A New Class of Wavelet Networks for Nonlinear System Identification

S.A. Billings and H.L. Wei

Research Report No. 857

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

March 2004

A New Class of Wavelet Networks for Nonlinear System Identification

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

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 detection 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.

Keywords: Nonlinear system identification; NARX models; orthogonal least squares; wavelet networks

1. Introduction

Wavelet theory (Chui, 1992; Daubechies, 1992; Meyer, 1993) has been extensively studied in recent years and has been widely applied in various areas throughout science and engineering. Dynamical system modelling 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 (Chen *et al.*, 1990, 1991, 1992; Narendra and Parthasarathy, 1990; Billings *et al.*, 1992; Chen and Billings, 1992; Haykin, 1994; Sastry et al., 1994; Lin, et al., 1996; Narendra and Mukhopadhyay, 1997). The idea of combining wavelets with neural networks has led to the development of wavelet networks, 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 wavelet networks were popularized by the work of Szu *et al.* (1992), Zhang and Benveniste (1992), and Pati and Krishnaprasad (1993) etc., the origin of wavelet networks can be traced back to the earlier work of Daugman (1988), 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 wavelet networks can therefore be catalogued into two types:

- *Adaptive wavelet networks*, 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 wavelet networks*, 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 wavelet network, the positions and dilations of the wavelets are fixed (pre-determined) 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 wavelet networks was introduced by Zhang and Benveniste (1992) 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 wavelet networks have been successfully applied to nonlinear static function approximation and classification (Szu *et al.*, 1992; Pittner *et al.*, 1998; Wong and Leung, 1998), and dynamical system modelling (Cao *et al.*, 1995; Alligham *et al.*, 1998). Clearly, to train an adaptive wavelet network, 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 (Oussar and Dreyfus, 2000). Another problem that needs to be considered for training an adaptive wavelet network is how to determine the initial number of wavelets associated with the network. These drawbacks often limit the application of adaptive wavelet networks to low-dimensional problems.

Unlike adaptive wavelet networks, in a fixed grid wavelet network, 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 wavelet network is now a linear-in-the-parameters regression, which can then be solved using least squares techniques. As will be discussed in Section 3.4, the number of candidate wavelet terms in a fixed grid wavelet network often increases dramatically with the model order. As a consequence, fixed grid wavelet networks are often limited to low-dimensions.

Inspired by the well known ANOVA (analysis of variance) expansions (Friedman, 1991; Chen, 1993), a new class of fixed grid wavelet networks is introduced in the present study for nonlinear system identification. In the new wavelet networks, 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 wavelet networks are therefore in structure different from either the existing wavelet networks (Zhang and Benveniste, 1992; Cao *et al.*, 1995; Zhang *et al.*, 1995; Zhang, 1997; Alligham *et al.*, 1998) or wavelet mutiresolution models (Coca and Billings, 1997; Billings and Coca, 1999). An efficient model term detection approach based on a forward orthogonal least squares (OLS) algorithm, along with the error reduction ratio (ERR) criterion (Billings *et al.*, 1989; Chen *et al.*, 1989; Chen *et al.*, 1991) is applied to solve the linear-in-the-parameters problem in the present study.

2. Representations of Nonlinear Dynamical Systems

A wide range of nonlinear systems can be represented using the NARX (N*onlinear AutoRegressive with eXogenous* inputs) model (Leontaritis and Billings, 1985; Pearson, 1995, 1997). Taking SISO systems as an example, this can be expressed by the following nonlinear difference equation

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

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 realise the representation (1) including polynomials (Chen *et al.*, 1989; Aguirre, 1994; Chiras, 2002), neural networks (Chen *et al.*, 1990, 1991, 1992; Billings *et al.*, 1992; Chen and Billings, 1992) and other complex models, see for example the book by Pearson (1999). 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 analysis of variance (ANOVA) expansions (Friedman, 1991; Chen, 1993)

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

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

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

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 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 is often sufficient to 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 (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. 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 (2). In the present study, however, 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 model in Wei and Billings (2004), where multiresolution wavelet decompositions were employed and a class of compactly supported wavelets were considered.

3. Wavelet Networks—Truncated Wavelet Decompositions

This section briefly reviews some results on wavelet decompositions and wavelet networks which are relevant to the present work. For more details about these results, see for example the work of Mallat (1989, 1998), Chui (1992), Daubechies (1992), Zhang and Benveniste (1992), Meyer (1993), Zhang (1997). In the following, it is assumed that the independent variable x of a function $f \in L^2(\mathbb{R})$ of interest is defined in the unit interval [0,1]. In addition, for the sake of simplicity, one-dimensional wavelets are considered as an example to illustrate related concepts.

