Monograph:
Wei, H.L., Billings, S.A. and Zhao, Y. (2007) Generalised additive multiscale wavelet models constructed using particle swarm optimisation and mutual information for spatio-temporal evolutionary system representation. Research Report. ACSE Research Report no. 961 . Automatic Control and Systems Engineering, University of Sheffield

Generalised Additive Multiscale Wavelet Models Constructed Using Particle Swarm Optimisation and Mutual Information for Spatio-Temporal Evolutionary System Representation

H. L. Wei, S. A. Billings, Y Zhao

Research Report No. 961

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

July 2007

Generalised Additive Multiscale Wavelet Models Constructed Using Particle Swarm Optimisation and Mutual Information for Spatio-Temporal Evolutionary System Representation

H. L. Wei, S. A. Billings, Y Zhao

Department of Automatic Control and Systems Engineering

The University of Sheffield

Mappin Street, Sheffield

S1 3JD, UK

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

Abstract: A new class of generalised additive multiscale wavelet models (GAMWMs) is introduced for high dimensional spatio-temporal evolutionary (STE) system identification. A novel two-stage hybrid learning scheme is developed for constructing such an additive wavelet model. In the first stage, a new orthogonal projection pursuit (OPP) method, implemented using a particle swarm optimisation (PSO) algorithm, is proposed for successively augmenting an initial coarse wavelet model, where relevant parameters of the associated wavelets are optimised using a particle swarm optimiser. The resultant network model, obtained in the first stage, may however be a redundant model. In the second stage, a forward orthogonal regression (FOR) algorithm, implemented using a mutual information method, is then applied to refine and improve the initially constructed wavelet model. The proposed two-stage hybrid method can generally produce a parsimonious wavelet model, where a ranked list of wavelet functions, according to the capability of each wavelet to represent the total variance in the desired system output signal is produced. The proposed new modelling framework is applied to real observed images, relative to a chemical reaction exhibiting a spatio-temporal evolutionary behaviour, and the associated identification results show that the new modelling framework is applicable and effective for handling high dimensional identification problems of spatio-temporal evolution sytems.

Keywords: Coupled map lattices, evolutionary algorithms, generalised additive models, orthogonal least squares, parameter estimation, particle swarm optimisation, spatio-temporal evolutionary systems, wavelets.

1. Introduction

Spatio-temporal evolutionary (STE) systems, considered in this study, are a class of complex dynamical systems where the system states evolve spatially as well as temporally. Spatio-temporal evolutionary phenomena widely exist in various areas of science and engineering including biology, chemistry, ecology, geography, medicine, physics, and sociology (Kaneko 1993, Jahne 1993, Silva and Principe 1997, Astic et al. 1998, Bascompte and Sole 1998, Czaran 1998, Spors and Grinvald 2002, Dimitrova and Berezney 2002, Berezney et al. 2005, Dolak and Schmeiser 2005). To replicate, imitate, or analyse STE phenomena, several efficient representations, for example the well known cellular automata (CA) (Wolfram 1994, Ilachinski 2001), coupled map lattice (CML) models (Kaneko 1993), and cellular neural networks (CNNs) (Chua and Yang 1988a, 1988b, Chua and Roska 2002, have been proposed. In these representations, it is often assumed that the associated mathematical model structure, along with the model parameters, is known, so that the model can be used to describe or imitate some specific phenomena. However, the evolution law of real-world STE phenomena may not always be completely known, and relative evolution rules need to be acquired from observed data of relevant images or patterns. Hence, in recent years, identification problems of spatio-temporal systems have received much attention and interest from researchers in diverse fields, and several efficient identification methods and algorithms have been proposed, see for example Adamatzky and Bronnikov (1990), Adamatzky (1994, 1997), Parlitz and Merkwirth (2000), Sitz et al. (2003), Coca and Billings (2001), Mandelj et al. (2001), Billings and Coca (2002), Veenman et al. (2003), Billings and Yang (2003), Xia and Leung (2005), Billings et al. (2005).

