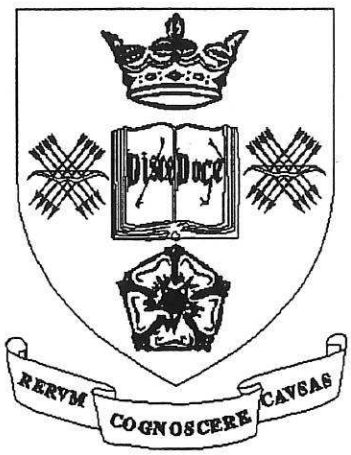
Long Term Prediction of Nonlinear Time Series Using Multiresolution Wavelet Models

H.L. Wei and S.A. Billings



Research Report No. 882

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

Feb. 2005

Long Term Prediction of Nonlinear Time Series Using Multiresolution Wavelet Models

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

The long term prediction of nonlinear dynamical time series, based on identified multiresolution wavelet models, from historically observed data sets is investigated and a new direct prediction approach is introduced. Prediction results based on the new direct scheme are compared with those from conventional iterative methods and it is shown that improved predictions can be obtained using the new approach.

Keywords: long term prediction, multiresolution wavelet decomposition, multi-step-ahead prediction, time series

1. Introduction

Forecasting future behaviour based on past observations has been a long standing topic in system identification and time series modelling (Box and Jenkins 1976, Pandit and Wu 1983). Techniques for prediction have been considerably advanced in recent years and many approaches have been proposed for predicting different types of nonlinear time series, see for example Zhang et al.(1998) and Judd and Small (2000). Existing approaches for nonlinear time series prediction include local approximations (Farmer and Sidorowich 1987, Jimensz et al. 1992, McNames 2002), the polynomial NARMA (Nonlinear Auto-Regressive Moving Average) model (Chen and Billings 1989), radial basis functions (Casdagli 1989, Lee and Billings 2003), neural networks (Weigend et al. 1990, Tang et al. 1991, Principe et al. 1992 , Hill et al. 1994, Lin et al. 1996, Atiya et al. 1999, Parlos et al. 2000) and wavelet models (Cao et al. 1995, Soltani 2002).

To obtain multi-step-ahead predictions of nonlinear time series, both the iterative and the direct approaches can be used. In theory, long-term predictions can be obtained from a short-term predictor, for example a one-step-ahead predictor, simply by applying the short predictor many times (steps) in an iterative way. This is called iterative prediction and is similar to the Box-Jenkins method (Box and Jenkins 1976). Direct prediction, however, provides a once-completed predictor with a long term prediction step and specified multi-step-ahead predictions can be obtained directly from the established predictor in a way that is similar to computing one-step-ahead predictions. Whilst it is possible and easy to compare the iterative and direct approaches for a given specific problem, it is difficult to give a general conclusion on which should be used in a given situation. In fact, there has been and still is a debate on the issue of whether one method is advantageous over the other. Farmer and Sidorowich (1987) believed that under certain conditions an iterative predictor would be better than a direct predictor and this was demonstrated by Weigend et al. (1992) in the specific case of the modelling and forecasting of sunspot data. General situations considered in practice, however, might not satisfy the required conditions. The argument that an iterative predictor is better than a direct approach for general cases therefore lacks sufficient evidence with theoretical foundations. Perhaps the main problem of iterative prediction is that

this method does not take into account the total errors accumulated from each step and the fact that it is difficult to estimate the distribution property of the accumulated errors (McNames 2002). Some authors (Zhang et al. 1998, Atiya et al. 1999) therefore argued that the direct approach could give a better prediction than the iterative method for nonlinear time series.

This study attempts to construct a new direct predictor using a class of wavelet based nonlinear ARMA models. Owing to the excellent approximation properties of wavelets (Chui 1992, Daubechies 1992, Meyer 1993, Mallat 1998), wavelet-based approaches have been introduced to describe nonlinear dynamical systems in recent years. In the literature, wavelet models for nonlinear dynamical systems can be broadly classified into three categories: wavelet networks (Cao et al. 1995, Zhang 1997, Allingham et al. 1998), multiresolution wavelet expansions (Coca and Billings 1997, Billings and Coca 1999), and wavelet-NARMAX models (Wei and Billings 2004, Billings and Wei 2005). These approaches have been applied to identify both time-invariant systems (Sureshbabu and Farrell 1999, Ghanem and Romeo 2001, Wei et al. 2004) and time-varying systems (Tsatsanis and Giannakis 1993, Zheng et al. 2001, Wei and Billings 2002). Motivated by the successful applications of these wavelet approaches in nonlinear dynamical modelling, especially by the flexibility and adaptability possessed by the wavelet-NARMAX models(Billings and Wei 2005), a class of multiresoluton wavelet models is introduced to construct a direct predictor for accurate multi-step-ahead predictions of nonlinear time series.

2. Problem representation

Most time series encountered in practice can be described using the general NARMA (Nonlinear Auto-Regressive Moving Average) model (Chen and Billings 1989)

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

where $y(t)$ (t=1,2, ...) is the output signal, $e(t)$ is an unobserved and unpredictable zero mean noise sequence, $f(\cdot)$ is some nonlinear function, n_y and n_e are the maximum lags for the output and noise signal, respectively. In some cases, the function $f(\cdot)$ is not dependent on the lagged noise signals $e(t-k)$ (k=1,2, ...), the NARMA model (1) then reduces to the NAR model

$$y(t) = f(y(t-1),\cdots,y(t-n_y)) + e(t) \tag{2}$$

For a real system, the nonlinear function $f(\cdot)$ in (1) and/or (2) is generally unknown and might be very complex. The objective of any practical identification procedure is to determine an approximation for $f(\cdot)$ based on a chosen model set of known functions. In practice, many types of functions have been chosen to approximate the unknown function $f(\cdot)$ including polynomials, neural networks, radial basis functions, neuro-fuzzy basis functions and wavelets. To obtain an approximator $\hat{f}(\cdot)$ for the unknown function $f(\cdot)$, most approximation (training) schemes take the following or a similar criteria

$$J = \|\{e(t)\}\| = \frac{1}{N}\sum_{t=1}^{N}\left|y(t) - \hat{f}(\varphi(t))\right|^2 \tag{3}$$

where N is the length of the observation of the time series, φ is some regression vector formed as

$$\varphi(t)=[y(t-1),\cdots,y(t-n_y),e(t-1),\cdots,e(t-n_e)]^T \tag{4}$$

Note that for a NAR model the regression vector φ does not contain any lagged noise signals $e(t-k)$ (k=1,2, ...). Also note that the approximator $\hat{f}(\cdot)$ is in effect a one-step-ahead predictor. The noise signal $e(t)$ in Eq. (1) is generally unobserved and is often replaced by the model residual sequence in the sense that

$$\varepsilon(t)=y(t)-\hat{y}(t)=y(t)-\hat{f}(y(t-1),\cdots,y(t-n_y),0,\cdots,0) \tag{5}$$