3.1 Wavelet decompositions

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

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

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 $\psi_{(a_t,b_t)}$ are preserved to be the same as that of ψ . Rearrange the elements of Ω so that

$$\Omega = \{\psi_t : t \in \Gamma\} \tag{5}$$

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 $f \in L^2(\mathbb{R})$ can be recovered from the mother wavelet ψ in the sense that (Chui, 1992; Daubechies, 1992; Zhang and Benveniste, 1992)

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

or

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

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

In practical applications the decomposition (7) is often discretized in both the scale and dilation parameters for computational efficiency. Based on this discretization, 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 $j,k\in\mathbb{Z}$ ($\mathbb{Z}$ is the set of all integers). Other non-dyadic ways of discretization are also available. For the dyadic lattice case, (7) becomes

$$f(x)=\sum_{j}\sum_{k} c_{j,k}\psi_{j,k}(x) \tag{8}$$

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

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* (Chui, 1992). 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 functions is given to a wavelet frame by relaxing the orthogonality.

3.2 Wavelet networks

In practical applications for either static function learning or dynamical system modelling, 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. Wavelet networks are in effect nothing but a truncated wavelet decomposition. Taking the decomposition (8) as an example, an approximation to a function $f\in L^2(\mathbb{R})$ using the truncated wavelet decomposition with the coarsest resolution j_0 and the finest resolution $j_{\max}$ can be expressed as below:

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

where K_j ($j=j_0, j_0+1,\cdots, j_{\max}$) are subsets of $\mathbb{Z}$ 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

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

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,..., *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 wavelet networks*, 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 4.

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

3.3 Extending to high-dimensions

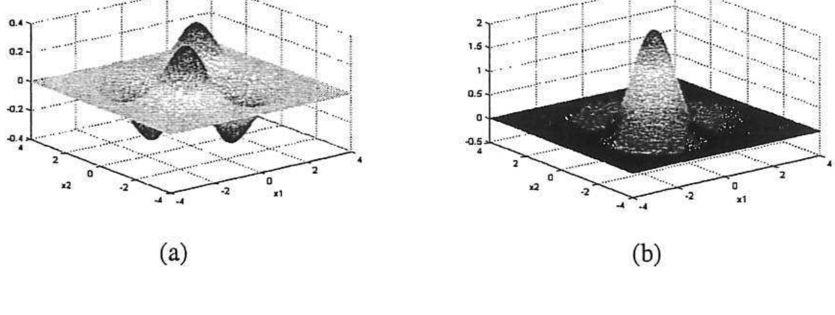
The results for one-dimensional case described above can be extended to high-dimensions. One commonly used approach is to generate separable wavelets by the tensor product of several one-dimensional wavelet functions. (Mallat, 1989; Zhang and Benveniste, 1992). For example, an *n*-dimensional wavelet $\psi^{[n]}: \mathbb{R}^n \mapsto \mathbb{R}$ can be constructed using a scalar wavelet ψ as follows

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

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

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

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 wavelet networks. The two-dimensional Gaussian and Mexican hat wavelets are shown in Figure 1.



(a)

(b)

Figure 1 The 2-D Gaussian and Marr mother wavelets. (a) Gaussian wavelet; (b) Marr wavelet.

3.4 Limitations of existing wavelet networks

It has been found that most exiting wavelet networks are limited to handling problems in low-dimensional space due to the difficulty of the so called *curse-of-dimensionality*. The following discussion will suffice the illustration that exiting wavelet networks are not suitable for high-dimensional problems.

Assume that a function $f \in L^2(\mathbb{R}^n)$ 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 3.3, this scalar wavelet can be used to generate an *n*-dimensional wavelet $\psi^{[n]} : \mathbb{R}^n \mapsto \mathbb{R}$ by (12). This multi-dimensional wavelet $\psi^{[n]}$ can then be used to approximate the *n*-dimensional function $f \in L^2(\mathbb{R}^n)$ using the wavelet network (9) as below:

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

where $k = [k_1, k_2 \cdots, k_n]^T \in \mathbb{Z}^n$ is an *n*-dimensional index. Noting that $x_i \in [0,1]$ for *i*=1,2,..., *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 $n_{term} = s^n$, where $s = 2^j + s_2 - s_1 - 2$. Setting *n*=5 and $s_2 - s_1$=5, this number will be n_{term} = 4^5, 5^5, 7^5, and 11^5 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_{term} = 4^{10}, 5^{10}, 7^{10} and 11^{10} for *j*=0,1,2 and 3, respectively. This implies that the total number of candidate wavelet terms involved in the wavelet network is very large even for some low resolution levels ($j \le 3$). This means that the computation task for a medium or high-dimensional wavelet network is too heavy to be implemented in a normal computer. Thus, it can be concluded that high-dimensional wavelet networks will be very difficult if not impossible to implement via 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 wavelet networks, where the wavelets are chosen to be radial wavelets, are also prohibited from high-dimensional problems by the above mentioned limitations. In addition, most existing radial wavelet networks 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 wavelet networks described above suggest that new wavelet networks need to be constructed to bypass the curse-of-dimensionality to enable the networks to handle high-dimensional problems.

4. A New Class of Wavelet Networks

The structure of the new wavelet networks 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.

