WNs were popularized by the work in [17]–[19], the origin of WNs can be traced back to the earlier work of Daugman [20], where Gabor wavelets were used for image classification and compression.

The wavelet analysis procedure is implemented with dilated and translated versions of a mother wavelet. Since signals of interest can usually be expressed using wavelet decompositions, signal processing algorithms can be performed by adjusting only the corresponding wavelet coefficients. In theory, the dilation (scale) parameter of a wavelet can be any positive real value and the translation (shift) can be an arbitrary real number. This is referred to as the continuous wavelet transform. In practice, however, in order to improve computation efficiency, the values of the shift and scale parameters are often limited to some discrete lattices. This is then referred to as the discrete wavelet transform.

Both continuous and discrete wavelet transforms have been introduced to implement neural networks. Existing WNs can, therefore, be catalogued into the following two types.

- *Adaptive WNs*, where wavelets as activation functions stem from the continuous wavelet transform and the unknown parameters of the networks include the weighting coefficients (the outer parameters of the network) and the dilation and translation factors of the wavelets (the inner parameters of the network). These parameters can be viewed as coefficients varying continuously as in conventional neural networks and can be learned by gradient type algorithms.
- *Fixed grid WNs*, where the activation functions stem from the discrete wavelet transforms and unlike in adaptive neural networks, the unknown inner parameters of the networks vary on some fixed discrete lattices. In such a WN, the positions and dilations of the wavelets are fixed (predetermined) and only the weights have to be optimized by training the network. In general, gradient type algorithms are not needed to train such a network. An alternative solution for training this kind of network is to convert the networks into a linear-in-the-parameters problem, which can then be solved using least squares type algorithms.

The concept of adaptive WNs was introduced in [18] as an approximation route which combined the mathematical rigor of wavelets with the adaptive learning scheme of conventional neural networks into a single unit. Adaptive WNs have been successfully applied to nonlinear static function approximation and classification [17], [21]–[24], and dynamical system modeling [25], [26]. Clearly, to train an adaptive WN, the gradients with respect to all the unknown parameters have to be expressed explicitly. The calculation of gradients may be heavy and complicated in some cases especially for high-dimensional models. In addition, most gradient type algorithms are sensitive to initial conditions, that is, the initialization of wavelet neural networks is extremely important to obtain a fast convergence for a given algorithm [27]. Another problem that needs to be considered for training an adaptive WN is how to determine the initial number of wavelets associated with the network. These drawbacks often limit the application of

Manuscript received March 22, 2004; October 26, 2004. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) (U.K.).

The authors are with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: s.billings@sheffield.ac.uk; w.hualiang@sheffield.ac.uk).

Digital Object Identifier 10.1109/TNN.2005.849842

adaptive WNs to low dimensions for dynamical identification problems.

Unlike adaptive WNs, in a fixed grid WN, the number of wavelets as well as the scale and translation parameters can be determined in advance. The only unknown parameters are the weighting coefficients, that is, the outer parameters, of the network. The WN is now a linear-in-the-parameters regression, which can then be solved using least squares techniques. As will be discussed in Section III-D, the number of candidate wavelet terms in a fixed grid WN often increases dramatically with the model order. As a consequence, fixed grid WNs are often limited to low dimensions.

Inspired by the well-known analysis of variance (ANOVA) expansions [28], [29], a new class of fixed grid WNs is introduced in the present study for nonlinear system identification. In the new WNs, the model structure of a high-dimensional system is initially expressed as a superimposition of a number of functions with fewer variables. By expanding each function using truncated wavelet decompositions, the multivariate nonlinear networks can then be converted into linear-in-the-parameter problems, which can be solved using least-squares type methods. The new WNs are, therefore, in structure different from either the existing WNs [18], [24]–[26], [30]–[32] or wavelet mutiresolution models [33], [34]. A wavelet multiresolution model is in structure similar to a fixed grid WN. The former, however, forms a wavelet multiresolution decomposition similar to an ordinary multiresolution analysis (MAR), which involves not only a wavelet, but also another function, the associated *scaling function*, where some additional requirements should be satisfied. An efficient model term detection approach based on a forward orthogonal least squares (OLS) algorithm, along with the error reduction ratio (ERR) criterion [35]–[37] is applied to solve the linear-in-the-parameters problem in the present study.

II. Presentation of Nonlinear Dynamical Systems

A wide range of nonlinear systems can be represented using the nonlinear autoregressive with exogenous inputs (NARX) model. Taking single-input–single-output (SISO) systems as an example, this can be expressed by the following nonlinear difference equation:

$$y(t) = f(y(t-1),\ldots,y(t-n_y) \times u(t-1),\ldots,u(t-n_u)) + e(t) \quad (1)$$

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 $e(t)$ is immeasurable but is assumed to be bounded and uncorrelated with the inputs.

Several approaches can be applied to realize the representation (1) including polynomials [36], [41], [42], neural networks [4]–[6], [8] and other complex models [43]. In the present study, an additive model structure will be adopted to represent the NARX model (1). The multivariate nonlinear function f in the model (1) can be decomposed into a number of functional components via the well-known functional ANOVA expansions [28], [29]

$$\begin{aligned} y(t) &= f(x_1(t), x_2(t), \ldots, x_n(t)) \\ &= f_0 + \sum_{i=1}^{n} f_i(x_i(t)) + \sum_{1\le i<j\le n} f_{ij}(x_i(t), x_j(t)) \\ &\quad + \sum_{1\le i<j<k\le n} f_{ijk}(x_i(t), x_j(t), x_k(t)) + \cdots \\ &\quad + \sum_{1\le i_1<\cdots<i_m\le n} f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_m}(t)) \\ &\quad + \cdots + f_{12\cdots n}(x_1(t), x_2(t), \ldots, x_n(t)) + e(t) \end{aligned} \quad (2)$$

where $x(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T$ and

$$x_k(t) = \begin{cases} y(t-k), & 1 \le k \le n_y \\ u(t-k+n_y), & n_y+1 \le k \le n = n_y + n_u \end{cases}. \quad (3)$$

The first functional component f_0 is a constant to indicate the intrinsic varying trend; $f_i, f_{ij}, \ldots$, 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 ith 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. The ANOVA expansion (2) can be viewed as a special form of the NARX model for input and output dynamical systems. Although the ANOVA decomposition of the NARX model (1) 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 [44], [45]. 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 (2) 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. Note that the function $f_{12\cdots j}(\cdot)(j = 1,2,\ldots,n)$ does not contain terms that can be written as functional components with an order smaller than j. It was also assumed that each functional component of the desired ANOVA expansion is square-integrable over the domain of interest for given data sets. In practice, the constant term f_0 can often set to be zero. If the constant term is different from zero for a given system, it can then be approximated by a wavelet expansion providing that the approximation is restricted to a compact subset of R^n.

It will generally be true that, whatever the data set and whatever the modeling approach, the structure of the final model will be unknown in advance. It is, therefore, not possible to know that expansions up to trivariate terms will always be sufficient in the ANOVA expansion. This is why model validation methods, which are independent of the model fitting procedure and the

model type, are an important part of the nonlinear autoregressive moving average with exogenous inputs (NARMAX) modeling methodology [9]. 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 bivariate or trivariate 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 bi- and tri-variate terms. The identification proceeds a stage at a time and uses model validation as the decision making process. This is the NARMAX methodology [9], which is implemented here, and which mimics the traditional approach to analytical modeling. 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 cutoff 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 cutoff, 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 [46] can be chosen to express the functional components in model (2). It is known that wavelet basis functions have the property of localization in both time and frequency. With the excellent approximation properties associated with multiscale decompositions, wavelet models outperform many other approximation schemes and are well-suited for approximating arbitrary functions [1], even functions with sharp discontinuities. It has been shown that the intrinsic nonlinear dynamics related to real nonlinear systems can easily be captured by an appropriately fitted wavelet model consisting of a small number of wavelet basis functions [31], [34], and this makes wavelet representations more adaptive compared with other basis functions. In the present study, therefore, wavelet decompositions, which are discussed in the next section, will be chosen to describe the functional components in the additive models (2), and this was referred to as the wavelet-NARX model, or the WANARX [45], where multiresolution wavelet decompositions were employed and a class of compactly supported wavelets was considered.

III. WNs and Truncated Wavelet Decompositions

This section briefly reviews some results on wavelet decompositions and WNs which are relevant to the present work. For more details about these results, see [1]–[3], [18], [31], [47], and [48]. In the following, it is assumed that the independent variable x of a function $f \in L^2(\mathbf{R})$ of interest is defined in the unit interval $[0, 1]$. In addition, for the sake of simplicity, one-dimensional (1-D) wavelets are considered as an example to illustrate related concepts.

A. Wavelet Decompositions

Let ψ be a mother wavelet and assume that there exists a denumerable family derived from ψ

$$\Omega = \left\{ \psi_{(a_t,b_t)} : \psi_{(a_t,b_t)}(x) = \frac{1}{\sqrt{a_t}} \times \psi\left(\frac{x-b_t}{a_t}\right), a_t \in R^+, b_t \in R \right\} \quad (4)$$

where a_t and b_t are the scale and translation parameters. The normalization factor $1/\sqrt{a_t}$ is introduced so that the energy of $\psi_{(a_t,b_t)}$ is preserved to be the same as that of ψ. Rearrange the elements of Ω so that

$$\Omega = \{\psi_t : t \in \Gamma\} \quad (5)$$

where Γ is an index set which might be finite or infinite. Note that the double index of the elements of Ω in (4) is replaced by a single index as shown in (5). Under the condition that ψ generates a frame, it is guaranteed that any function $f \in L^2(\mathbf{R})$ can be expanded in terms of the elements in Ω in the sense that [1], [2], [18]

$$f(x) = \sum_{t\in\Gamma} c_t \psi_t(x) \quad (6)$$

$$f(x) = \sum_{t\in\Gamma} c_t \psi_{(a_t,b_t)}(x) = \sum_{t\in\Gamma} c_t \frac{1}{\sqrt{a_t}} \psi\left(\frac{x-b_t}{a_t}\right) \quad (7)$$

where c_t are the decomposition coefficients or weights. Equation (7) is called the *wavelet frame decomposition*.