The regression vector (4) becomes

$$\varphi(t)=[y(t-1),\cdots,y(t-n_y),\varepsilon(t-1),\cdots,\varepsilon(t-n_\varepsilon)]^T \tag{6}$$

Once the one-step-ahead approximator $\hat{f}(\cdot)$ has been obtained, this is often used to calculate multi-step-ahead (say s ahead) prediction in an iterative way

$$\hat{y}(t+s)=\hat{f}(\hat{\varphi}(t+s)) \tag{7}$$

where

$$\hat{\varphi}(t+s)=[\hat{y}(t+s-1),\cdots,\hat{y}(t),y(t-1),\cdots,y(t-n_y),\varepsilon(t+s-1),\cdots,\varepsilon(t-n_\varepsilon)]^T \tag{8}$$

for $s=1,2,\cdots$. Note that in each iteration step, $\hat{y}(t+s)=\hat{f}(\hat{\varphi}(t+s))\neq\hat{f}(\varphi(t+s))$, where $\varphi(t+s)$ is defined as in (4). The errors for the s-step-ahead prediction are the accumulation of the errors of the previous (s-1)th steps. Generally, the longer the forecasting horizon, the larger the accumulated errors are and the less accurate the iterative method is.

The direct approach can be considered by identifying the direct relation

$$y(t+s)=f^{(s)}(y(t),\cdots,y(t-n_y+1),\varepsilon^{(s)}(t),\cdots,\varepsilon^{(s)}(t-n_\varepsilon+1)) \tag{9}$$

where $f^{(s)}(\cdot)$ for $s=1,2,\cdots$ are some complex nonlinear functions, $\varepsilon^{(s)}(\cdot)$ are s-step-ahead prediction errors similar to the model residuals defined by (5). Note that the analytical representation for nonlinear functions $f^{(s)}(\cdot)$ may be very complex even for a system described by a very simple model. To illustrate this, consider a second-order deterministic system described by

$$y(t)=1-ay^2(t-1)+by(t-2) \tag{10}$$

The analytical representation of the s-step-ahead prediction function for this system is of the form

$$f^{(s)}(y(t),y(t-1))=g_s(y(t),y^2(t),\cdots,y^{2^s}(t),y(t-1),y^2(t-1),\cdots,y^{2^{s-1}}(t-1)) \tag{11}$$

where $g_s(\cdot)$ are complex functions which will become very complicated for large s. Traditional models may thus be inappropriate to describe the s-step-ahead predictor $f^{(s)}(\cdot)$ and may therefore fail to provide good multi-step predictions. A class of models that are both flexible, have excellent approximation capabilities, and which can represent a broad class of highly complex systems are therefore required to ensure accurate direct s-step-ahead predictions. The model class that uses wavelets as the basis functions to approximate the s-step-ahead

predictor $f^{(s)}(\cdot)$ satisfies all these conditions and will therefore be investigated in the present study as a new approach of achieving accurate direct s-step-ahead predictions.

3. The multiresolution wavelet model for long term prediction

Following Billings and Wei (2005), the unknown nonlinear function $f^{(s)}(\cdot)$ in Eq. (9) is assumed to consist of two parts with the form

$$\begin{aligned} y(t+s) &= f^{(s)}(y(t),\cdots,y(t-n_y+1),\varepsilon^{(s)}(t),\cdots,\varepsilon^{(s)}(t-n_\varepsilon+1)) \\ &= P^{(s)}(y(t),\cdots,y(t-n_y+1))+Q^{(s)}(\varepsilon^{(s)}(t),\cdots,\varepsilon^{(s)}(t-n_\varepsilon+1)) \end{aligned} \tag{12}$$

where $P^{(s)}(\cdot)$ and $Q^{(s)}(\cdot)$ are some nonlinear functions associated with the process and the errors terms, respectively. In this study the function $Q^{(s)}(\cdot)$ will be approximated using a traditional polynomial model. The function $P^{(s)}(\cdot)$, however, will be approximated using a multiresolution wavelet model. This is described below.

The nonlinear function $P^{(s)}(\cdot)$ can be decomposed into a number of functional components via the well known functional analysis of variance (ANOVA) expansions (Friedman 1991,Chen 1993, Wei and Billings 2004)

$$\begin{aligned} P^{(s)}(t) &= P^{(s)}(x_1(t),x_2(t),\cdots,x_n(t)) \\ &= P_0^{(s)}+\sum_{i_1=1}^{n}P_{i_1}^{(s)}(x_{i_1}(t))+\sum_{1\le i_1<i_2\le n}P_{i_1i_2}^{(s)}(x_{i_1}(t),x_{i_2}(t)) \\ &+\sum_{1\le i_1<i_2<i_3\le n}P_{i_1i_2i_3}^{(s)}(x_{i_1}(t),x_{i_2}(t),x_{i_3}(t))+\cdots \end{aligned} \tag{13}$$

where $x_p(t)=y(t-p+1)$ for p=1,2, ..., n with $n=n_y$. The first functional component $P_0^{(s)}$ is a constant to indicate the intrinsic varying trend; $P_i^{(s)}$, $P_{ij}^{(s)},\cdots$, are univariate, bivariate, etc., functional components. The ANOVA expansion (13) can be viewed as a special form of the NAR model for dynamical time series. Although the ANOVA decomposition of the NAR model (13) involves up to 2^n different functional components, experience shows that a truncated representation containing the low-order components, say up to the bivariate or tri-variate functional terms is often sufficient to provide a satisfactory description for many high dimensional problems providing that the input variables are properly selected (Wei and Billings 2004). An exhaustive search for all the possible submodel structures of (13) 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. 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 $P_0^{(s)}$ can often be set to zero. If the constant term is different from zero for a given system, this can then be approximated by a wavelet expansion providing that the approximation is restricted to a compact subset of R^n.