4.1 The structure of the new wavelet networks

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

$$f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), \cdots, x_{i_m}(t)) = \sum_{j=j_m}^{J_m} \sum_{k_1 \in K_j} \cdots \sum_{k_m \in K_j} c_{j;k_1,\cdots,k_m} \psi^{[m]}_{j;k_1,\cdots,k_m}(x_{i_1}(t), \cdots, x_{i_m}(t)) \tag{15}$$

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

$$\begin{aligned} f_{pq}(x_p(t), x_q(t)) &= \sum_{j=j_2}^{J_2} \sum_{k_1} \sum_{k_2} c_{j;k_1,k_2} \psi^{[2]}_{j;k_1,k_2}(x_p(t), x_q(t)) \\ &= \sum_{j=j_2}^{J_2} \sum_{k_1} \sum_{k_2} c_{j;k_1,k_2} 2^j \{2 - [2^j x_p(t) - k_1]^2 - [2^j x_q(t) - k_2]^2\} e^{-\frac{1}{2}\{[2^j x_p(t) - k_1]^2 + [2^j x_q(t) - k_2]^2\}} \end{aligned} \tag{16}$$

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

By expanding each functional component in (2) using a radial wavelet network (15), a nonlinear wavelet network 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.

4.2 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 (Wei and Billings, 2004).

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 4th order B-spline wavelet (Chui, 1992) is [0,7]. At a resolution scale *j*, the variation range for the translation parameter *k* is $-6 \le k \le 2^j - 1$. The number of total candidate wavelet terns at different resolution scales in a wavelet network can then be determined.

Most radial wavelets are not compactly supported but rapidly vanishing. The variation range for the translation parameter *k* at a given scale level *j* can be determined as follows. Taking the one and two-dimensional Gaussian wavelets (13) as an example, it can be found that

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

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

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

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

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

The compactly supported Gaussian wavelets can be defined in the same way. The support for 3-dimensional Gaussian and Mexican wavelet can be defined as $S^{[3]} = [-3,3] \times [-3,3] \times [-3,3]$.

For the scalar Gaussian or Mexican hat wavelet, given a resolution scale *j*, since $\left|2^j x - k\right| \le 4$ and $0 \le x \le 1$, the choice for the translation parameter *k* should satisfy $-3 \le k \le 2^j + 3$. This means that the number of candidate one-dimensional 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.

4.3 Significant term detection

Assume that *M* candidate wavelet terms are involved in a wavelet network. The wavelet network 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,...,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 Eq. (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 $M_0 << M$), 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 orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) criterion, which was originally introduced to determine which terms should be included in a model (Billings *et al.*, 1989; Chen *et al.*, 1989). This approach has been extensively studied and widely applied in nonlinear system identification, see for example, the papers by Chen *et al.* (1991), Wang and Mendel (1992), Zhang (1997), Hong and Harris (2001), Billings and Wei (2003). The forward OLS algorithm involves a stepwise orthogonalization of the regressors and a forward selection of the relevant terms in (21) based on the error reduction ratio (ERR) (Chen *et al.*, 1989).

4.4 A procedure to implement the new wavelet networks

Two schemes can be adopted to implement the new wavelet network. 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 wavelet network is over-completed. The aim of the estimation procedure is to select the most significant wavelet terms from the deterministic over-completed library, and the selected model terms can often describe a given system well. Another scheme starts from a low-order submodel, where the library of wavelet basis functions (wavelet terms) used to construct a wavelet network may or may not be completed. The estimation procedure then selects the most significant wavelet terms from the given library. If model tests suggest that the selected wavelet terms cannot adequately describe a given system over the range of interests, higher dimensional wavelet terms should then be added to the wavelet network (library). Significant terms are then re-selected from the new library. This procedure may repeat several times until a satisfactory model is obtained. These two identification procedures to implement the wavelet network are summarized below.

4.4.1 Implement a wavelet network starting from an over-constructed model

This identification procedure contains in general 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 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:

(*i*) Select initial values for n_y and n_u.

(*ii*) Select the significant variables from all candidate lagged output and input variables {$y(t\text{-}1)$, $y(t\text{-}2)$, …, $y(t-n_y)$, $u(t\text{-}1)$, $u(t\text{-}2)$, …, $u(t-n_u)$.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 wavelet network consisting of functional components up to m-dimensions*

(*i*) Determine the coarsest and finest resolution scales $j_1,\cdots,j_m$ and $J_1,\cdots,J_m$, where J_k $(1 \le k \le m)$ indicates the scales of the associated *k*-dimensional wavelets. Generally the initial resolution scales $j_k = 0$, and the finest resolution scales J_k $(1 \le k \le m)$ 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 models terms and then form a parsimonious model of the form (22).

4.4.2 Implement a wavelet network starting from low-order submodels

This identification procedure can be summarized below:

***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 tests*

If the identified *m*th 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…, m+1), go to and repeat from Step 3.

5. Examples

Three bench test examples are provided to illustrate the performance of the new wavelet networks. The first data set comes from a simulated continuous-time input-output system, the second come from a high-dimensional chaotic time series, and the third is the sunspot time series.