In practical applications the decomposition (7) is often discretized for computational efficiency by constricting both the scale and dilation parameters to some fixed lattices. In this way, wavelet decompositions can be obtained to provide an alternative basis function representation. The most popular approach to discetize (7) is to restrict the dilation and translation parameters to a dyadic lattice as $a_t = 2^{-j}$ and $b_t = k2^{-j}$ with $j, k \in \mathbf{Z}$ ($\mathbf{Z}$ is the set of all integers). Other nondyadic ways of discretization are also available. For the dyadic lattice case, (7) becomes

$$f(x) = \sum_j \sum_k c_{j,k} \psi_{j,k}(x) \quad (8)$$

where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and $j, k \in \mathbf{Z}$.

Note that in general a frame provides a redundant basis. Therefore, the decompositions (7) and (8) are usually not unique, even for a tight frame. Under some conditions, it is possible to make the decomposition (8) to be unique and in this case this decomposition is called a *wavelet series* [1]. An orthogonal wavelet decomposition, which requires stronger restrictions than a wavelet frame, is a special case of a wavelet series. Although orthogonal wavelet decompositions possess several attractive properties and provide concise representations for arbitrary signals, most functions are excluded from being candidate wavelets for orthogonal decompositions. On the contrary, much more freedom on the choice of the wavelet

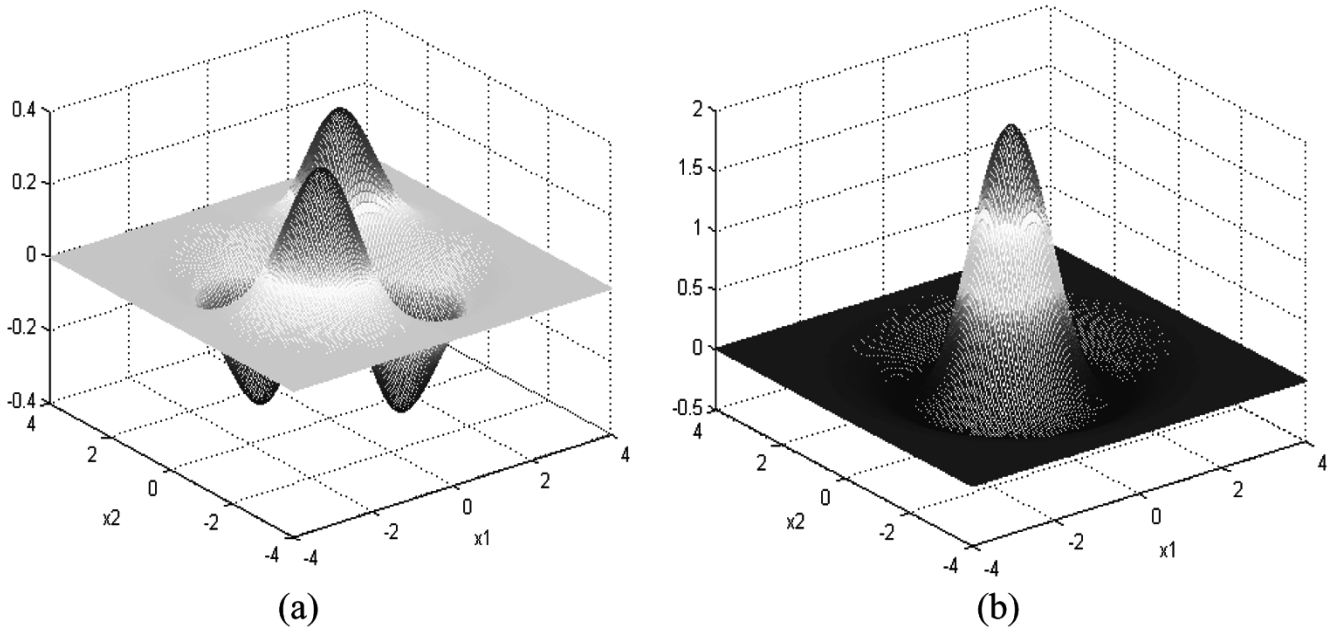

Fig. 1. Two-dimensional Gaussian and Marr mother wavelets. (a) Gaussian wavelet. (b) Marr wavelet.

functions is given to a wavelet frame by relaxing the orthogonality.B.

B. WNs

In practical applications for either static function learning or dynamical system modeling, it is unnecessary and impossible to represent a signal using an infinite decomposition of the form (7) or (8) in terms of wavelet basis functions. The decompositions (7) and (8) are therefore often truncated at an appropriate accuracy. WNs are in effect nothing but a truncated wavelet decomposition. Taking the decomposition (8) as an example, an approximation to a function $f \in L^2(\mathbf{R})$ using the truncated wavelet decomposition with the coarsest resolution j_0 and the finest resolution $j_{\max}$ can be expressed in the following:

$$f(x) = \sum_{j=j_0}^{j_{\max}} \sum_{k \in K_j} c_{j,k} \psi_{j,k}(x) \tag{9}$$

where $K_j (j = j_0, j_0 + 1, \ldots, j_{\max})$ are subsets of $\mathbf{Z}$ and often depend on the resolution level j for all compactly supported wavelets and for most rapidly vanishing wavelets that are not compactly supported. The details on how to determine K_j at a given level j will be discussed later. Define

$$\Omega_1 = \{\psi_{j,k} : j = j_0, j_0 + 1, \ldots, j_{\max}, k \in K_j\}. \tag{10}$$

Assume that the number of wavelets in Ω_1 is M. For convenience of description, rearrange the elements of Ω_1 so that the double index (j, k) can be indicated by a single index $m = 1, 2, \ldots, M$ in the sense that

$$f(x) = \sum_{m=1}^{M} c_m \psi_m(x). \tag{11}$$

The truncated wavelet decompositions (9) and (11) are referred to as *fixed grid WNs*, which can be implemented using neural network schemes by choosing different types of wavelets and employing different training/learning algorithms. This will be discussed in Section IV.

Note that although the WN (9) or (11) involves different resolutions or scales, it cannot be called a multiresolution decomposition related to wavelet MAR, which involves not only a wavelet, but also another function, the associated *scaling function*, where some additional requirements should be satisfied.

C. Extending to High Dimensions

The results for the 1-D case described previously can be extended to high dimensions. One commonly used approach is to generate separable wavelets by the tensor product of several 1-D wavelet functions. For example, an n-dimensional wavelet $\psi^{[n]} : \mathbf{R}^n \mapsto \mathbf{R}$ can be constructed using a scalar wavelet ψ as follows:

$$\psi^{[n]}(x) = \psi^{[n]}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \psi(x_i). \tag{12}$$

Another popular scheme is to choose the wavelets to be some radial functions. For example, the n-dimensional Gaussian type functions can be constructed as

$$\psi^{[n]}(x) = \psi^{[n]}(x_1, x_2, \ldots, x_n) = x_1 x_2 \cdots x_n e^{-(1/2)\|x\|^2} \tag{13}$$

where $\|x\|^2 = x^T x = \sum_{i=1}^{n} x_i^2$. Similarly, the n-dimensional Mexican hat (also called the Marr) wavelet can be expressed as $\psi^{[n]}(x) = (n - \|x\|^2) \exp(-\|x\|^2/2)$. In the present study, the radial wavelets are used to implement WNs. The two-dimensional (2-D) Gaussian and Mexican hat wavelets are shown in Fig. 1.

D. Limitations of Existing WNs

It has been found that most exiting WNs are limited to handling problems in low-dimensional space due to the difficulty

of the so called *curse-of-dimensionality*. The following discussion will illustrate why existing WNs are not readily suitable for high-dimensional problems.