One prominent feature of STE systems, compared with classical pure temporal signals or static images, is that there exists, in any given STE system, an inherent evolution law that determines the dynamical variation of relevant patterns with time. The individual value of a state at a local position of the current pattern, at the present time instant, is determined by individual values at several local positions of one or more previous patterns. In this sense, approaches for dealing with STE systems are often significantly different from those for processing classical pure temporal signals or static images, even though spatio-temporal approaches are also involved in some complex image processing, see for example Kim and J. Woods (1997, 1998), Ricquebourg and Bouthemy (2000), Sanchez-Marin et al. (2001), Caspi and Irani (2002), Ngo et al. (2003), Yang and Parvin (2003), and Nguyen et al. (2007). Compared to classical pure temporal signal modelling and static image processing, the identification and modelling of high dimensional STE systems are more challenging.

The central task in any STE system identification is to learn, from available observations of patterns or images of the relevant system, nonlinear models that can represent, as close as possible, the observed spatio-temporal evolution behaviours. The evolution law in real-world STE systems often involves many local state variables at past times and at different local positions, thus the identification procedure of STE systems may need to construct, based on available data, very high dimensional

nonlinear models containing a great number of 'input' variables. However, because of the *curse-of-dimensionality*, which is ubiquitously involved in any high dimensional nonlinear function learning and nonlinear modelling procedures, most existing STE system identification approaches can only handle low dimensional problems, where only a small size of neighbourhood and a very short time lag are considered.

Additive models and generalised additive models (GAMs) (Stone 1985, 1986, Buja et al. 1989, Hastie and Tibshirani 1990) are an important class of representations for high dimensional nonlinear signals. It follows that GAMs can not only avoid the curse of dimensionality, but also provide the ability to detect nonlinear dynamics and nonlinear patterns, without sacrificing interpretability of the relevant component functions. Also, GAMs, combined with other modelling techniques, have recently become extremely popular and have been widely applied in high dimensional data processing and modelling (Aerts et al. 2002, Ruppert et al. 2003, Wood 2004, Brezger and Lang 2006, Lado et al. 2006, Wei and Billings 2006a). Wavelet transforms (Daubechies 1992, Mallat 1998), due to their inherent properties and excellent capability for time-frequency domain representations of arbitrary signals (Unser 1995, Van De Ville et al. 2004), should be one of the best candidates to form the most powerful elementary building blocks to implement generalised additive models.

The construction of wavelet-based adaptive additive models may need to solve some nonlinear-in-the-parameters problems. Traditionally, Gaussian-Newton type nonlinear optimisation methods are often applied to estimate the unknown model parameters, with a stipulation that the gradients of the associated object functions are differentiable and easy to explicitly calculate. In this study, however, the recently developed particle swarm optimisation (PSO) algorithm (Eberhart and Kennedy 1995, Kennedy and Eberhart 1995) is employed as an alternative to solve complex nonlinear optimisation problems. Compared with classical nonlinear least squares algorithms, the PSO algorithm, as a population-based evolutionary method, possesses several desirable attractive properties, for example, this type of algorithm is easy to implement but quite efficient in dealing with a wide class of nonlinear optimisation problems. As a stochastic algorithm, PSO does not need any information on the gradients of the relevant object functions, this ensures that PSO is highly suitable for nonlinear optimisation problems where the relevant object functions are not differentiable or the gradients are computationally expensive or very difficult to obtain (van den Bergh 2002).

Starting with these observations, this study aims to introduce a novel class of generalised additive multiscale wavelet models (GAMWMs), which can be used to handle the identification problems of high dimensional STE systems. The construction procedure of the GAMWM is composed of two stages. At the first stage, a new constructive learning method, called the orthogonal projection pursuit (OPP), implemented with a particle swarm optimisation (PSO) algorithm, is used to form an initial coarse additive multiscale wavelet model by recruiting a number of optimised wavelets into the model in a stepwise manner. The OPP learning algorithm, which is in mechanism similar to conventional projection pursuit regression (Friedman and Stuetzle 1981), may produce a redundant model. Thus, at

the second stage, a forward orthogonal regression (FOR) learning algorithm (Billings and Wei 2007a, Wei and Billings 2007), implemented using a mutual information estimation method, is then applied to refine and improve the initially obtained wavelet model by removing redundant wavelet functions from the model.