The one-dimensional functions $P_{i_1}^{(s)}(x_{i_1})$ in the model (13) can be decomposed using a multiresolution decomposition (Mallat 1989, Chui 1992, Daubechies 1992) as

$$P_{i_1}^{(s)}(x_{i_1}(t)) = \sum_k \alpha_{j_0,k}^{(i_1)} \phi_{j_0,k}(x_{i_1}(t)) + \sum_{j \geq j_0} \sum_k \beta_{j,k}^{(i_1)} \psi_{j,k}(x_{i_1}(t)), \; i_1 = 1,2,\cdots,n_y \tag{14}$$

where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, $j,k \in \mathbf{Z}$ are the wavelet and associated scaling functions. The results of multiresolution wavelet decomposition for one-dimensional case can be extended to higher-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, a d-dimensional wavelet $\psi^{[d]}: \mathbf{R}^d \mapsto \mathbf{R}$ can be constructed using a scalar wavelet ψ as follows

$$\psi^{[d]}(x) = \psi^{[d]}(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} \psi(x_i) \tag{15}$$

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

$$\psi^{[d]}(x) = \psi^{[d]}(x_1, x_2, \cdots, x_d) = x_1 x_2 \cdots x_d e^{-\frac{1}{2}\|x\|^2} \tag{16}$$

where $\|x\|^2 = x^T x = \sum_{i=1}^{d} x_i^2$. Similarly, the n-dimensional Mexican hat (also called the Marr) wavelet can be expressed as $\psi^{[d]}(x) = (d - \|x\|^2)\exp(-\|x\|^2/2)$.

In the present study, the tensor product approach is used to implement multi-variate nonlinear functions. Following Wei and Billings (2004), the multivariate functional component $P^{(s)}{}_{i_1 i_2 \cdots i_d}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_d}(t))$ in the model (13) can be approximated using a truncated wavelet series as

$$\begin{aligned} & P_{i_1 i_2 \cdots i_d}^{(s)}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_d}(t)) \\ & = \sum_{j=j_d}^{J_d} \sum_{k_1 \cdots k_d} c_{j;k_1,k_2,\cdots k_d}^{(i_1 i_2 \cdots i_d)} B_d(2^j x_{i_1}(t) - k_1, 2^j x_{i_2}(t) - k_2, \cdots, 2^j x_{i_d}(t) - k_d) \\ & = \sum_{j=j_d}^{J_d} \sum_{k_1,k_2,\cdots,k_d} c_{j;k_1,k_2,\cdots,k_d}^{(i_1 i_2 \cdots i_d)} \prod_{m=1}^{d} B(2^j x_{i_m}(t) - k_m) \end{aligned} \tag{17}$$

where $k = [k_1, k_2 \cdots, k_d]^T \in \mathbf{Z}^d$ is an d-dimensional index, $B_d(x)$ is an d-dimensional wavelet or scaling function and can be expressed using a one-dimensional function $B(x)$ by taking the direct product. Generally, all the initial coarsest resolutions j_d $(d = 1,2,\cdots,n_y)$ are chosen to be the same, that is, $j_1 = j_2 = \cdots = j_{n_y} = j_0$. Similarly, all the finest resolution levels can also be set to the same value, that is, $J_1 = J_2 = \cdots = J_{n_y} = J$. In addition, in the high-dimensional wavelet decomposition (17), the initial coarsest and finest resolutions can generally be set to be $j_d = J_d = J$ for $d = 2,3,\cdots,n_y$.

Although many functions can be chosen as scaling and/or wavelet functions, most of these are not suitable for system identification applications, especially in the case of multidimensional and multiresolution expansions. An

implementation, which has been tested with very good results, involves B-spline and B-wavelet functions in multiresolution wavelet decompositions (Coca and Billings 1997, Billings and Coca 1999, Wei and Billings 2004, Billings and Wei 2005). The main advantage of adopting B-spline wavelets is that this class of wavelet and scaling functions form multiresolution decompositions for arbitrary given functions, and more importantly, each wavelet and scaling basis function in such decompositions is compactly supported. This is quite desirable for nonlinear system identification since all the candidate model terms can be determined beforehand once all the resolution levels are chosen (Wei and Billings 2004). For a comprehensive discussion on B-splines and B-wavelets, see Chui (1992).

Assume that M_P wavelet model terms (produced by the combination of mother wavelet and/or scaling functions) are required to expand the function $P^{(s)}(\cdot)$ and M_Q polynomial model terms are required to approximate the function $Q^{(s)}(\cdot)$ in the NARMA model (12), and for convenience of representation also assume that the $M = M_P + M_Q$ model terms are ordered with a single index m, (12) can then be expressed as a linear-in-the-parameters form as below:

$$y(t+s) = \sum_{m=1}^{M_P} \theta_m^{(s)} p_m^{(s)}(x(t)) + \sum_{m=M_P+1}^{M_Q} \theta_m^{(s)} p_m^{(s)}(\xi(t)) \tag{18}$$

where $x(t) = [y(t), \cdots, y(t-n_y+1)]^T$, $\xi(t) = [\varepsilon^{(s)}(t), \cdots, \varepsilon^{(s)}(t-n_\varepsilon+1)]^T$, $p_m^{(s)}(\cdot)$ are regressors produced by $x(t)$ and $\xi(t)$, $\theta_m^{(s)}$ are unknown parameters. Eq. (18) can be solved using linear regression techniques. Note that for large n_y, the model (18) might involve a great number of model terms or regressors. Practice and experience shows that very often many of the model terms are redundant and therefore are insignificant to the system output and can be removed from the model. An efficient orthogonal least squares (OLS) algorithm and an error reduction ratio (ERR) criterion (Korenberg et al. 1988, Billings et al. 1988, Chen et al. 1989) can be used to determine which terms should be included in the model.

4. A procedure to implement the multiresolutoin wavelet model

Two schemes are often be applied to implement the multiresolution wavelet model. 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 model 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 model 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 interest, higher dimensional wavelet terms should be added to the wavelet model (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 are combined here to implement the wavelet model and this is summarized below.