Assume that a function $f \in L^2(\mathbf{R}^n)$ of interest is defined in the unit hypercube $[0,1]^n$. Let ψ be a scalar wavelet function that is compactly supported on $[s_1, s_2]$. From Section III-C, this scalar wavelet can be used to generate an n-dimensional wavelet $\psi^{[n]} : \mathbf{R}^n \mapsto \mathbf{R}$ by (12). This multidimensional wavelet $\psi^{[n]}$ can then be used to approximate the n-dimensional function $f \in L^2(\mathbf{R}^n)$ using the WN (9) in the following:

$$\begin{aligned} f(x) &= f(x_1, x_2, \ldots, x_n) \\ &= \sum_{j=j_0}^{j_{\max}} \sum_{k \in K_j} c_{j,k} \psi_{j,k}^{[n]}(x_1, x_2, \ldots, x_n) \\ &= \sum_{j=j_0}^{j_{\max}} \sum_{k_1 \in K_j} \cdots \sum_{k_n \in K_j} c_{j;k_1,\ldots,k_n} 2^{jn/2} \prod_{i=1}^{n} \psi_{j,k_i}(x_i) \end{aligned} \tag{14}$$

where $k = [k_1, k_2 \ldots, k_n]^T \in \mathbf{Z}^n$ is an n-dimensional index. Noting that $x_i \in [0,1]$ for $i = 1, 2, \ldots, n$ and that the wavelet ψ is compactly supported on $[s_1, s_2]$. Then for a given resolution level j, it can easily be proved that the possible values for k_i should be between $-(s_2 - 1)$ and $2^j - s_1 - 1$, that is, $-(s_2 - 1) \le k_i \le 2^j - s_1 - 1$. Therefore, the number of candidate wavelet terms to be considered at scale level j will be $n_{\text{term}} = s^n$, where $s = 2^j + s_2 - s_1 - 1$. Setting $n = 5$ and $s_2 - s_1 = 5$, this number will be $n_{\text{term}} = 5^5, 6^5, 8^5$, and 12^5 for $j =$ 0,1,2, and 3, respectively. If n and $(s_2 - s_1)$ are set to be 10 and 5, the number of candidate wavelets will then become $n_{\text{term}} = 5^{10}, 6^{10}, 8^{10}$, and 12^{10} for $j =$ 0,1,2, and 3, respectively. This implies that the total number of candidate wavelet terms involved in the WN can become very large even for some low resolution levels ($j \le 3$). This means that the computation task for a medium or high-dimensional WN can become very high. Thus, it can be concluded that high-dimensional WNs will be very difficult if not impossible to implement via a tensor product approach. This is the case where an n-dimensional wavelet is constructed by the tensor product of n scalar wavelets.

Similarly, applications of existing WNs, where the wavelets are chosen to be radial wavelets, are also prohibited from high-dimensional problems by the previously mentioned limitations. In addition, most existing radial WNs possess an inherent drawback, that is, every wavelet term includes all the process variables as in the Gaussian and the Marr mother wavelets. This is unreasonable since in general it is not necessary that every variable of a process interacts directly with all the other variables. Moreover, experience shows that inclusion of the *total-variable-involved* wavelet terms (here a *total-variable-involved term* refers to a model term that involves all the process variables simultaneously) may produce a deleterious effect on the resulting model of a dynamical process and will often induce spurious dynamics. From the point of view of identification studies, it is therefore desirable to exclude the total-variable-involved wavelet terms.

The limitations and drawbacks associated with existing WNs described previously suggest that new WNs need to be constructed to bypass the curse-of-dimensionality to enable the networks to handle more realistic and high-dimensional problems.

IV. New Class of WNs

The structure of the new WNs is based on the ANOVA expansion (2), where it is assumed that the additive functional components can be described using truncated wavelet decompositions. The construction and implementation procedure of the new networks is described as follows.

A. Structure of the New WNs

Consider the m-dimensional functional component $f_{i_1 i_2 \ldots i_m}(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_m}(t))$ in the ANOVA expansion (2). From (9) or (11), $f_{i_1 i_2 \ldots i_m}(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_m}(t))$ can be expressed using an m-dimensional WN as

$$\begin{aligned} &f_{i_1 i_2 \ldots i_m}(x_{i_1}(t), \ldots, x_{i_m}(t)) \\ &\quad = \sum_{j=j_m}^{J_m} \sum_{k_1 \in K_j} \cdots \sum_{k_m \in K_j} c_{j;k_1,\ldots,k_m} \psi_{j;k_1,\ldots,k_m}^{[m]} \\ &\qquad (x_{i_1}(t), \ldots, x_{i_m}(t)) \end{aligned} \tag{15}$$

where the m-dimensional wavelet function $\psi_{j;k_1,\ldots,k_m}^{[m]}(x_{i_1}(t), \ldots, x_{i_m}(t))$ can be generated from a scalar wavelet as in (12) or (13). Taking the 2-D component $f_{pq}(x_p(t), x_q(t))(1 \le p \le q \le n)$ in (2) as an example, this can be expressed using a radial WN as

$$\begin{aligned} f_{pq}(x_p(t), x_q(t)) &= \sum_{j=j_2}^{J_2} \sum_{k_1} \sum_{k_2} c_{j;k_1,k_2} \psi_{j;k_1,k_2}^{[2]} \\ &\quad \times (x_p(t), x_q(t)) \\ &= \sum_{j=j_2}^{J_2} \sum_{k_1} \sum_{k_2} c_{j;k_1,k_2} 2^j \\ &\quad \times \left\{ 2 - [2^j x_p(t) - k_1]^2 \right. \\ &\qquad \left. - [2^j x_q(t) - k_2]^2 \right\} \\ &\quad \times e^{-(1/2)\{[2^j x_p(t) - k_1]^2 + [2^j x_q(t) - k_2]^2\}} \end{aligned} \tag{16}$$

where the Mexican hat function is used. Other wavelets can also be employed.

By expanding each functional component in (2) using a radial WN (15), a nonlinear WN can be obtained and this will be used for nonlinear system identification in the present study. Note that in (16) the scale parameters for each variable of an m-dimensional wavelet are the same. In fact, the scales for different variables of an m-dimensional wavelet are permitted to be different. This may enable the network to be more adaptive and more flexible. However, this will also make the number of candidate wavelet terms increase drastically and even lead to prohibitive calculations for high-dimensional systems. Therefore, the same scales for different variables will be considered here.

B. Determining the Number of Candidate Wavelet Terms

Assume that both the input and the output of a nonlinear system are limited to be in the unit interval $[0, 1]$. If not, both the input and output can be normalized into $[0, 1]$ under the condition that the input and output are bounded in finite intervals [45].

The number of candidate wavelet terms is determined by both the scale levels and translation parameters. For a wavelet with a compact support, it is easy to determine the parameters at a given scale level j. For example, the support of the fourth-order B-spline wavelet [1] is $[0, 7]$. At a resolution scale j, the variation range for the translation parameter k is $-6 \leq k \leq 2^j - 1$. The number of total candidate wavelet terms at different resolution scales in a WN can then be determined.

Most radial wavelets are not compactly supported but rapidly vanishing. Using this property, a radial wavelet can often be truncated at some points such that this radial wavelet becomes quasi-compactly supported. Under this case, the support boundaries are design parameters and some good reference results were obtained for the boundary values given in the following:

$$\left|\psi^{[1]}(x)\right| = |\psi(x)| \leq 0.0013, \quad |x| \geq 4 \tag{17}$$

$$\left|\psi^{[2]}(x_1, x_2)\right| \leq 0.0202, \quad |x_1| \geq 3 \text{or} |x_2| \geq 3. \tag{18}$$

The support of the one and 2-D Gaussian wavelets can then be defined as $S^{[1]} = [-4, 4]$ and $S^{[2]} = [-3, 3]\times[-3, 3]$. Similarly, for the 1-D and 2-D Mexican hat wavelets, $|\psi(x)| \leq 0.005$ for $|x| \geq 4$ and $|\psi^{[2]}(x_1, x_2)| \leq 0.08$ for $|x_1| \geq 3$ or $|x_2| \geq 3$. Therefore, the one and 2-D Mexican hat wavelets can also be defined as $S^{[1]}$ and $S^{[2]}$. The compactly supported one and 2-D Mexican hat wavelets can be defined as

$$\psi^{[1]}(x) = \begin{cases} (1 - x^2)e^{(1/2)x^2} & x \in S^{[1]} = [-4, 4] \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

$$\psi^{[2]}(x) = \begin{cases} (2 - \|x\|^2)e^{(1/2)\|x\|^2} & x \in S^{[2]} = [-3, 3] \times [-3, 3] \\ 0 & \text{otherwise} . \end{cases} \tag{20}$$

The compactly supported Gaussian wavelets can be defined in the same way. The support for three-dimensional (3-D) Gaussian and Mexican wavelet can be defined as $S^{[3]} = [-3, 3] \times [-3, 3] \times [-3, 3]$. Note that from experience the wavelet support boundaries are not critical design parameter, this means that the proposed identification techniques enjoys some robustness with respect to the choice of wavelet boundaries.

For the scalar Gaussian or Mexican hat wavelet, given a resolution scale j, since $|2^j x - k| \leq 4$ and $0 \leq x \leq 1$, the choice for the translation parameter k should satisfy $-3 \leq k \leq 2^j + 3$. This means that the number of candidate 1-D wavelets at a given scale j can be determined beforehand. Similarly, the number of candidate m-dimensional candidate wavelets terms can be determined. Therefore, the number of the total candidate wavelet terms is now deterministic.

C. Significant Term Detection

Assume that M candidate wavelet terms are involved in a WN. The WN can then be converted into a linear-in-the-parameters form

$$y(t) = \sum_{m=1}^{M} \theta_m p_m(t) + e(t) \tag{21}$$

where $p_m(t)(m = 1, 2, \ldots, M)$ are regressors (model terms) produced by the dilated and translated versions of some mother wavelets. For a high-dimensional system, where n_y and/or n_u in (1) are large numbers, the model (21) may involve a great number of model terms. 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. In other words, only a small number of significant terms are necessary to describe a given nonlinear system with a given accuracy. Therefore, there exists an integer M_0 (generally $M_0 \ll M$), such that the model

$$y(t) = \sum_{k=1}^{M_0} \theta_{i_k} p_{i_k}(t) + e(t) \tag{22}$$

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

A fast and efficient model structure determination approach has been implemented using the forward OLS algorithm and the ERR criterion, which was originally introduced to determine which terms should be included in a model [35], [36]. This approach has been extensively studied and widely applied in nonlinear system identification [31], [35], [36], [49]–[52]. The forward OLS algorithm involves a stepwise orthogonalization of the regressors and a forward selection of the relevant terms in (21) based on the ERR [36]. See the Appendix for more details of the forward OLS algorithm.