As will be seen from the illustrative example, by combining the PSO based nonlinear OPP learning method with the effective mutual information aided FOR algorithm, the resultant additive wavelet model can provide very good representations for a class of high dimensional STE systems. One feature of the new GAMWM, produced by the above two-stage hybrid learning algorithm, is that now the resultant model is transparent to model users. Involved wavelets are ranked according to the capability in representing the total variance in the system output signal. This is desirable for many application cases where physical insight on the individual variables and associated model basis functions are of interest. Also, notice that the proposed GAMWM is nearly self-implemented, that is, all model parameters can automatically be adjusted by the proposed algorithms. This is desirable for any structure-unknown or black-box modelling problem. In summary, the main contribution of this work is that it provides, for the first time, an effective automatic and adaptive model identification approach for STE systems involving very high dimensional modelling procedures, by means of the proposed two stage hybrid constructive learning scheme that can produce sparse and transparent models with good generalisation properties.

This paper is organised as follows. In section 2, the general form of STE systems is briefly described. In section 3, the structure of the new GAMWM is presented. In section 4, a two-stage hybrid learning scheme, involving both the PSO based orthogonal projection pursuit approach and the mutual information aided forward orthogonal regression algorithm, is addressed in detail. In section 5, an example, relative to real observations for a chemical experiment, is presented to demonstrate the application of the new modelling framework. Some conclusions are given in section 6.

2. Spatio-Temporal Evolutionary Systems

The general form of spatio-temporal evolutionary (STE) systems is briefly introduced. In this study, the 2-D case, which has obvious physical meaning and is widely applied in practice, is taken as an example. For simplicity, only the zero-input (autonomous) class of STE systems is considered here. Model representations for these situations can easily be extended to other more complex cases in a straightforward way.

Assume that the 2-D image or pattern produced by an STE system, at the time instant t, consists of a $I \times J$ rectangular array of cells, $C^t(i, j)$, with Cartesian coordinates (i,j), i=1,2, …, I, j=1,2, …, J. Following Chua and Roska (2002), let $S_r^t(i, j)$ be the sphere of influence of the radius r of cell $C^t(i, j)$, at the time instant t, defined as

$$S_r^t(i,j)=\{C^t(i,j):\max_{1\le p\le I,1\le q\le J}\{|i-p|,|j-q|\}\le r\} \tag{1}$$

where t=1,2, …, i=1,2, …, I, j=1,2, …, J, and r is a non-negative integer number indicating how many neighborhood cells are involved in the evolution procedure. The sphere $S_r^t(i,j)$ is sometimes referred to as the $(2r+1)\times(2r+1)$ neighbourhood. Let $s_{i,j}(t)\in \mathrm{R}$ be the state variable representing the cell $C^t(i,j)\in S_r^t(i,j)$. From the definition of $S_r^t(i,j)$, a total of $(2r+1)^2$ state variables are involved in (1), see Table 1, where the symbol $C(i,j)$ will be used to indicate cells at arbitrary evolution time instants.

Table 1. The $(2r+1)\times(2r+1)$ neighbourhood defined by (1)

$C(i\text{-}r, j\text{-}r)$ x_1	…	$C(i\text{-}r, j)$ x_r	…	$C(i\text{-}r,j\text{+}r)$ x_{2r+1}
…	…	…	…	…
$C(i, j\text{-}r)$ $x_{r(2r+1)+1}$	…	$C(i,j)$ $x_{r(2r+1)+(r+1)}$	…	$C(i,j\text{+}r)$ $x_{(r+1)(2r+1)}$
	…		…	…
$C(i\text{+}r,j\text{-}r)$ $x_{2r(2r+1)+1}$	…	$C(i\text{+}r,j)$ $x_{2r(2r+1)+(r+1)}$	…	$C(i\text{+}r,j\text{+}r)$ $x_{(2r+1)\,(2r+1)}$