Step 1: *Determining the model initial conditions*. This includes:

(*i*) Select the maximum lag n_y for the wavelet-autoregressive model $P^{(s)}(\cdot)$ and the maximum lag n_ε for the polynomial moving average model $Q^{(s)}(\cdot)$, where s indicates how many step ahead values are to be predicted.

(*ii*) Select the significant variables from all the candidate lagged variables $\{y(t\text{-}1), y(t\text{-}2), \ldots, y(t-n_y)\}$.

(*iii*) Determine m, the highest dimension of all the submodels (functional components) in (13). A typical recommended start value for m is 2 or 3.

Step 2: *Identify the wavelet model 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_d $(1\le d\le m)$ indicates the scales of the associated k-dimensional wavelets. Generally the initial resolution scales $j_d=0$, and the finest resolution scales $J_d\le 5\,(1\le d\le m)$.

(*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 (18).

Step 3: *An iterative loop to identify a wavelet model*

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

$$\begin{aligned}\varepsilon^{(s)(0)}(t) &= y(t+s)-\hat{P}^{(s)}(t)\\ &= y(t+s)-\hat{P}^{(s)}(y(t),\cdots,y(t-n_y+1))\\ &= y(t)-\sum_{i=1}^{M_0} g_i^{(k)} w_i^{(k)}(t)\end{aligned} \tag{19}$$

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

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

$$\begin{aligned}\varepsilon^{(s)(k)}(t) &= y(t+s)-\hat{f}(y(t),\cdots,y(t-n_y+1),\varepsilon^{(s)(k-1)}(t),\cdots,\varepsilon^{(s)(k-1)}(t-n_\varepsilon+1))\\ &= y(t+s)-\sum_{m=1}^{M_0+m_e}\theta_{\ell_m}^{(k)} p_{\ell_m}(t)\end{aligned} \tag{20}$$

where m_e is the number of noise terms selected. The above recursive calculation can be terminated after a number of iterations. Numerous tests have shown that less than 10 iterations, typically 3-5 iterations, are sufficient for the algorithm to converge.

Step 4: *Model validity tests*

If the identified model in Step 3 provides a satisfactory representation, then terminate the procedure. Otherwise, go to Step 1 to change one or more key conditions and repeat the procedure.

To check whether an identified model is adequate to provide a sufficient description for a given data set, the validity of the model will be testified using the following tests (Billings and Tao 1991)

$$\begin{cases} \gamma_{\varepsilon\varepsilon}(\tau) = \delta(\tau), & \forall \tau \\ \gamma_{\varepsilon^2\varepsilon^2}(\tau) = \delta(\tau), & \forall \tau \end{cases} \tag{21}$$

where $\delta(\tau)$ is the Kronecker delta function, $\varepsilon(t)$ are the model residuals, the correlation function $\gamma_{\xi\eta}(\tau) = E\{[\xi(t) - E(\xi(t))][\eta(t+\tau) - E(\eta(t))]\}$ can be estimated as

$$\gamma_{\xi\eta}(\tau) = \frac{\sum_{t=1}^{N-\tau}[\xi(t) - \bar{\xi}][\eta(t+\tau) - \bar{\eta}]}{\sqrt{\left(\sum_{t=1}^{N-\tau}[\xi(t) - \bar{\xi}]^2\right)\left(\sum_{t=1}^{N-\tau}[\eta(t+\tau) - \bar{\eta}]^2\right)}} \tag{22}$$

In practice, if these correlation functions fall within the confidence intervals at a given significance level $\alpha(0 < \alpha < 1)$, say α =0.05, which corresponds to the 95% confidence interval, the model is viewed as adequate and acceptable. For large N, these confidence intervals are approximately $\pm 1.96/\sqrt{N}$.

Alternatives to the above correlation tests for long term predictions are also available, for example, the direct comparisons between the model outputs and the corresponding measurements.

5. Examples

Two examples are provided to illustrate the application of the proposed direct approach for long term prediction of nonlinear time series. In both examples the 4th-order B-spline wavelet and associated scaling functions (Chui 1992) were used. Note that in the following the symbol $y(t)$ will be used to represent both continuous and discrete-time signals, the meaning should be clear from the context.

5.1 Example 1 –a two dimensional autonomous system

The following model was taken from Billings and Chen (1998)

$$\begin{aligned} y(t) = {} & (0.8 - 0.5e^{-y^2(t-1)})y(t-1) - (0.3 + 0.9e^{-y^2(t-1)})y(t-2) \\ & + 0.1\sin(\pi y(t-1)) + e(t) \end{aligned} \tag{23}$$

where $e(t)$ is a Gaussian white noise with a probability distribution $N(0,0.02^2)$. This model which has been used as a benchmark example by several authors was simulated starting with the initial condition $y(-1)=0$, $y(0)=1$ and 1200 data points were sampled after the time series settled down. The first 600 points were used for model estimation and the remaining 600 data points were used for model testing. The objective here was to obtain direct multi-step-ahead predictions using the proposed wavelet modelling approach. However, prior to identifying a multi-step-ahead prediction model using the proposed direct approach, two models, one polynomial NARMA model and another, multiresolution wavelet model, were initially estimated and these were used to predict the future behaviour of the time series in an iterative way.

5.1.1 Iterative models

The significant variables were detected to be $y(t)$ and $y(t\text{-}1)$ using a variable detection algorithm (Wei et al. 2004). The initial wavelet-NARMA model was chosen to be

$$\begin{aligned} y(t+s) &= P^{(s)}(y(t), y(t-1)) + Q^{(s)}(\varepsilon^{(s)}(t), \cdots, \varepsilon^{(s)}(t-20)) \\ &= \sum_k \alpha_{0,k}^{(1)} \phi_{0,k}(y(t)) + \sum_{j=0}^{5} \sum_k \beta_{j,k}^{(1)} \psi_{j,k}(y(t)) \\ &+ \sum_k \alpha_{0,k}^{(2)} \phi_{0,k}(y(t-1)) + \sum_{j=0}^{5} \sum_k \beta_{j,k}^{(2)} \psi_{j,k}(y(t-1)) \\ &+ \sum_{j=0}^{2} \sum_{k_1,k_2} \gamma_{j;k_1,k_2} \phi_{j,k_1}(y(t)) \phi_{j,k_2}(y(t-1)) \\ &+ \sum_{k=0}^{20} c_k \varepsilon(t-k) \end{aligned} \tag{24}$$

where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, $j, k \in Z$ are the 4th-order B-spline wavelet and associated scaling functions.