D. Procedure to Implement the New WNs

Two schemes can be adopted to implement the new WN. One scheme starts from an over constructed model consisting of both low and high dimensional submodels. This means that the library of wavelet basis functions (wavelet terms) used to construct a WN is over-completed. The aim of the estimation procedure is to select the most significant wavelet terms from the deterministic over-completed library, so that the selected model terms describe the system well. Another scheme starts from a low-order submodel, where the library of wavelet basis functions (wavelet terms) used to construct a WN may or may not be completed. The estimation procedure then selects the most significant wavelet terms from the given library. If model validity tests [53], [54] suggest that the selected wavelet terms cannot adequately describe a given system over the range of interest, higher dimensional wavelet terms should then be added to the WN (library). Significant terms are then reselected from the new library. This procedure may repeat several times until a satisfactory model is obtained. These two identification procedures to implement the WN are summarized in the following.

1) Implement a WN Starting From an Over-Constructed Model: This identification procedure contains in general of the following steps.

Step 1) *Data preprocessing*. For convenience of implementation, convert the original observational input–output data $u(t)$ and $y(t)(t = 1, 2, \ldots, N)$ into the unit interval $[0, 1]$. The converted input and output are still denoted by $u(t)$ and $y(t)$.

Step 2) *Determining the model initial conditions*. This includes the following.

i) Select initial values for n_y and n_u.

ii) Select the significant variables from all candidate lagged output and input variables $y(t-1), y(t-2), \ldots, y(t-n_y), u(t-1), u(t-2), \ldots, u(t-n_u)$. This involves the model order determination and variable selection problems.

iii) Determine m, the highest dimension of all the submodels (functional components) in (2).

Step 3) *Identify the WN consisting of functional components up to m-dimensions.*

i) Determine the coarsest and finest resolution scales $j_1, \ldots, j_m$ and $J_1, \ldots, J_m$, where $J_k (1 \leq k \leq m)$ indicates the scales of the associated k-dimensional wavelets. Generally the initial resolution scales $j_k = 0$, and the fines resolution scales $J_k (1 \leq k \leq m)$ can be chosen in a heuristic way.

ii) Expand all the functional components of up to m-dimensions using selected mother wavelets of up to m-dimensions.

iii) Select the significant model terms from the candidate model terms and then form a parsimonious model of the form (22).

Step 4) *Model validity tests*. If the identified mth-order model in Step 3) provides a satisfactory representation over the range considered for the measured input–output data, then terminate the procedure. Otherwise, set $m = m + 1$ and/or $J_k = J_k + 1 (k = 1, 2 \ldots, m+1)$, go to and repeat from Step 3.

2) Implement a WN Starting From Low-Order Submodels: This identification procedure can be summarized in the following.

Step 1) *The same as in 4.4.1.*

Step 2) *Determining the model initial conditions*. This includes: i) and ii) The same as in 4.4.1. iii) Set $m = 1$.

Step 3) *The same as in 4.4.1.*

Step 4) *Model validity tests.*

E. Noise Modeling

In many cases the noise signal $e(t)$ in (1) may be a correlated or colored noise sequence. This is likely to be the case for most real data sets The NARX model (1) will then become the NARMAX model [38]

$$y(t) = f(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u), e(t-1), \ldots, e(t-n_e)) + e(t). \quad (23)$$

Model (23) is obviously more general than the NARX model (1) and which includes as special cases several linear and nonlinear representations [43]. The NARMAX model (23) is easily accommodated in the ANOVA expansion (2) by defining $x_t(t)$ in (3) to include noise terms

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

where $n = n_y + n_u + n_e$. Note that the noise signal $e(t)$ in model (23) is generally unobserved and is often replaced by the model residual sequence. Let $\hat{f}(\cdot)$ represent an estimator for the model $f(\cdot)$, the residuals $\varepsilon(t)$ can then be estimated as

$$\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t) \\ &= y(t) - \hat{f}(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, \\ &\qquad u(t-n_u), \varepsilon(t-1), \ldots, \varepsilon(t-n_e)). \end{aligned} \quad (25)$$

In this case the algorithm in Sections IV-D.1 and II will include an extra step in Step 3) which consists of the following:

- compute the prediction errors $\varepsilon(t)$;
- use the value of $\varepsilon(\cdot)$ from the previous iteration so that noise model terms are included in model $f(\cdot)$.

In some situations it may be possible to use just a linear noise model where

$$\varepsilon(t) = \alpha_1 \varepsilon(t-1) + \cdots + \alpha_{n_e} \varepsilon(t-n_e). \quad (26)$$

But if this is insufficient then $\varepsilon(t-p)$ for $p = 1, 2, \ldots, n_e$ can be included in the ANOVA expansion (2) where $x_k(t)$ is defined as

$$x_k(t) = \begin{cases} y(t-k), & 1 \leq k \leq n_y \\ u(t-k-n_y), & n_y + 1 \leq k \leq n_y + n_u \\ \varepsilon(t-k+n_y+n_e), & n_y + n_u + 1 + \leq k \leq n\,. \end{cases} \quad (27)$$

The model validity tests [53], [54] can be used to determine if the process and noise models are adequate.

V. Examples

Three bench test examples are provided to illustrate the performance of the new WNs. The first data set comes from a simulated continuous-time input–output system, the second is from a high-dimensional chaotic time series, and the third is the sunspot time series. Note that the original data sets used for identification were initially normalized to $[0, 1]$, the identification procedure is therefore performed using normalized variables. The outputs of an identified model can then be recovered to the original system operating domain. The varying bounds of a variable in the original system operating domain were determined by inspecting the data sets available for identification rather than by physical insight.

A. Nonlinear Continuous-Time Input–Output System

Consider the Goodwin equation described by a nonlinear time-invariant continuous-time model [55]

$$\ddot{y}(t) + a \frac{y^2(t) - 1}{y^2(t) + 1} \dot{y}(t) + by(t) + cy^3(t) = u(t) \quad (28)$$

where a, b, and c are time-invariant parameters. Under the initial conditions $\dot{y}(0) = y(0) = 0$ and with $u(t) = A\cos(t)$, $a = 0.1$, $b = -0.5$, $c = 0.5$, $A = 37$, a fourth-order Runge–Kutta algorithm was used to simulate this model with the integral step size $\Delta t = 0.01$, and 3000 equi-spaced samples were obtained from the input and output with a sampling interval of $T = 0.02$ time units. The sampled input and output, $u(k)$ and $\{y(k)\}$ for $k = 1, 2, \ldots, 3000$, were normalized into the unit interval $[0, 1]$ using the fact that $u(k) \in [-37, 37]$ and $y(k) \in [-7, 7]$. The normalized input and output sequences were still designated by $u(k)$ and $y(k)$.

The 3000 data points of input–output samples were divided into two parts: the estimation set consisting of the first 1000 data points was used for WN training and the test set consisting of the remaining 2000 data points was used for model testing. A variable selection algorithm [56] was performed on the estimation data set and three significant variables $\{y(k-1), y(k-2), u(k-1)\}$ were selected. The initial WN was chosen as

$$\begin{aligned} y(k) &= f(y(k-1), y(k-2), u(k-1)) \\ &= \sum_{p=1}^{3} f_p(x_p(k)) + \sum_{p=1}^{2}\sum_{q=2}^{3} f_{pq}(x_p(k), x_q(k)) \\ &\quad + f_{123}(x_1(k), x_2(k), x_3(k)) \end{aligned} \tag{29}$$

where $x_p(k) = y(k-p)$ for $p = 1,2$ and $x_3(k) = u(k-1)$. The 1-D, 2-D, and 3-D Mexican hat radial WNs were used in this example to approximate the univariale functions f_p, the bivariate functions f_{pq}, and the tri-variate function f_{123}, respectively, with the coarsest resolutions $j_1 = j_2 = j_3 = 0$ and finest resolutions $J_1 = 3$ and $J_2 = J_3 = 2$. A forward OLS algorithm, together with the ERR criterion [35]–[37] was applied to select significant model terms. The final identified model was found to be

$$\begin{aligned} y(k) = {} & 0.070\,723\psi_{2,1}(y(k-1)) \\ & + 1.846\,539\psi_{0,3}(y(k-2)) \\ & + 1.734\,865\psi^{[2]}_{0;2,0}(y(k-1), y(k-2)) \\ & - 1.300\,637\psi_{3,6}(u(k-1)) \end{aligned} \tag{30}$$

where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and $\psi^{[2]}_{j,k_1,k_2}(x_1, x_2) = 2^j\psi(2^j x_1 - k_1, 2^j x_2 - k_2)$ are the one and two dimensional compactly supported Gaussian wavelets, j, k, k_1 and k_2 are some integer numbers.