Let $s_{i,j}(t)$ be the (i,j)th cell to be updated at time t. A wide range of STE systems can be described by the discrete-time, discrete-space and continuous-state spatio-temporal difference equation of the form below

$$\begin{aligned} s_{i,j}(t) &= f(\mathbf{s}(t-1),\mathbf{s}(t-2),\mathrm{L}\ ,\mathbf{s}(t-n_{lag})) \\ &= f(s_{i-r,j-r}(t-1),\mathrm{L}\ ,s_{i,j}(t-1),\mathrm{L}\ ,s_{i+r,j+r}(t-1),\mathrm{L}\ , \\ &\quad s_{i-r,j-r}(t-2),\mathrm{L}\ ,s_{i,j}(t-2),\mathrm{L}\ ,s_{i+r,j+r}(t-2),\mathrm{L}\ , \\ &\quad s_{i-r,j-r}(t-n_{lag}),\mathrm{L}\ ,s_{i,j}(t-n_{lag}),\mathrm{L}\ ,s_{i+r,j+r}(t-n_{lag})) \end{aligned} \tag{2}$$

where f is some nonlinear function, n_{lag} is the time lag, defined as a positive integer, indicating how many past images or patterns are involved in the evolution procedure, and $\mathbf{s}(t-k)$ is the state vector formed by the $(2r+1)^2$ state variables relative to the patterns at the time instant (t-k) with k=1,2, …, n_{lag}, that is,

$$\mathbf{s}(t-k)=[s_{i-r,j-r}(t-k),\mathrm{L}\ ,s_{i,j}(t-k),s_{i+r,j+r}(t-k)] \tag{3}$$

Note that the general representation form (2) includes, as special cases, most typical coupled map

lattice models. For convenience of description, introduce d single-indexed variables $x_k(t)$ as below

$$\mathbf{x}(t) = [x_1(t), x_2(t), \mathrm{L}\ , x_d(t)] = [\mathbf{s}(t-1), \mathbf{s}(t-2), \mathrm{L}\ , \mathbf{s}(t-n_{lag})] \tag{4}$$

where $\mathbf{s}(t-k) = [x_{1+(k-1)(2r+1)^2}(t), \mathrm{L}\ , x_{k(2r+1)^2}(t)]$ for k=1,2, …, n_{lag} . For the case n_{lag} =1, the description (4) is shown in Table 1. Also, let $y(t)$ represent the state variable $s_{i,j}(t)$ corresponding to the central cell $C^t(i, j)$. Then, Eq. (2) becomes

$$y(t) = f(\mathbf{x}(t)) = f(x_1(t), x_2(t), \mathrm{L}\ , x_d(t)) \tag{5}$$

In conventional coupled map lattice models, the nonlinear function f in model (2) is often assumed to be known as some deterministic function, and the model is mainly used to imitate or produce some specific phenomena. However, for real-word complex STE systems, a pre-determined function f may not sufficiently characterise the underlying dynamics. It may be better to learn, from available real observations, an appropriate model for a given STE system.

The task of STE system identification is to construct, based on available data, a model that can represent, as close as possible, the observed evolution behaviour. Unlike constructing static models for typical data fitting, the objective of dynamical modelling is not merely to seek a model that fits the given data well, it also requires, at the same time, that the model should be capable of capturing the underlying system dynamics carried by the observed data, so that the resultant model can be used in simulation, analysis, and control studies.

Notice that equation (2) involves a total of $d = (2r+1)^2 n_{lag}$ variables. A large value for either r or n_{lag} will mean that a great number of variables may be involved in the representation (2). For example, let r=2 and n_{lag} =4, then a total of 100 variables will be involved. This means that the identification procedure of STE systems may require constructing a very high dimensional nonlinear model involving a great number of 'input' (or 'independent') variables. This property of STE systems prohibits most traditional identification and modelling frameworks that are suitable for classical pure temporal dynamical process modelling, and new identification approach for STE systems need to be developed.

3. The New Generalised Additive Multiscale Wavelet Model

The new generalised additive multiscale wavelet models (GAMWMs) are a special implementation of the typical generalized additive models, where the additive functional components are approximated using a family of multiscale wavelet functions. Starting with the discretisation of the wavelet transform, this section represents the architecture of the new GAMWMs.