Set s=1. Starting from the initial model (24), a multiresolution wavelet-NARMA model was identified based on the estimation data set by performing the identification procedure proposed in Section 4, to give the wavelet model

$$\begin{aligned} y(t+1) = {} & 1.318223\phi_{0,0}(y(t)) - 1.020386\phi_{0,-4}(y(t)) - 0.813415\psi_{0,-1}(y(t)) \\ & - 0.502918\phi_{0,-2}(y(t-1)) + 1.838834\phi_{0,-2}(y(t))\phi_{0,-3}(y(t-1)) \\ & - 1.156073\phi_{0,-3}(y(t))\phi_{0,-1}(y(t-1)) - 3.799698\phi_{0,-2}(y(t))\phi_{0,0}(y(t-1)) \\ & + 1.298586\phi_{0,-1}(y(t))\phi_{0,-2}(y(t)) - 1.317584\phi_{0,-2}(y(t))\phi_{0,-1}(y(t)) \\ & + 1.222945\phi_{0,-1}(y(t))\phi_{0,-3}(y(t-1)) + 2.480638\phi_{0,-2}(y(t))\phi_{0,-4}(y(t-1)) \\ & - 0.565649\phi_{1,-5}(y(t))\phi_{1,-4}(y(t-1)) + \Phi(\varepsilon^t) \end{aligned} \tag{25}$$

where $\Phi(\varepsilon^t)$ is a collection of noise model terms relating to the moving average part of the model with time lags up to time t. The moving average noise model terms are necessary to avoid bias on the estimated parameters. These terms will be neglected when the model is used for iterative prediction. As a comparison, a polynomial NARMA model was also identified and is given by

$$\begin{aligned} y(t+1) = {} & -1.198787y(t) + 0.603279y(t-1) \\ & + 0.852159y^2(t)y(t-1) - 0.290071y^5(t) \\ & + 0.012730y^4(t)y(t-1) + 0.003551y(t)y^8(t-1) + \Psi(\varepsilon^t) \end{aligned} \tag{26}$$

where $\Psi(\varepsilon^t)$ was defined as in (25). The validity tests, over the validation data set, for the identified models (25) and (26) are shown in figures 1 and 2, and clearly indicate that the two models provide a sufficient description of the given data set. These two models will therefore be used for predictions in an iterative way, and will be compared with the results from the proposed direct approach in the next section.

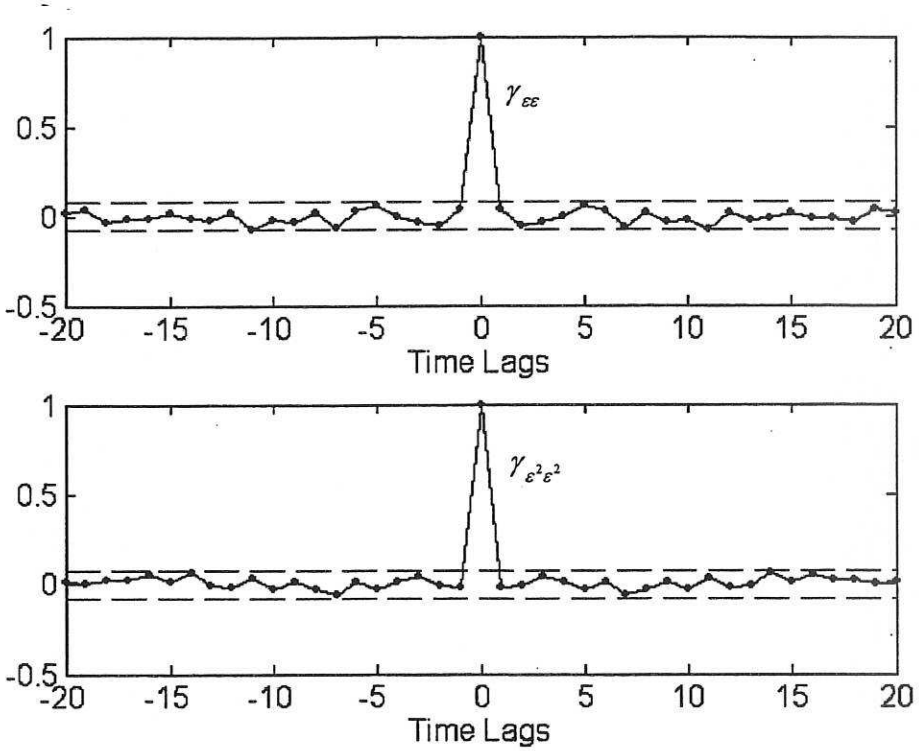


Figure 1 Model validity tests for the identified wavelet NARMA model (25).

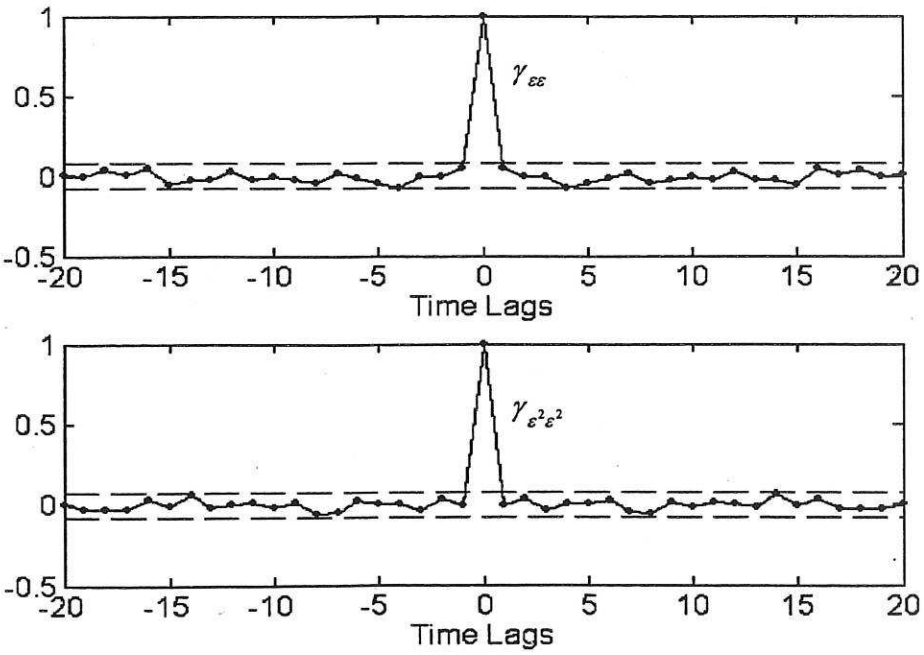


Figure 2 Model validity tests for the identified polynomial NARMA model (26).