5.1 Example 1—a nonlinear continuous-time input-output system

Consider the Goodwin equation described by a nonlinear time-invariant continuous-time model (Coca, 1996)

$$\ddot{y}(t) + a\frac{y^2(t)-1}{y^2(t)+1}\dot{y}(t) + by(t) + cy^3(t) = u(t) \tag{23}$$

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 4th-order Runge-Kutta algorithm was used to simulate this model with the integral step $\Delta t = 0.01$, 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 sequences, $\tilde{u}_k = \tilde{u}(k) = \tilde{u}(kT)$ and $\tilde{y}_k = \tilde{y}(k) = \tilde{y}(kT)$, were normalised into the unit interval [0,1] using the fact that $\tilde{u}_k \in [-37,37]$ and $\tilde{y}_k \in [-7,7]$. Designate the normalised input and output sequences by $u_k = u(k) = u(kT)$ and $y_k = y(k) = y(kT)$.

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 wavelet network training and the test set consisting of the rest 2000 data points was used for model tests. A variable selection algorithm (Wei *et al.*, 2004) was performed on the estimation data set and three significant variables {*y*(*k*-1), *y*(*k*-2), *u*(*k*-1)} were selected. The initial wavelet network was chosen as

$$\begin{aligned} y(k) &= f(y(k-1), y(k-2), u(k-1)) \\ &= \sum_{p=1}^{3} f_p(x_p(k)) + \sum_{p=1}^{2}\sum_{q=2}^{3} f_{pq}(x_p(k), x_q(k)) + f_{123}(x_1(k), x_2(k), x_3(k)) \end{aligned} \tag{24}$$

where $x_p(k) = y(k-p)$ for *p*=1,2 and $x_3(k) = y(k-3)$. The one, two and three-dimensional Mexican hat radial wavelet networks 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 orthogonal least squares (OLS) algorithm, together with the error reduction ratio (ERR) criterion (Billings *et al.*, 1989; Chen *et al.*, 1989; Chen *et al.*, 1991) was applied to select significant model terms. The final identified model was found to be

$$\begin{aligned} y(k) &= 0.070723\psi_{2,1}(y(k-1)) + 1.846539\psi_{0,3}(y(k-2)) \\ &\quad + 1.734865\psi^{[2]}_{0;2,0}(y(k-1), y(k-2)) - 1.300637\psi_{3,6}(u(k-1)) \end{aligned} \tag{25}$$

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, where $j, k, k_1, k_2 \in \mathbb{Z}$.

Setting the input signal *u*(*k*)=cos(*kT*), *T*=0.02, and starting from the initial value *y*(1)=*y*(2)=0.5 (this is equivalent to the original $\tilde{y}(1) = \tilde{y}(2) = 0$, where $\tilde{y}$ indicates the original signal), the model (25) was simulated and the output y was recovered to its original amplitude by the inverse transform $\tilde{y}_{WN}(k) = y_{\min} + (y_{\max} - y_{\min})y(k)$, where $y_{\max} = -y_{\min} = 7$. The system output $\tilde{y}_{WN}$ from the model (25) was compared with that from the original model (23) over the validation set and is shown in Figure 2(a) and (b), which clearly indicates that the model (25) provides an excellent representation for the original system (23). 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 Figure 2(c). It follows that a conventional wavelet network, where only the total-variable-involved wavelet terms in (24) were considered, could not reach a result as good as that shown in Figure 2 under the condition that only the 3-dimensional wavelets were used.

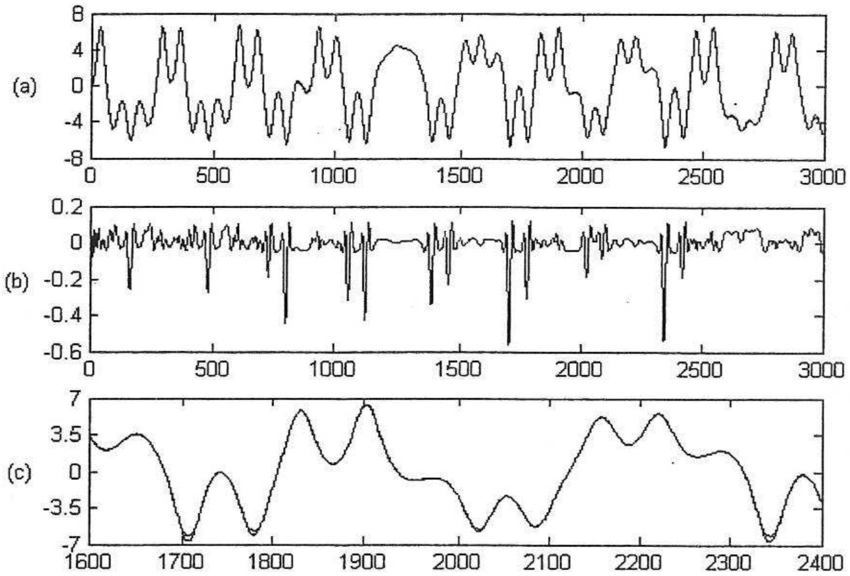


Figure 2 Comparison of the model output based on the wavelet network (25) with the measurements over the test set. (a) Overlap of the output of the wavelet network (25) and the measurements. (b) Discrepancy between the output of the wavelet network (25) and the measurements. (c) The interval [1600, 2400] 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.