Setting the input signal $u(t) = \cos(k/50)$, and starting from the initial value $y(1) = y(2) = 0.5$ [this is equivalent to the original initial condition $\dot{y}(0) = y(0) = 0$ for (28)], the model (30) was simulated and the output was recovered to its original amplitude by the inverse transform $y_{WN}(k) = y_{\min} + (y_{\max} - y_{\min})y(k)$, where $y_{\max} = -y_{\min} = 7$. The recovered system output from the model (30) was compared with that from the original model (28) over the validation set and is shown in Fig. 2(a) and (b), which clearly indicates that the model (30) provides an excellent representation for the input–output data set generated from the system (28) with an input of sine wave. For a closer inspection of the result, the interval [1600, 2400], where the maximum errors appear as shown in (b), was expanded and this is shown in Fig. 2(c). Note that model predicted outputs or

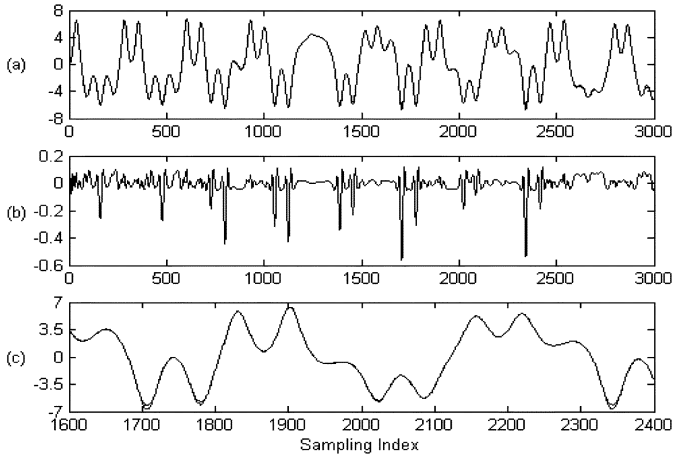

Fig. 2. Comparison of the model output based on the WN (30) with the measurements over the test set. (a) Overlap of the output of the WN (30) and the measurements. (b) Discrepancy between the output of the WN (30) and the measurements. (c) The interval $[1600, 2400]$ was expanded for a closer inspection. In (a) and (c), the solid lines indicate the measurements and the dashed lines indicate the model predicted outputs.

the long term model predictions are used here as a much more severe test compared with one-step-ahead predicted outputs. For comparison, we have also tried other wavelet models using only the total-variable-involved functional exponent $f_{123}(y(k-1), y(k-2), u(k-1))$ in (29) by expanding this exponent using a 3-D radial wavelet decomposition (a traditional 3-D fixed grid WN). For example, with the same input–output data set and the same wavelet parameters as did in the model (29), the 3-D Mexican hat radial wavelet was used to fit a model. It was calculated that the root-mean-square-error (RMSE) of the model prediction over the test data set (points from 1001 to 3000) is 0.0889 for the traditional wavelet model. The value of RMSE with respect to the same test data set based on the proposed method, however, is only 0.0213, which is much smaller. This implies that for this example the new proposed WN may be advantageous over a conventional fixed grid WN, where only the total-variable-involved functional exponent were considered.

B. High-Dimensional Chaotic System

Consider the Mackey–Glass delay-differential equation [57]

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1 + x^{10}(t-\tau)} \tag{31}$$

where the time delay τ was chosen to be 30 in this example. This example was chosen to facilitate comparisons with other results [25], [58]. Setting the initial condition $x(t) = 0.9$ for $0 \le t \le \tau$, a Runge–Kutta integral algorithm was applied to calculate (31) with an integral step $\Delta t = 0.01$ and 6000 equi-spaced samples, $x(k)$, $(k = 1, 2, \ldots, 6000)$ were recorded with a sampling interval of $T = 0.06$ time units.

The recorded sequence was normalized into the unit interval $[0, 1]$ using the *a priori* knowledge $x_k \in [0.2, 1.4]$. Designate the normalized sequence still by $y(k)$. The 6000 points were then divided into two parts: the estimation set consisting of the first 500 points was used for WN training and the validation set consisting of the remaining 5500 points was used for model

tests. Following [59], the dimension of the recorded time series was assumed to be $n = 6$, and the significant variables were therefore chosen to be $\{y(k-1), y(k-2), \ldots, y(k-6)\}$. Similar to the previous example, the initial WN was chosen to be

$$\begin{aligned} y(k) &= f(y(k-1), y(k-2), \ldots, y(k-6)) \\ &= \sum_{p=1}^{6} f_p(y(k-p)) + \sum_{p=1}^{5}\sum_{q=2}^{6} f_{pq}(y(k-p), y(k-q)) \\ &\quad + \sum_{p=1}^{4}\sum_{q=2}^{5}\sum_{r=3}^{6} f_{pqr}(y(k-p), y(k-q), y(k-r)) \end{aligned} \tag{32}$$

where the 1-D, 2-D, and 3-D compactly supported Mexican hat radial WNs were used in this example to approximate the univariale functions f_p, the bivariate functions f_{pq}, and the tri-variate function f_{pqr}, respectively, with the coarsest resolutions $j_1 = j_2 = j_3 = 0$ and finest resolutions $J_1 = 3$, $J_2 = 1$ and $J_3 = 0$. A forward OLS-ERR algorithm was used to select significant model terms. The final identified model was found to be

$$\begin{aligned} y(k) = &-2.126\,564\,37 \times 10^{-3}\psi_{3,0}(y(k-1)) \\ &- 3.222\,047\,81 \times 10^{-2}\psi_{3,10}(y(k-1)) \\ &+ 1.636\,196\,75 \times 10^{-1}\psi_{3,-3}(y(k-3)) \\ &+ 1.154\,344\,12 \times 10^{-3}\psi_{3,2}(y(k-3)) \\ &- 6.932\,382\,61 \times 10^{-2}\psi_{1,3}(y(k-5)) \\ &+ 1.391\,115\,31 \times 10^{-3}\psi_{3,9}(y(k-6)) \\ &+ 8.653\,133\,32 \times 10^{-3}\psi_{2,4}(y(k-6)) \\ &+ 3.608\,562\,65 \times 10^{-2}\psi^{[2]}_{0;1,1}(y(k-1), y(k-2)) \\ &- 2.377\,188\,15 \times 10^{0}\psi^{[2]}_{0;-1,1}(y(k-1), y(k-3)) \\ &- 1.711\,817\,90 \times 10^{-1}\psi^{[2]}_{1;1,2}(y(k-1), y(k-5)) \\ &- 8.499\,305\,98 \times 10^{-2}\psi^{[2]}_{1;4,4}(y(k-1), y(k-6)) \\ &+ 9.930\,006\,74 \times 10^{-2}\psi^{[2]}_{1;1,2}(y(k-2), y(k-6)) \\ &+ 2.095\,854\,03 \times 10^{-1}\psi^{[2]}_{0;1,1}(y(k-3), y(k-4)) \\ &+ 4.947\,355\,53 \times 10^{-1}\psi^{[2]}_{0;-1,1}(y(k-3), y(k-5)) \end{aligned} \tag{33}$$

where $\psi_{j,k}(u) = 2^{j/2}\psi(2^j u - k)$ and $\psi^{[2]}_{j,k_1,k_2}(u_1, u_2) = 2^j\psi(2^j u_1 - k_1, 2^j u_2 - k_2)$ are the 1-D and 2-D compactly supported Mexican hat wavelets, $j, k, k_1, k_2 \in Z$.

Most of the results in the literature concern one-step-ahead predictions of the sampled time series. In this example, however, two-step-ahead predictions were considered and the predicted results were compared with previous studies [25], [58], where only one-step-ahead predictions were considered. To facilitate comparisons, a measurement index, the relative error [25], was used to measure the performance of the identified WN. This index is defined as

$$E_k = \frac{|x_k - \hat{x}_k|}{|x_k|} \tag{34}$$

where x_k and $\hat{x}_k$ are the measurements on the test set and associated two-step-ahead predictions, respectively.

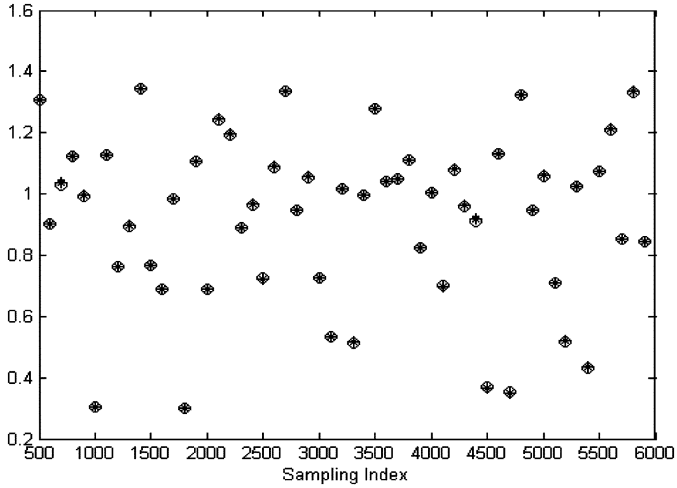

Fig. 3. Two-step-ahead predictions for the Mackey–Glass delay-differential (31) using the identified WN (33) over the validation set. The stars "$*$" indicate the measurements and the circles "o" indicate the predications. To allow a clear inspection, the data are plotted once every 100 points.