5.1.2 Direct approach

To test the performance of direct models, a number of different direct wavelet models were identified starting from the initial wavelet NARMA model (24) by performing the procedure proposed in Section 4. The performance of 10 direct models for multi-step-ahead predictions corresponding to s=1,2,5, ..., 100 is shown in Table 1, where the standard derivations (STDs) and the normalized-mean-square-errors (NMSEs) are calculated over the validation test data set. For comparison, information on the associated iterative wavelet and polynomial NARMA models (25) and (26), together with different direct polynomial NARMA models, are also listed in Table 1. The initial polynomial NARMA model was restricted to second order and tenth degree of nonlinearity (here nonlinear degree is defined as the highest order among all the model terms, for example, model (26) is a second order, ninth degree nonlinear NARMA model). The 100-step-ahead-prediction result based on the identified direct wavelet model over the test data set is shown in figure 3, where only the data points from 800 to 1000 are plotted to allow a close inspection.

It can be concluded from Table 1 that: i) For a short prediction, for example one or two-step-ahead-predictions, the polynomial models are slightly advantageous over the wavelet models for the cases of both the direct and the iterative approaches; ii) for long term predictions, wavelet models are advantageous over polynomial models; iii) the direct approach is advantageous over the iterative approach and gives significantly improved prediction accurately for many step ahead predictions.

5.2 Example 2—a high-dimensional chaotic time series

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

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} \tag{27}$$

Table 1 Performance of identified wavelet and polynomial NARMA models for the system described by (23). (NMT—number of model terms included in the identified model)

	Direct wavelet NARMA models			Direct polynomial NARMA models			Iterative wavelet NARMA model (25)		Iterative polynomial NARMA model (26)	
Steps	NMT	STDs	NMSEs	NMT	STDs	NMSEs	STDs	NMSEs	STDs	NMSEs
1	12	0.0198	7.222×10^{-4}	6	0.0194	7.196×10^{-4}	0.0198	7.222×10^{-4}	0.0194	7.196×10^{-4}
2	11	0.0313	0.0028	8	0.0263	0.0023	0.0335	0.0023	0.0329	0.0021
5	14	0.0371	0.0026	10	0.0373	0.0026	0.0516	0.0049	0.0518	0.0051
10	15	0.0438	0.0045	12	0.0500	0.0047	0.0540	0.0053	0.0543	0.0056
20	15	0.0700	0.0097	14	0.0717	0.0098	0.0772	0.0105	0.0776	0.0107
40	12	0.0822	0.0111	10	0.0869	0.0144	0.1090	0.0221	0.1096	0.0229
60	16	0.1010	0.0189	9	0.1042	0.0206	0.1063	0.0212	0.1069	0.0217
80	18	0.1198	0.0224	12	0.1214	0.0235	0.1311	0.0295	0.1316	0.0298
100	17	0.1319	0.0270	11	0.1336	0.0281	0.1634	0.0489	0.1638	0.0510

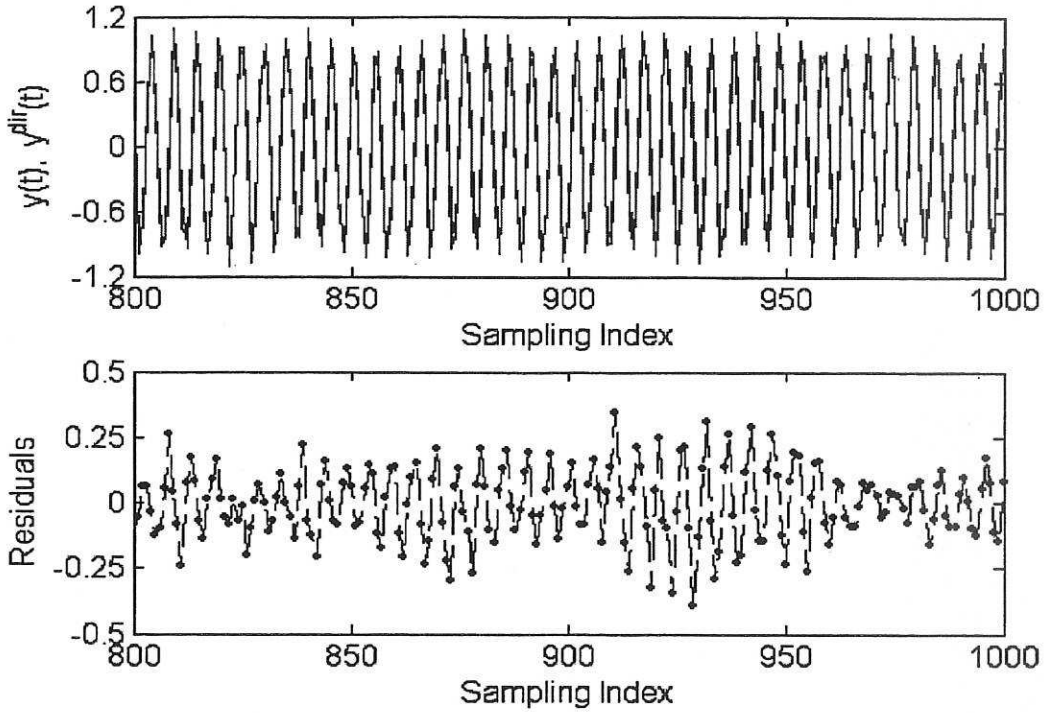


Figure 3 One hundred-step-ahead-predictions and prediction errors from the direct wavelet model for the system (23) over the validation data set of the points from 800 to 1000. (In the top figure, the solid line indicates the measurements and dashed line indicates 100-step-ahead direct predictions)

where the time delay τ was chosen to be 17 in this example. Setting the initial condition $x(t) = 0.9$ for $0 \le t \le \tau$, a Runge-Kutta integral algorithm was used to simulate Eq. (27) with an integral step Δt =0.01 and 2000 equi-spaced samples, $y(t)$, (t=1,2, …, 2000) were recorded after the time series settled down with a sampling interval of T =1 time units. A Gaussian white noise with a probability distribution $N(0,0.01^2)$ was added to the recorded data set. The first 1000 points were used for model estimation and the remaining 1000 data points were used for model validation.