5.2 Example 2—a high-dimensional chaotic system

Consider the Mackey-Glass delay-differential equation (Mackey and Glass, 1977)

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} \tag{26}$$

where the time delay τ was chosen to be 30 in this example. This example was chosen to facilitate comparisons with other results (Cao *et al*, 1995; Bone *et al*, 2002). Setting the initial condition $x(t) = 0.9$ for $0 \le t \le \tau$, a numerical integral algorithm was applied to calculate Eq. (26) with an integral step Δt =0.01 and 6000 equi-spaced samples, $x_k = x(k) = x(kT)$, (*k*=1,2, …, 6000) were recorded with a sampling interval of T =0.06 time units.

The recorded sequence was normalised into the unit interval [0,1] using the a priori knowledge $x_k \in [0.2, 1.4]$. Designate the normalised sequence by $y_k = y(k) = y(kT)$. The 6000 points was divided into two parts: the estimation set consisting of the first 500 points was used for wavelet network training and the validation set consisting of the remaining 5500 points was used for model tests. Following Casdagli (1989), 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), …, *y*(*k*-6)}. Similar to Example 1, the initial wavelet network was chosen to be

$$\begin{aligned} y(k) = f(y(k-1), y(k-2), \cdots, y(k-6)) &= \sum_{p=1}^{6} f_p(y(k-p)) + \sum_{p=1}^{5}\sum_{q=2}^{6} f_{pq}(y(k-p), y(k-q)) \\ &+ \sum_{p=1}^{4}\sum_{q=2}^{5}\sum_{r=3}^{6} f_{pqr}(y(k-p), y(k-q), y(k-r)) \end{aligned} \tag{27}$$

where the one, two and three-dimensional compactly supported Mexican hat radial wavelet networks 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 (Billings *et al.*, 1989; Chen *et al.*, 1989) was used to select significant model terms. The final identified model was found to be

$$\begin{aligned} y(k) = & -2.12656437\times10^{-3}\psi_{3,0}(y(k-1)) - 3.22204781\times10^{-2}\psi_{3,10}(y(k-1)) \\ & +1.63619675\times10^{-1}\psi_{3,-3}(y(k-3)) + 1.15434412\times10^{-3}\psi_{3,2}(y(k-3)) \\ & -6.93238261\times10^{-2}\psi_{1,3}(y(k-5)) + 1.39111531\times10^{-3}\psi_{3,9}(y(k-6)) \\ & +8.65313332\times10^{-3}\psi_{2,4}(y(k-6)) + 3.60856265\times10^{-2}\psi^{[2]}_{0;1,1}(y(k-1), y(k-2)) \\ & -2.37718815\times10^{0}\psi^{[2]}_{0;-1,1}(y(k-1), y(k-3)) - 1.71181790\times10^{-1}\psi^{[2]}_{1;1,2}(y(k-1), y(k-5)) \\ & -8.49930598\times10^{-2}\psi^{[2]}_{1;4,4}(y(k-1), y(k-6)) + 9.93000674\times10^{-2}\psi^{[2]}_{1;1,2}(y(k-2), y(k-6)) \\ & +2.09585403\times10^{-1}\psi^{[2]}_{0;1,1}(y(k-3), y(k-4)) + 4.94735553\times10^{-1}\psi^{[2]}_{0;-1,1}(y(k-3), y(k-5)) \end{aligned} \quad (28)$$

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 one and two dimensional compactly supported Mexican hat wavelets, where $j, k, k_1, k_2 \in \mathbb{Z}$.

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 other work (Cao *et al*, 1995; Bone *et al*, 2002), where only one-step-ahead predictions were considered. To facilitate comparisons, a measurement index, the relative error (Cao *et al*, 1995) , was used to measure the performance of the identified wavelet network. This index is defined as

$$E_k = |x_k - \hat{x}_k| / |x_k| \quad (29)$$

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

The results of two-step-ahead predictions of the wavelet network (28) were compared with the measurements and these are shown in Figure 3, where the data are plotted once every 100 points to obtain a clear inspection. The relative error E_k is shown in Figure 4, which clearly indicates that the underlying dynamics have been captured by the identified wavelet network (28). Notice that from Figure 4 the result of two-step-ahead predictions of the wavelet network (28) is by far better even than that of the one-step-ahead predictions provided by the wavelet networks proposed by Cao *et al* (1995), In fact, simulation results show that the relative error E_k with respect to the one-step-ahead predictions provided by the wavelet network (28) is by far smaller than those with respect to the two-step-ahead predictions. The standard derivation was calculated to be 0.0029 with respect to the two-step-ahead predictions of the wavelet network (28), which is by far smaller than that given by Bone *et al* (2002), where the one-step-ahead predictions were considered. These results obviously show that the new wavelet networks are more effective and efficient than conventional networks.