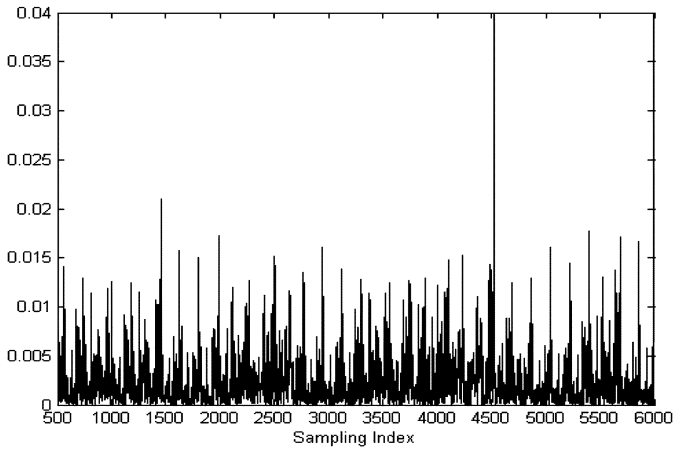

Fig. 4. Relative errors between the two-step-ahead predictions from the identified WN (33) and the measurements for the Mackey–Glass delay-differential (31) over the validation set.

The results of two-step-ahead predictions of the WN (33) were compared with the measurements and these are shown in Fig. 3, where the data are plotted once every 100 points to allow a clear inspection. The relative error E_k is shown in Fig. 4, which clearly indicates that the underlying dynamics have been captured by the identified WN (33). Notice that from Fig. 4 the result of two-step-ahead predictions of the WN (33) is by far better even than that of the one-step-ahead predictions provided by the WNs proposed in [25]. In fact, simulation results show that the relative error E_k with respect to the one-step-ahead predictions provided by the WN (33) are by far smaller than those with respect to the two-step-ahead predictions. The standard derivation over the test data set was calculated to be 0.0029 with respect to the two-step-ahead predictions of the WN (33), which is much smaller than 0.041 and is equivalent to 0.0016 given by [58], where the one-step-ahead predictions were considered. These results obviously show that the new WNs are more effective than conventional fixed grid WNs and are equivalent to adaptive WNs.

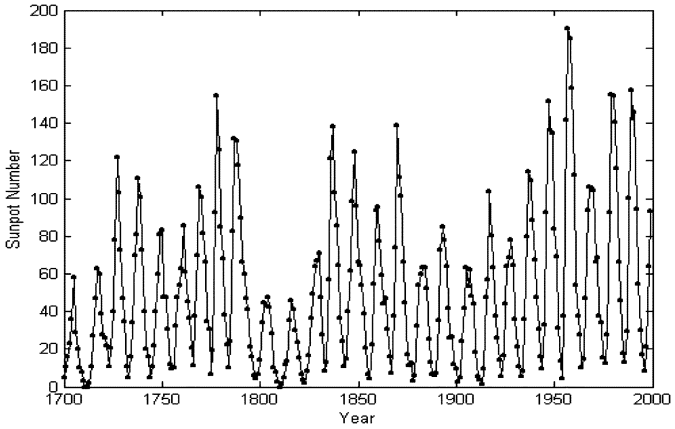

Fig. 5. Sunspot time series for the period from 1700 to 1999.

C. The Sunspot Time Series

The sunspot time series considered in this example consists of 300 annually recorded Wolf sunspots of the period from 1700 to 1999, see Fig. 5. The objective here is to identify a WN model to produce one-step-ahead predictions for the sunspot data set. Again, the original measurements $y(t)(1 \leq t \leq 300\}$ were initially normalized into the unit interval $[0, 1]$ using the information $0 \leq y(t) \leq 200$. Designate the normalized sequence by $y(t)$. The data set was separated into two parts: the training set consisted of 250 data points corresponding to the period 1700–1949, and the test set consisted of 50 data points corresponding to the period 1950–1999.

Following [56], the model order was chosen to be $n = 9$ here, and the most significant variables were chosen to be $y(t-1), y(t-2)$ and $y(t-9)$. The initial WN model was therefore chosen to be

$$\begin{aligned} y(t) &= f(y(t-1), y(t-2), \ldots, y(t-9)) \\ &= \sum_{p=1}^{9} f_p(x_p(t)) + \sum_{p=1}^{2}\sum_{q=2}^{3} f_{pq}(z_p(t), z_q(t)) \\ &\quad + f_{123}(z_1(t), z_2(t), z_3(t)) \end{aligned} \tag{35}$$

where $x_p(t) = y(t-p)$ for $p = 1, 2, \ldots, 9$, $z_k(t) = y(t-k)$ for $k = 1,2$, and $z_3(t) = y(t-9)$. The 1-D, 2-D, and 3-D compactly supported Gaussian radial WNs were used in this example to approximate the univariate functions f_p, the bivariate functions f_{pq}, and the tri-variate function f_{123}, respectively, with the coarsest resolutions $j_1 = j_2 = j_3 = 0$ and finest resolutions $J_1 = 2J_2 = J_3 = 0$. A forward OLS-ERR algorithm [35], [36] was used to select significant model terms. The final identified model was found to be

$$y(t) = \sum_{k=1}^{24} \theta_k p_k(t) \tag{36}$$

where $p_k(t)$ are the wavelet terms formed by compactly supported Gaussian wavelets. The identified wavelet terms, the corresponding parameters, and the associated ERRs are listed in Table I. Roughly speaking, the values of the ERRs provide an index indicting the contribution made by the corresponding model term to a signal of interest, and in general, the larger a ERR value is, the more significant the corresponding model term is for representing a given signal. For details about the meaning of ERR, see [25] and [36]. The result of the one-step-ahead predictions based on the WN (36) over the test set is shown in Fig. 6 (the dashed-star line), which clearly shows that the identified model provides an excellent representation for the sunspot time series.

TABLE I
WAVELET TERMS, PARAMETERS, AND ASSOCIATED ERROR REDUCTION RATIOS FOR THE SUNSPOT TIME SERIES

No.	Terms $p_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	p_1	1.51298827e+000	9.24528866e+001
2	$\psi_{2,3}(x_2)$	5.41812968e-003	1.19512777e+000
3	$\psi_{2,-3}(x_2)$	5.37704026e-001	8.34773530e-001
4	$\psi^{[2]}_{0;0,1}(x_1, x_9)$	-2.65654782e+000	4.17087285e-001
5	$\psi_{2;-3}(x_1)$	-2.27744202e-001	3.15221310e-001
6	$\psi_{2,7}(x_1)$	-6.67540058e+000	2.82572747e-001
7	$\psi^{[2]}_{0;-2,0}(x_2, x_9)$	6.89573412e+000	1.01940451e-001
8	$\psi^{[2]}_{0;-2,0}(x_1, x_9)$	-8.10206178e+000	1.37984742e-001
9	$\psi^{[2]}_{0;3,-2}(x_1, x_9)$	2.04730652e+001	1.11760613e-001
10	$\psi_{2,0}(x_4)$	6.65007554e-002	1.24850577e-001
11	$\psi^{[2]}_{0;2,2}(x_1, x_9)$	1.45025760e+000	6.36718843e-002
12	$\psi_{2,1}(x_3)$	-4.43226358e-002	8.08322664e-002
13	$\psi_{2,2}(x_7)$	-6.81401148e-002	1.00811211e-001
14	$\psi_{1,5}(x_3)$	7.99121974e+000	5.07384329e-002
15	$\psi_{1,4}(x_2)$	-9.59979407e-001	6.99589742e-002
16	$\psi_{2,2}(x_4)$	5.19552579e-002	6.77180550e-002
17	$\psi_{2,3}(x_8)$	4.47442151e-002	3.71432968e-002
18	$\psi_{2,4}(x_7)$	-2.09739189e-001	4.93876251e-002
19	$\psi_{2,-3}(x_8)$	1.31932809e+000	6.76831534e-002
20	$\psi_{2,-3}(x_6)$	4.02514721e+000	3.82428793e-002
21	$\psi_{2,-2}(x_6)$	-4.46511248e-001	5.79397319e-002
22	$\psi_{2,4}(x_5)$	5.92505836e-002	6.85987322e-002
23	$\psi_{2,-2}(x_9)$	-1.43930547e-001	3.48667548e-002
24	$\psi_{2,7}(x_9)$	1.95271113e+000	1.73543459e-002

$x_j = y(t-j)$ for j=1,2, …, 9,

$p_1 = \psi^{[3]}_{0;0,-1,2}(x_1, x_2, x_9)$,

$\psi_{j,k}(u) = 2^{j/2}\psi(2^j u - k)$,

$\psi^{[2]}_{j;k_1,k_2}(u_1, u_2) = 2^j \psi^{[2]}(2^j u_1 - k_1, 2^j u_2 - k_2)$,

$\psi^{[3]}_{j;k_1,k_2,k_3}(u_1, u_2, u_3)$
$= 2^{3j/2}\psi^{[3]}(2^j u_1 - k_1, 2^j u_2 - k_2, 2^j u_3 - k_3)$,

$\psi_{j,k}(u)$, $\psi^{[2]}_{j;k_1,k_2}(u_1, u_2)$ and $\psi^{[3]}_{j;k_1,k_2,k_3}(u_1, u_2, u_3)$ are the compactly supported Gaussian wavelets.