Following Casdagli (1989), the embedding dimension of the recorded time series was assumed to be n=4, and the significant variables were therefore chosen to be $y(t)$, $y(t\text{-}1)$, $y(t\text{-}2)$ and $y(t\text{-}3)$. Long term predictions for chaotic time series are often impossible or inaccurate because of the local instability, i.e., the sensitive dependency on the initial conditions. It will be shown, however, that satisfactory long term predictions can be obtained for this system using the proposed direct approach. The initial wavelet-NARMA model was chosen to be

$$y(t+s) = P^{(s)}(y(t),\cdots,y(t-3)) + Q^{(s)}(\varepsilon^{(s)}(t),\cdots,\varepsilon^{(s)}(t-20))$$
$$= \sum_{d=0}^{3}\left\{\sum_{k}\alpha_{0,k}^{(d)}\phi_{0,k}(y(t-d)) + \sum_{j=0}^{5}\sum_{k}\beta_{j,k}^{(d)}\psi_{j,k}(y(t-d))\right\}$$

$$+\sum_{0\le p<q\le 3}\left\{\sum_{j=0}^{2}\sum_{k_1,k_2}\gamma_{j;k_1,k_2}^{(p,q)}\phi_{j;k_1}(y(t-p))\phi_{j;k_2}(y(t-q))\right\}$$

$$+\sum_{0\le p<q<r\le 3}\left\{\sum_{j=0}^{1}\sum_{k_1,k_2,k_3}\gamma_{j;k_1,k_2,k_3}^{(p,q,r)}\phi_{j;k_1}(y(t-p))\phi_{j;k_2}(y(t-q))\phi_{j;k_3}(y(t-r))\right\}$$

$$+\sum_{k=0}^{20}c_k\varepsilon(t-k) \tag{28}$$

where $\psi_{j,k}(x)=2^{j/2}\psi(2^j x-k)$ and $\phi_{j,k}(x)=2^{j/2}\phi(2^j x-k)$, $j,k\in Z$ are the 4th-order B-spline wavelet and associated scaling functions.

The case of 100-step-ahead-predictions was considered here, that is, s was set to 100 in (28). Starting from the initial model (28), a multiresolution wavelet-NARMA model was identified, which consisted of only 6 model terms

$$\begin{aligned}y(t+100)=&-0.044559172\psi_{2,-1}(y(t))-0.92268458\psi_{0,-1}(y(t-1))\\&+0.45156224\psi_{1,-3}(y(t-3))-0.17060753\psi_{3,8}(y(t-3))\\&+3.40163941\phi_{0,0}(y(t))\phi_{0,0}(y(t-1))\\&+2.02765699\phi_{0,-1}(y(t))\phi_{0,-1}(y(t-3))+\Phi(\varepsilon^t)\end{aligned} \tag{29}$$

where $\Phi(\varepsilon^t)$ was defined as in (25). Note that the variable $y(t\text{-}2)$ was not included in the identified model (29). This may be because this variable is not significant for 100-step-ahead-predictions using a wavelet model of this form. The performance of model (29) was checked by considering 100-step-ahead-predictions over the validation data set (points from 1000 to 2000) and this is shown in figure 4. The standard derivation and the normalized-mean-square-error over the validation data set were calculated to be 0.1277 and 0.3220, respectively. For comparison, a polynomial NARMA model with an order of 4 and nonlinear degree of 5 was also identified for 100-step-ahead direct predictions. The standard derivation and the normalized-mean-square-error over the validation data set for this identified polynomial NARMA model were calculated to be 0.1402 and 0.3632, respectively.

6. Conclusions

A direct modelling approach for long term predictions of nonlinear time series has been proposed by introducing multiresolution wavelet-based NARMA models. It has been shown from the examples that the proposed direct scheme provides better long term predictions compared with the corresponding results from traditional iterative models under the assumption that the evolution of a time series is governed by a complex deterministic law, which might be corrupted by noise. While in most cases it is difficult, if not impossible, to obtain satisfactory long term predictions for chaotic time series using an iterative scheme, the proposed direct approach provided good results for the examples considered in this study.

Compared with the traditional iterative approach, where only one model is required for both short and long term predictions, the main disadvantage of the direct modelling approach is that it requires different models for different step-ahead predictions. That is, different values of s in Eqs (9) or (12) will require different models in

each case. This will undoubtedly increase the computations associated with model building if many different step ahead-prediction values are required. If only one particular step ahead prediction is required, the procedure and computation associated with building a direct model is exactly the same as identifying an iterative model.