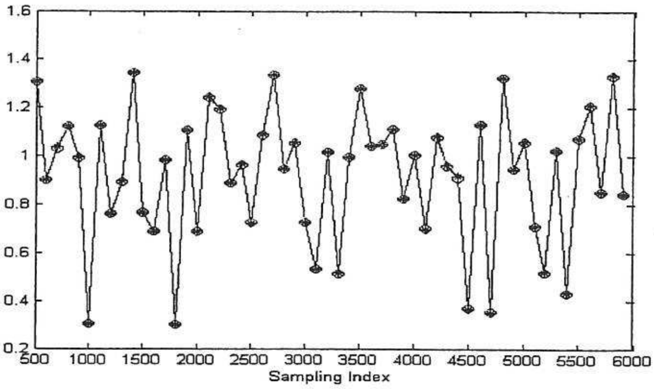


Figure 3 Two-step-ahead predictions for the Mackey-Glass delay-differential equation (26) using the identified wavelet network (28) over the validation set. The stars '*' indicate the measurements and the circles 'o' indicate the predications. For a clear inspection, the data are plotted once every 100 points.

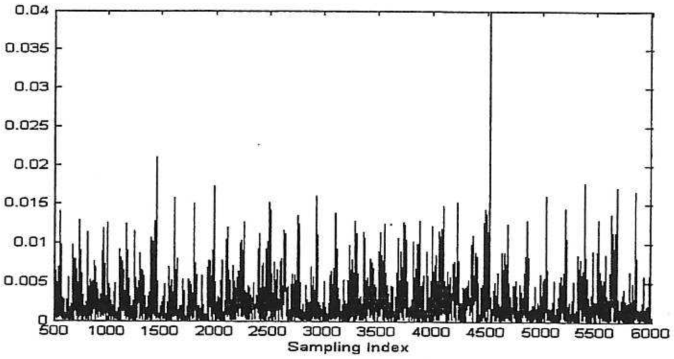


Figure 4 Relative errors between the two-step-ahead predictions from the identified wavelet network (28) and the measurements for the Mackey-Glass delay-differential equation (26) over the validation set.

5.3 Example 3—the sunspot time series

The sunspot time series considered in this example consists of 300 records of the dark spots on the sun from 1700 to 1999, see Figure 5. The objective here is to identify a wavelet network model to produce one-step-ahead predications for the sunspot data set. Again, the original measurements $\{\tilde{y}(t)\}(1 \le t \le 300\}$ were initially normalized into the unit interval [0,1] using the information $0 \le \tilde{y}(t) \le 200$. Design the normalised sequence by $\{y(t)\}$. The data set was separated into two parts: the training set consists of 250 data points corresponding to the period 1700-1949, and the test set consists of 50 data points corresponding to the period 1950-1999.

Following Wei *et al.* (2004), the model order was chosen to be *n*=9 here. It was also pointed out that *y*(*t*-1), *y*(*t*-2) and *y*(*t*-9) are the three most significant variables (Wei *et al.* 2004). The initial wavelet network model was therefore chosen to be

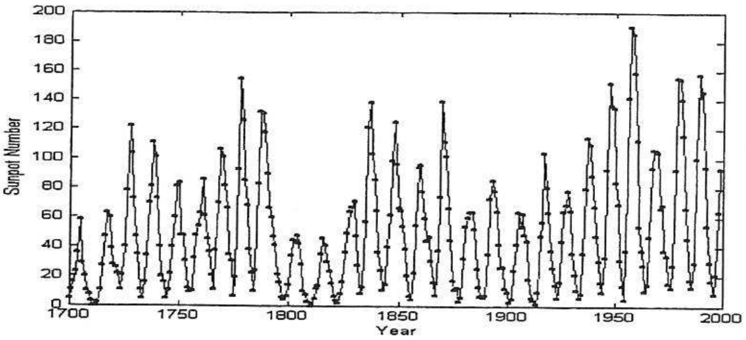


Figure 5 The sunspot time series for the period from 1700 to 1999.

$$
\begin{aligned}
y(t) &= f(y(t-1), y(t-2), \cdots, y(t-9)) \\
&= \sum_{p=1}^{9} f_p(x_p(t)) + \sum_{p=1}^{2}\sum_{q=2}^{3} f_{pq}(z_p(t), z_q(t)) \ + f_{123}(z_1(t), z_2(t), z_3(t))
\end{aligned} \tag{30}
$$

where $x_p(t) = y(t-p)$ for p=1,2,...,9, $z_k(t) = y(t-k)$ for k=1,2, and $z_3(t) = y(t-9)$. The one, two and three-dimensional compactly supported Gaussian radial wavelet networks 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 $J_1 = 2$, $J_2 = J_3 = 0$. A forward OLS-ERR algorithm (Billings *et al.*, 1989; Chen *et al.*, 1989) was used to select significant model terms. The final identified model was found to be

$$
y(t) = \sum_{k=1}^{24} \theta_k p_k(t) \tag{31}
$$

where $p_k(t)$ are the wavelet terms formed by compactly supported Gaussian wavelets. The identified wavelet terms, the corresponding parameters, and the associated error reduction ratios (ERRs) are listed in Table 1. 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 for example Billings *et al.* (1989) and Chen *et al.* (1989). The result of the one-step-ahead predictions based on the wavelet network (31) over the test set is shown in Figure 6 (the dashed-star line), which clearly shows that the identified model provides an excellent representation for the sunspot time series.