In order to compare the predicted result of the WN with other work [60], the following index, the mean-square-error on the test set, was used to measure the performance of the identified WN

$$\bar{E} = \frac{\sum_{k=1}^{N_{\text{test}}} |x_k - \hat{x}_k|^2}{\sum_{k=1}^{N_{\text{test}}} |x_k - \bar{x}|^2} \tag{37}$$

where N_{test} is the length of the test set, x_k and $\hat{x}_k$ are the measurements over the data set and associated one-step-ahead pre-

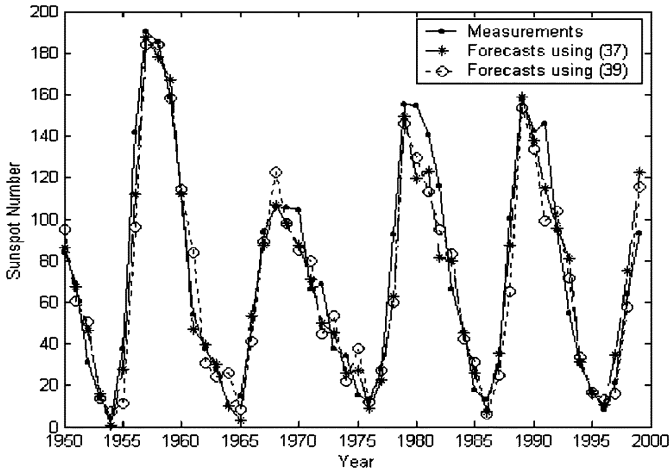

Fig. 6. One-step-ahead predictions for the sunspot time series based the WNs (36) and (38) over the test set. The point-solid line indicates the measurements, the dashed-star line indicates the predictions from (36), and the dotted-circle line indicates the predications from (38).

dictions, respectively, and $\bar{x} = (1/N_{\text{test}}) \sum_{k=1}^{N_{\text{test}}} x_k$. It was calculated $\bar{E} = 0.0651$ for the identified WN (36) that is smaller than 0.076 (for the period of years 1921–1954) and 0.23 (for the period of years 1955–1979) which are given by a wavelet decomposition model proposed in [60].

An important point revealed by Table I is that the three variables $y(t-1)$, $y(t-2)$ and $y(t-9)$ are far more significant than the other variables. This is consistent with the result given in [55]. In fact, the sunspot time series can be satisfactory described using a WN with respect to only these three significant variables. This model is given in the following:

$$\begin{aligned} y(t) = & -0.127\,59\psi_{2,-3}(y(t-1)) \\ & -9.414\,69\psi_{2,7}(y(t-1)) \\ & +0.055\,04\psi_{2,3}(y(t-2)) \\ & +1.308\,97\psi_{2,-3}(y(t-2)) \\ & -3.679\,94\psi^{[2]}_{0;-2,0}(y(t-1), y(t-9)) \\ & -1.577\,33\psi^{[2]}_{0;0,1}(y(t-1), y(t-9)) \\ & +6.474\,37\psi^{[2]}_{0;3,-2}(y(t-1), y(t-9)) \\ & +4.225\,07\psi^{[2]}_{0;-2,0}(y(t-2), y(t-9)) \\ & -2.412\,46\psi^{[3]}_{0;0,-1,2}(y(t-1), y(t-2), y(t-9)). \end{aligned} \tag{38}$$

The one-step-ahead predictions from the WN (38) over the test set is shown in Fig. 6 (the dotted-circle line), where the normalized error $\bar{E}$ was calculated to be 0.1044, which is still very small.

VI. Conclusion

A new class of WNs has been introduced for nonlinear system identification. The main advantage of the new identification approach compared with existing WNs, is that the new WNs are more practical and can be applied to problems in medium and high dimensions. This property arises due to the fact that the structure of the new WNs are based on ANOVA expansions, for which the high-dimensional subfunctions (submodels) can often be neglected for many nonlinear systems.

It has been noted that a conventional WN always includes the total-variables-involved wavelet terms, even though this is not necessary for most systems in the real world. In addition, a model that includes only high-order terms is liable to produce a deleterious effect on the output behavior of the model which can induce spurious dynamics. The new WNs avoid most of these problems by decomposing a multidimensional function into a number of low-dimensional submodels.

In theory, many types of wavelets can be used to approximate the low-dimensional submodels by a scheme of taking tensor products or adopting radial functions. In network training, however, it is often preferable to use a wavelet that is compactly supported, since the number of compactly supported wavelets at a given resolution scale can be determined beforehand and, thus, the total number of candidate wavelet terms involved in the network becomes known. Radial wavelets are not compactly supported but rapidly vanishing. It is therefore reasonable to truncate a radial wavelet to make it quasi-supported, this can then be used as a normal compactly supported wavelet to implement a WN. Most radial wavelets including the Gaussian and Mexican hat wavelets are easy to calculate with a very small computational load and can therefore be chosen to implement the WN. Other nonradial wavelets, which are either compactly supported or not, can also be used if there is strong evidence that these wavelets can easily be used to implement a WN.

A WN may involve a great number of wavelet terms for a high-dimensional system. However, in most cases many of the model terms are redundant and only a small number of significant terms are necessary to describe a given nonlinear system with a given accuracy. In the present study, an efficient term detection algorithm was employed to train the new WNs to yield parsimonious models.

In summary, the new WNs appear to be advantageous compared to conventional wavelet modeling schemes and provide an effective approach for nonlinear system identification. The results obtained from the bench test examples demonstrate the effectiveness of the new identification procedure.

Appendix
Forward OLS Algorithm and the ERR

The OLS algorithm [35], [36] was initially introduced to select the most significant model terms and estimate the model parameters simultaneously for all linear-in-the-parameter models. Consider the linear-in-the-parameters model (21), where the regression matrix $P = [p_1, p_2, \ldots, p_M]$ with, $p_i = [p_i(1), p_i(2), \ldots, p_i(N)]^T$, 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

$$P = WA \tag{39}$$

where A is an $M \times M$ unit upper triangular matrix and W is an $N \times M$ matrix with orthogonal columns $w_1, w_2, \ldots, w_M$ in the sense that $W^T W = D = \text{diag}[d_1, d_2, \ldots, d_M]$ with $d_m = w_m^T w_m$. Model (21) can then be expressed as

$$Y = (PA^{-1})(A\Theta) + \Xi = WG + \Xi \tag{40}$$

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

$$g_i = \frac{Y^T w_i}{w_i^T w_i}, \quad i = 1, 2, \ldots, M. \tag{41}$$

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 (21) 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 ($M_0 \ll M$). Detection of the significant model terms can be achieved using the OLS procedures described in the following.

Assume that the residual signal $\varepsilon(t)$ in the model (21) is uncorrelated with the past outputs of the system, then the output variance can be expressed as

$$\frac{1}{N} Y^T Y = \frac{1}{N} \sum_{i=1}^{M} g_i^2 w_i^T w_i + \frac{1}{N} \Xi^T \Xi. \tag{42}$$

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 ith ERR_i, introduced by p_i, can be defined as

$$\begin{aligned} \mathrm{ERR}_i &= \frac{g_i^2 (w_i^T w_i)}{Y^T Y} \times 100\% \\ &= \frac{(Y^T w_i)^2}{(Y^T Y)(w_i^T w_i)} \times 100\%, \quad i = 1, 2, \ldots, M. \end{aligned} \tag{43}$$

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 selection procedure can be terminated at the M_0th step ($M_0 \le M$) when $1 - \sum_{i=1}^{M_0} \mathrm{ERR}_i < \rho$, where ρ is a desired error tolerance, or cutoff value, 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 $\{p_i\}_{i=1}^{M}$

$$y(t) = \sum_{i=1}^{M_0} g_i^0 w_i^0(t) + \varepsilon(t) \tag{44}$$

which is equivalent to

$$y(t) = \sum_{i=1}^{M_0} \theta_{\ell_i} p_{\ell_i}(x(t)) + \varepsilon(t) \tag{45}$$

where the parameters $[\theta_{\ell_1}, \theta_{\ell_2}, \ldots, \theta_{\ell_{M_0}}]^T$ can be calculated in the selection procedure. Note that, since most significant model terms can be selected in a forward-regression manner step by step, or a term at a time, the assumption that the regression matrix P is full rank in columns becomes unnecessary [56].