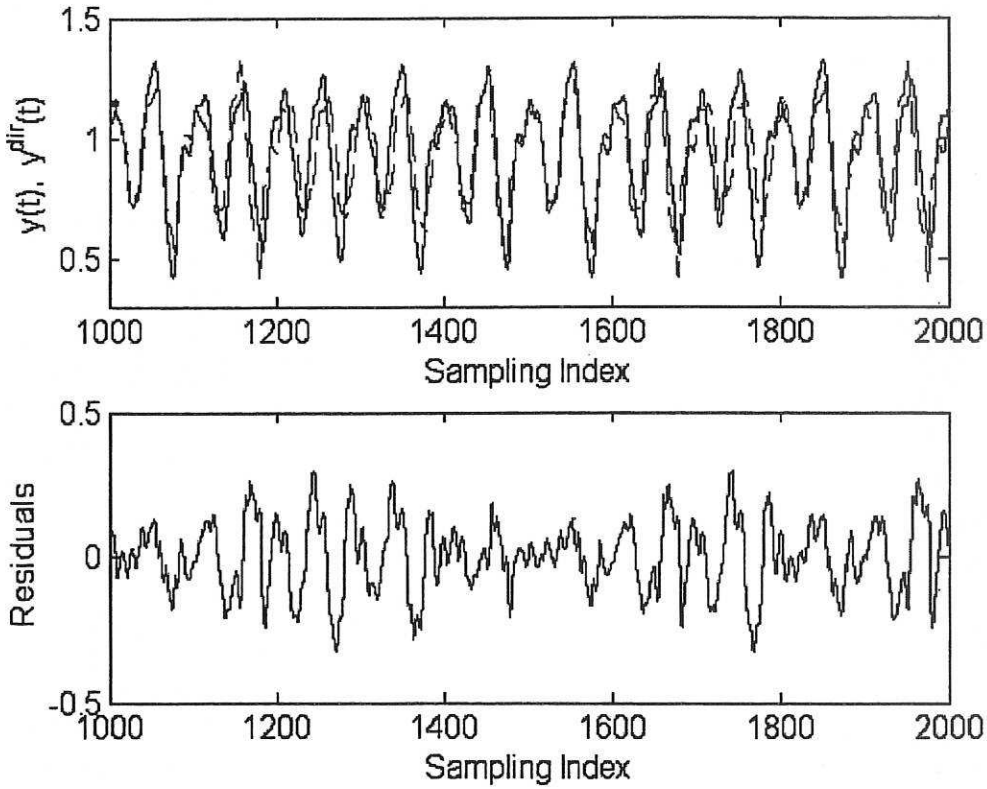


Figure 4 One hundred-step-ahead-predictions and prediction errors from the direct wavelet model for the system (27) over the validation data set of the points from 1000 to 2000. (In the top figure, the solid line indicates the measurements and dashed line indicates 100-step-ahead direct predictions)

Acknowledgment

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

References

Abarbanel, H.D.I., Brown, R. and Kadtke, J.B.(1989). Prediction and system identification in chaotic nonlinear systems: time series with broadband Fourier spectra. *Physics Letters A*, **138(8)**, 401-408.

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

Atiya, A., El-Shura, S.M. Shaheen, S. I. and El-Sherif, M.S. (1999). A comparison between neural-network forecasting techniques—case study: river flow forecasting. *IEEE Transactions on Neural Networks*, **10(2)**, 402-409.

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

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

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

Billings, S.A. and Tao, Q.H.(1991). Model validation tests for nonlinear signal processing applications. *International Journal of Control*, **54(1)**,235-244.

Billings, S.A. and Wei, H.L.(2005). The wavelet-NARMAX representation: a hybrid model structure combining polynomial models and multiresolution wavelet decompositions. (*Accepted for publication*).

Box, G.E.P. and Jenkins, G.M.(1976). *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

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. (1989). Modelling and analysis of nonlinear time series. *International Journal of Control*, **50(6)**, 2151-2171.

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.

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

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.

Farmer, J.D. and Sidorowich, J.J.(1987). Predicting chaotic time series. *Physical Review Letters*, **59(8)**, 845-848.

Ghanem, R. and Romeo, F. (2001). A wavelet-based approach for model and parameter identification of non-linear systems. *International Journal of Non-Linear Mechanics*, **36(5)**, 835-859.

Hill, T. , Marquez, L., O'Connor, M. and Remus, W. (1994). Artificial neural networks for forecasting and decision making. *International Journal of Forecasting*, **10(1)**, 5-15.

Jimenez, J., Moreno, J.A. and Ruggeri, G.J. (1992). Forecasting on chaotic time series: a local optimal linear reconstruction method. *Physics Reviewer A*, **45(6)**, 3553-3558,

Judd, K. and Small, M. (2000). Towards long-term prediction. *Physica D*, **136**, 31-44.

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

Lee, K.L. and Billings, S.A. (2003). A new direct approach of computing multi-step ahead predictions for nonlinear models. *International Journal of Control*, **76(8)**, 810-822.

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.

McNames, J. (2002). Local averaging optimization for chaotic time series prediction. *Neurocomputing*, **48**, 279-297.

Pandit, S.M. and Wu, S.M. (1983). *Time Series and System Analysis with Applications*. New York: Wiley & Sons.

Parlos, A.G., Rais, O.T. and Atiya, A.F. (2000). Multi-step-ahead prediction using dynamic recurrent neural networks. *Neural Networks*, **13(7)**, 765-786.

Principe, J.C., Rathie, A. and Kuo, J.M.(1992). Prediction of chaotic time series with neural networks and the issue of dynamic modelling. *International Journal of Bifurcation and Chaos*, **2(4)**, 989-996.

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

Sureshbabu, N. and Farrell, A. (1999). Wavelet-based system identification for nonlinear control. *IEEE Transactions on Automatic Control*, **44(2)**, 412-417.

Tang, Z., Almeida, C. and Fishwick, P.A. (1991). Time series forecasting using neural networks vs Box-Jenkins methodology. *Simulation*, **57(5)**, 303-310.

Tsatsanis, M.K., and Giannakis,G.B.(1993). Time-varying system identification and model validation using Wavelets. *IEEE Trans. Signal Processing,* **41(12)**,3512-3523.

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

Wei, H.L. and Billings, S.A. (2004). A unified wavelet-based modelling framework for nonlinear system identification: the WANARX model structure. *International Journal of Control*, **77(4)**, 351-366.

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.

Wei, H.L., Billings, S.A. and Balikhin, M.A.(2004). Prediction of the Dst index using multiresolution wavelet models. *Journal of Geophysical Research-Space Physics*, 109A, No.07212.

Weigend, A.S., Huberman, B.A. and Rumelhart, D.E. (1990). Predicting the future: a connectionist approach. *International Journal of Neural Systems*, **1**, 193-209.

Weigend, A.S., Huberman, B.A. and Rumelhart, D.E. (1992). Predicting sunspots and exchange rates with connectionist networks. In M. Casdagli and S. Eubank (Eds) *Nonlinear Modelling and Forecasting*, Redwood City, CA: Addison-Wesley, pp. 395-432.

Zhang G.Q., Patuow, B.E. and Hu, M.Y. (1998). Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting*, **14(1)**, 35-62.

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

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

Zheng, Y., Lin, Z. and Tay, D. B. H. (2001).Time-varying parametric system multiresolution identification by wavelets. *International Journal of Systems Science,* **32(6)**, 775-793.