In order to compare the predicted result of the wavelet network with other work (Soltani, 2002), the following index, the normalized error on the test set, was used to measure the performance of the identified wavelet network

$$\overline{E} = \sum_{k=1}^{N_{test}} \left|x_k - \hat{x}_k\right|^2 \Big/ \sum_{k=1}^{N_{test}} \left|x_k - \overline{x}\right|^2 \tag{32}$$

where x_k and $\hat{x}_k$ are defined as in Eq. (29), N_{test} is the length of the test set , and $\overline{x} = \sum_{k=1}^{N_{test}} x_k$. It was calculated that $\overline{E}$ =0.0651 for the identified wavelet network (31) and this is smaller than that given by a wavelet decomposition model proposed in Soltani (2002).

Table 1 The wavelet terms, parameters and associated error reduction ratios for the sunspot time series.

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

Note: $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, $\psi^{[2]}_{j;k_1,k_2}(x_1, x_2) = 2^j \psi^{[2]}(2^j x_1 - k_1, 2^j x_2 - k_2)$, and $\psi^{[3]}_{j;k_1,k_2,k_3}(x_1, x_2, x_3) = 2^{3j/2}\psi^{[3]}(2^j x_1 - k_1, 2^j x_2 - k_2, 2^j x_3 - k_3)$ are the compactly supported Gaussian wavelets.

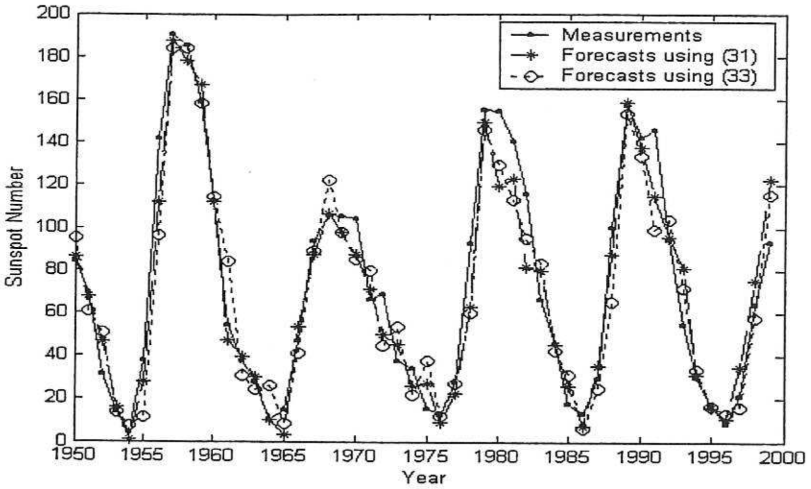


Figure 6 One-step-ahead predictions for the sunspot time series based the wavelet networks (31) and (33) over the test set. The point-solid line indicates the measurements, the dashed-star line indicates the predictions from (31), and the dotted-circle line indicates the predications from (33).

An important point revealed by Table 1 is that the three variables $y(t\text{-}1)$, $y(t\text{-}2)$ and $y(t\text{-}9)$ are far more significant than the other variables. This is consistent with the result given in Wei *et al.* (2004). In fact, the sunspot time series can be satisfactory described using a wavelet network with respect to only these three significant variables. This model is given below

$$\begin{aligned}
y(t) = &-0.12759\psi_{2,-3}(y(t-1)) - 9.41469\psi_{2,7}(y(t-1)) \\
&+ 0.05504\psi_{2,3}(y(t-2)) + 1.30897\psi_{2,-3}(y(t-2)) \\
&- 3.67994\psi^{[2]}_{0;-2,0}(y(t-1), y(t-9)) - 1.57733\psi^{[2]}_{0;0,1}(y(t-1), y(t-9)) \\
&+ 6.47437\psi^{[2]}_{0;3,-2}(y(t-1), y(t-9)) + 4.22507\psi^{[2]}_{0;-2,0}(y(t-2), y(t-9)) \\
&- 2.41246\psi^{[3]}_{0;0,-1,2}(y(t-1), y(t-2), y(t-9))
\end{aligned} \tag{33}$$

The one-step-ahead predictions from the wavelet network (33) over the test set is shown in Figure 6 (the dotted-circle line), where the normalized error $\overline{E}$ was calculated to be 0.1044, which is still very small.

6. Conclusions

A new class of wavelet networks (WNs) has been introduced for nonlinear system identification. The main advantage of the new identification approach compared with existing wavelet networks is that the new wavelet networks are more practical and applicable for handling problems in medium and high-dimensions as well as in low-dimensions due to the fact that the structure of the new wavelet networks bypass the dilemma of the curse-of-dimensionality.

It has been noted that a conventional wavelet network always includes the total-variable-involved wavelet terms, but this is not the case for most systems in the real world. In addition, a model that includes only the high-order terms is liable to produce a deleterious effect on the output behaviour of the model and can often induce spurious dynamics. From this point of view, the new wavelet networks are more reasonable compared with conventional wavelet networks by decomposing a multi-dimensional function into a number of low-dimensional submodels.