REFERENCES

[1] C. K. Chui, *An Introduction to Wavelets*. New York: Academic, 1992.
[2] I. Daubechies, *Ten Lectures on Wavelets*. Philaelphia, PA: SIAM, 1992.
[3] Y. Meyer, *Wavelets: Algorithms and Applications*. Philaelphia, PA: SIAM, 1993.
[4] S. Chen, S. A. Billings, and P. M. Grant, "Nonlinear system identification using neural networks," *Int. J. Control*, vol. 51, pp. 1191–1214, Jun. 1990.
[5] S. Chen and S. A. Billings, "Neural networks for nonlinear system modeling and identification," *Int. J. Control*, vol. 56, pp. 319–346, Aug. 1992.
[6] S. Chen, S. A. Billings, and P. M. Grant, "Recursive hybrid algorithm for nonlinear system identification using radial basis function networks," *Int. J. Control*, vol. 55, pp. 1051–1070, May 1992.
[7] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 252–262, Mar. 1991.
[8] S. A. Billings, H. B. Jamaluddin, and S. Chen, "Properties of neural networks with applications to modeling nonlinear dynamic systems," *Int. J. Control*, vol. 55, pp. 193–224, Jan. 1992.
[9] S. A. Billings and S. Chen, "The determination of multivariable nonlinear models for dynamic systems using neural networks," in *Neural Network Systems Techniques and Applications*, C. T. Leondes, Ed. New York: Academic, 1998, pp. 231–278.
[10] P. S. Sastry, G. Santharam, and K. P. Unnikrishnan, "Memory neuron networks for identification and control of dynamical systems," *IEEE Trans. Neural Netw.*, vol. 5, no. 3, pp. 306–319, May 1994.
[11] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.
[12] T. N. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 7, no. 6, pp. 1329–1338, Nov. 1996.
[13] K. S. Narendra and S. Mukhopadhyay, "Adaptive control using neural networks and approximate models," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 475–485, May 1997.
[14] N. B. Karayiannis and M. M. Randolph, "On the construction and training of reformulated radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 835–846, Jul. 2003.
[15] I. Rivals and L. Personnaz, "Neural-network construction and selection in nonlinear modeling," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 804–819, Jul. 2003.
[16] R. J. Wai and H. H. Chang, "Backstepping wavelet neural network control for indirect field-oriented induction motor drive," *IEEE Trans. Neural Netw.*, vol. 15, no. 2, pp. 367–382, Mar. 2004.
[17] H. H. Szu, B. Telfer, and S. Kadambe, "Neural network adaptive wavelets for signal representation and classification," *Opt. Eng.*, vol. 31, pp. 1907–1916, Sep. 1992.
[18] Q. H. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Netw.*, vol. 3, no. 6, pp. 889–898, Nov. 1992.
[19] Y. C. Pati and P. S. Krishnaprasad, "Analysis and synthesis of feedforward neural networks using discrete affine wavelet transforms," *IEEE Trans. Neural Netw.*, vol. 4, no. 1, pp. 73–85, Jan. 1993.
[20] J. G. Daugman, "Complete discrete 2-D gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp. 1169–1179, Jul. 1988.
[21] H. Dickhaus and H. Heinrich, "Classfying biosignals with wavlet networks," *IEEE Eng. Med. Biol. Mag.*, vol. 15, no. 5, pp. 103–111, Sep. 1996.
[22] S. Pittner, S. V. Kamarthi, and Q. G. Gao, "Wavelet networks for sensor signal classification in flank wear assessment," *J. Intell. Manufact.*, vol. 9, pp. 315–322, Aug. 1990.
[23] K. W. Wong and A. C.-S. Leung, "On-line successive synthesis of wavelet networks," *Neural Process. Lett.*, vol. 7, pp. 91–100, Apr. 1998.
[24] E. A. Rying, G. L. Bilbro, and J. C. Lu, "Focused local learning with wavelet neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 304–319, Mar. 2002.
[25] L. Y. Cao, Y. G. Hong, H. P. Fang, and G. W. He, "Predicting chaotic time series with wavelet networks," *Phys. D*, vol. 85, pp. 225–238, Jul. 1995.
[26] D. Allingham, M. West, and A. Mees, "Wavelet reconstruction of nonlinear dynamics," *Int. J. Bifurcat. Chaos*, vol. 8, pp. 2191–2201, Nov. 1998.

[27] Y. Oussar and G. Dreyfus, "Initialization by selection for wavelet network training," *Neurocomput.*, vol. 34, pp. 131–143, Sep. 2000.
[28] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, pp. 1–67, Mar. 1991.
[29] Z. H. Chen, "Fitting multivariate regression functions by interaction spline models," *J. Roy. Stat. Soc. B Met.*, vol. 55, pp. 473–491, 1993.
[30] J. Zhang, G. G. Walter, Y. Miao, and W. N. W. Lee, "Wavelet neural networks for function learning," *IEEE Trans. Signal Process.*, vol. 43, no. 6, pp. 1485–1497, Jun. 1995.
[31] Q. H. Zhang, "Using wavelet network in nonparametric estimation," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 227–236, Mar. 1997.
[32] S. Ferrari, M. Maggioni, and N. A. Borghese, "Multoresolution approximation with hierarchical radial basis functions networks," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 178–188, Jan. 2004.
[33] D. Coca and S. A. Billings, "Continuous-time system identification for linear and nonlinear system identification using wavelet decompositions," *Int. J. Bifurcat. Chaos*, vol. 7, pp. 87–96, Jan. 1997.
[34] S. A. Billings and D. Coca, "Discrete wavelet models for identification and qualitative analysis of chaotic systems," *Int. J. Bifurcat. Chaos*, vol. 9, pp. 1263–1284, Jul. 1999.
[35] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO nonlinear systems suing a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, pp. 2157–2189, Jun. 1989.
[36] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to nonlinear system identification," *Int. J. Control*, vol. 50, pp. 1873–1896, Nov. 1989.
[37] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least-squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 302–309, Mar. 1991.
[38] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for nonlinear systems-part I: Deterministic nonlinear systems; part II: Stochastic nonlinear systems," *Int. J. Control*, vol. 41, pp. 303–344, 1985.
[39] R. K. Pearson, "Nonlinear input/output modeling," *J. Process Contr.*, vol. 5, pp. 197–211, Aug. 1995.
[40] ——, "Nonlinear process identification," in *Nonlinear Process Control*, M. A. Henson and D. E. Seborg, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1997, pp. 11–110.
[41] L. A. Aguirre, "Application of global models in the identification, analysis and control of nonlinear dynamics and chaos," Ph.D. dissertation, Dept. Autom. Control Syst. Eng., Univ. Sheffield, Sheffield, U.K., 1994.
[42] N. Chiras, "Linear and nonlinear modeling of gas turbine engines," Ph.D. dissertation, School Electron., Univ. Glamorgan, Wales, U.K., 2002.
[43] R. K. Pearson, *Discrete-Time Dynamic Models*. Oxford, U.K.: Oxford Univ. Press, 1999.
[44] G. Y. Li, C. Rosenthal, and H. Rabits, "High dimensional model representations," *J. Phys. Chem. A*, vol. 105, pp. 7765–7777, Aug. 2001.
[45] H. L. Wei and S. A. Billings, "A unified wavelet-based modeling framework for nonlinear system identification: The WANARX model structure," *Int. J. Control*, vol. 77, pp. 351–366, Mar. 2004.
[46] J. Gonzalez, I. Rojas, J. Ortega, H. Pomares, F. J. Fernandez, and A. F. Diaz, "Multiobjective evolutionary optimization of the size, shape, and position parameters of radial basis functions networks for function approximation," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1478–1495, Nov. 2003.
[47] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, Jul. 1989.
[48] ——, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.
[49] L. X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximations, and orthogonal least squares learning," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 807–814, Sep. 1992.
[50] X. Hong and C. J. Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 435–439, Mar. 2001.
[51] X. Hong, P. M. Sharkey, and K. Warwick, "A robust nonlinear identification algorithm using PRESS statistic and forward regression," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 454–458, Mar. 2003.
[52] S. A. Billings and H. L. Wei, "The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions," Int. J. Syst. Sci., to be published.
[53] S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, vol. 44, pp. 235–244, Jul. 1986.
[54] S. A. Billings and Q. M. Zhu, "Nonlinear model validation using correlation tests," *Int. J. Control*, vol. 60, pp. 1107–1120, Dec. 1994.
[55] D. Coca, "A class of wavelet multiresolution decompositions for nonlinear system identification and signal processing," Ph.D., Dept. Autom. Control Syst. Eng., Univ. Sheffield, Sheffield, U.K., 1996.
[56] H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for nonlinear system identification," *Int. J. Control*, vol. 77, pp. 86–110, Jan. 2004.
[57] M. C. Mackey and L. Glass, "Oscillation and chaos in physiological control systems," *Science*, vol. 197, pp. 287–289, Jul. 1977.
[58] R. Bone, M. Crucianu, and J.-P. A. de Beauville, "Learning long-term dependences by the selective additional time-delayed connetions to recurrent neural networks," *Neurocomput.*, vol. 48, pp. 251–266, Oct. 2002.
[59] M. Casdagli, "Nonlinear prediction of chaotic time series," *Phys. D*, vol. 35, pp. 335–356, May 1989.
[60] S. Soltani, "On the use of wavelet decomposition for time series prediction," *Neurocomput.*, vol. 48, pp. 267–277, Oct. 2002.

Stephen A. Billings received the B.Eng. degree in electrical engineering (first class honors) from the University of Liverpool, Liverpool, U.K., in 1972, the Ph.D. degree in control systems engineering from the University of Sheffield, Sheffield, U.K., in 1976, and the D.Eng. degree from the University of Liverpool in 1990.

He was appointed as Professor in the Department of Automatic Control and Systems Engineering, University of Sheffield, in 1990 and leads the Signal Processing and Complex Systems research group. His research interests include system identification and information processing for nonlinear systems, narmax methods, model validation, prediction, spectral analysis, adaptive systems, nonlinear systems analysis and design, neural networks, wavelets, fractals, machine vision, cellular automata, spatio–temporal systems, fMRI and optical imagery of the brain, metabolic systems engineering, systems biology, and related fields.

Dr. Billings is a Chartered Engineer (C.Eng.), Chartered Mathematician (C.Math.), a Fellow of the Institution of Electrical Engineers (IEEE-U.K.) and a Fellow of the Institute of Mathematics and Its Applications.

Hua-Liang Wei received the B.Sc. degree in mathematics from Liaocheng University, Liaocheng, China, in 1989, the M.Sc. degree in automatic control from the Beijing Institute of Technology, Beijing, China, in 1992, and the Ph.D. degree in signal processing and control engineering from the University of Sheffield, Sheffield, U.K., in 2004.

He previously held academic appointments at the Beijing Institute of Technology, Beijing, China, from 1992 to 2000. He joined the Department of Automatic Control and Systems Engineering, University of Sheffield, initially as a Research Associate in 2004. He has been a Research Fellow working with the Signal Processing and Complex Systems research group since 2005. His recent research interests include identification and signal processing of nonlinear systems, NARMAX methodology and its applications, analysis of nonlinear systems in the frequency domain, wavelets and neural networks in nonlinear system identification, regression analysis, and data-driven modeling.