In theory, any wavelets can be used to approximate the low-dimensional submodels by a scheme of taking tensor products or adopting radial functions. For convenience of network training, 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 is deterministic. Radial wavelets are not compactly supported but rapidly vanishing. It is therefore reasonable to truncate a radial wavelet to make it become quasi-supported, this can then be used as a normal compactly supported wavelet to implement a wavelet network. Most radial wavelets including the Gaussian and Mexican hat wavelets are easy to calculate with very slight computation load and can therefore be chosen to implement a wavelet network. Other non-radial wavelets, which are either compactly supported or not, can also be available if there is strong evidence that these wavelets can easily be performed to implement a wavelet network.

A wavelet network 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. An efficient term detection algorithm was employed to train wavelet networks and finally parsimonious models were obtained.

In summary, the new wavelet networks are advantageous over existing conventional wavelet modelling 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.

Acknowledgment

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

References

Aguirre, L.A. (1994). *Application of Global Models in the Identification, Analysis and Control of Nonlinear Dynamics and Chaos*. PhD Thesis, Department of Automatic Control and Systems Engineering, the University of Sheffield, UK.

Allingham, D., West, M., and Mees, A. (1998). Wavelet reconstruction of nonlinear dynamics. *International Journal of Bifurcation and Chaos*, **8(11)**, 2191-2201.

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

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

Billings, S.A., 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 Wei, H.L. (2003). The wavelet-NARMAX representation: a hybrid model structure combining polynomial models with multiresolution wavelet decompositions. (*submitted for publication*).

Bone, R., Crucianu, M. and de Beauville, J.-P. A. (2002). Learning long-term dependences by the selective additional time-delayed connetions to recurrent neural networks. *Neurocomputing*, **48**, 251-266.

Cao, L.Y., Hong, Y.G., Fang, H.P. and He, G.W. (1995). Predicting chaotic time series with wavelet networks. *Physica* **D**, **85(1-2)**, 225-238.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, **35(3)**, 335-356.

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

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

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

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

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

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

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

Chiras, N. (2002). *Linear and Nonlinear Modelling of Gas Turbine Engines*. PhD Thesis, the University of Glamorgan, Brazil.

Coca, D. (1996). *A Class of Wavelet Multiresolution Decompositions for Nonlinear System Identification and Signal Processing*. PhD Thesis, Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield, UK.

Coca, D. and Billings, S.A. (1997). Continuous-time system identification for linear and nonlinear system identification using wavelet decompositions. *International Journal of Bifurcation and Chaos*, **7(1)**, 87-96.

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

Daugman, J.G. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics Speech and Signal Processing*, **36(7)**, 1169-1179.

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.

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). *International Journal of Control*, **41(2)**,303-344.

Lin, T.N., Horne, B.G., Tino, P. and Giles, C.L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, **7(6)**, 1329-1338.

Mackey, M.C. and Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science*, **197**, 287-289.

Mallat,S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11(7)**, 674-693.

Mallat, S.G. (1998). *A Wavelet Tour of Signal Processing*. San Diego: Academic Press.

Meyer, Y. (1993). *Wavelets: Algorithms and Applications*. Philaelphia, Pennsylvania: SIAM.

Narendra, K.S. and Mukhopadhyay, S. (1997). Adaptive control using neural networks and approximate models. *IEEE Transactions on Neural Networks*, **8(3)**, 475-485.

Narendra, K.S. and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, **1(1)**, 252-262.

Oussar, Y. and Dreyfus, G. (2000). Initialization by selection for wavelet network training. *Neurocomputing*, **34**, 131-143.

Pati, Y.C. and Krishnaprasad, P.S. (1993). Analysis and synthesis of feedforward neural networks using discrete affine wavelet transforms. *IEEE Transactions on Neural Networks*, **4(1)**, 73-85.

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

Pearson, R.K. (1997). Nonlinear Process Identification. In *Nonlinear Process Control* (Eds: M.A. Henson and D.E. Seborg). New Jersey: Prentice Hall.

Pearson, R.K. (1999). *Descrete-Time Dynamic Models.* Oxford: Oxford University Press.

Pittner, S., Kamarthi, S.V. and Gao, Q.G. (1998). Wavelet networks for sensor signal classification in flank wear assessment. *Journal of Intelligent Manufacturing*, **9(4)**, 315-322.

Sastry, P.S., Santharam, G. and Unnikrishnan, K.P. (1994). Memory neuron networks for identification and control of dynamical systems. *IEEE Transactions on Neural Networks*, **5(2)**, 306-319.

Soltani, S. (2002). On the use of wavelet decomposition for time series prediction. *Neurocomputing*, **48**, 267-277.

Szu, H.H., Telfer, B. and Kadambe, S. (1992). Neural network adaptive wavelets for signal representation and classification. *Optical Engineering*, **31(9)**, 1907-1916.

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

Wei, H.L. and Billings, S.A. (2004). A unified wavelet-based modelling framework for nonlinear system identification: the WANARX model structure. (*Accepted for publication by the International Journal of Control*).

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

Wong, K.W. and Leung, A.C-S. (1998). On-line successive synthesis of wavelet networks. *Neural Processing Letters*, **7(2)**, 91-100.

Zhang, J., Walter, G.G., Miao, Y. and Lee, W.N.W. (1995). Wavelet neural networks for function learning. *IEEE Transactions on Signal Processing*, **43(6)**, 1485-1497.

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

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