3.1 Wavelet frames and wavelet series

Consider a wavelet family below

$$\psi^{a,b}(x) = a^{-1/2}\psi\left(\frac{x-b}{a}\right) \tag{6}$$

where $a \in \mathrm{R}^+$, $b \in \mathrm{R}$, and the mother wavelet ψ is admissible. The admissibility condition is depicted using the Fourier transform $\hat{\psi}(\xi)$ of the function ψ as $C_\psi = \int_{-\infty}^{\infty} \xi^{-1} |\hat{\psi}(\xi)|^2 d\xi < \infty$. It has been shown (Daubechies 1992) that for reasonable ψ, there exists a grid $G = \{(a_m, b_n) : a_m \in \mathrm{R}^+, b_n \in \mathrm{R}; m, n \in \mathrm{Z}\}$, such that the family $\psi_{m,n}(x) = a_m^{-1/2}\psi(a_m^{-1}x - b_n)$, with $(a_m, b_n) \in G$, constitute a frame for $L^2(\mathrm{R})$ (the space of all square integrable functions), with frame bounds *A*, *B*; that is, for all $g \in L^2(\mathrm{R})$

$$A \| f \|^2 \le \sum_{m,n} |< g, \psi_{m,n} >|^2 \le B \| f \|^2 \tag{7}$$

where the symbols '<, >' and '|| ||' denote the inner product and the norm, respectively, following the ordinary definitions. The fact that $\psi_{m,n}$, whose parameters are restricted to a grid *G*, constitute a frame for $L^2(\mathrm{R})$ can guarantee that for any $g \in L^2(\mathrm{R})$, there exits a sequence $\{c_{m,n} : m, n \in \mathrm{Z}\} \in l^2(\mathrm{Z}^2)$ (the set of all double square summable sequences of complex numbers indexed by integers) such that

$$g(x) = \sum_m \sum_n c_{m,n} \psi_{m,n}(x) \tag{8}$$

A special choice of the grid *G* is to let $a_m = a_0^m$, $b_n = n b_0 a_0^m$, with $a_0 > 1$, $b_0 > 0$. Daubechies (1992) gave a theoretical approach for calculating the wavelet coefficients $c_{m,n}$ in (8). For some very special choices of ψ and *G*, the family $\psi_{m,n}$ can constitute an orthogonal basis for $L^2(\mathrm{R})$. The most popular choice is $a_0 = 2$, $b_0 = 1$, for which there exists ψ, with good time-frequency localisation properties, such that $\psi_{m,n}(x) = 2^{-m/2}\psi(2^{-m}x - n)$ constitute an orthogonal basis for $L^2(\mathrm{R})$.

Notice that this study considers nonlinear spatio-temporal dynamical modeling problems, where relative observations are often sparse and where the independent (input) variables involved in the *dynamical* model are often formed by some variables representing the past states in time and at different spatial locations. This is different from a typical signal decomposition, where a given signal is represented using a *static* model formed by some wavelet-based elementary building blocks. For nonlinear dynamical modeling, the choice $a_0 = 2$ and $b_0 = 1$ may not usually be optimal. In fact, the choices of optimal values for a_0 and b_0 are still an open problem when wavelet decompositions are used for nonlinear dynamical modelling. One best alternative is perhaps to let the data speak for themselves, that is, to let the relevant observed data themselves adaptively and automatically choose the dilation and translation parameters.

3.2 The new GAMWM

Generalised additive models (GAMs) (Stone 1985, 1986, Buja et al. 1989, Hastie and Tibshirani 1990) provide an efficient approach for dealing with data fitting problems in some high dimensional space. GAMs can not only avoid the curse of dimensionality, but also provide the ability to detect nonlinear dynamics and nonlinear patterns, without sacrificing interpretability of the relevant component functions. A general representation of GAMs for a *d*-dimensional function *f* of the form (5) is given as

$$\begin{aligned} y(t) &= f(x_1(t), x_2(t), \mathrm{L}\ , x_d(t)) \\ &= f_0 + f_1(x_1(t)) + f_2(x_2(t)) + \mathrm{L}\ + f_d(x_d(t)) \end{aligned} \tag{9}$$

where f_0 is a constant, generally set to be the mean value of the desired 'output' signal $y(t)$, and $f_i(\cdot)$ are some univariate nonlinear functions that need to be identified. In practice, several specific functions, including splines and wavelets, have been introduced as the elementary building blocks to construct GAMs (Aerts et al. 2002, Ruppert et al. 2003, Wood 2004, Billings and Wei 2005, Wei and Billings 2006b).

In this study, wavelet frames will be adopted to represent these *d* functional components $f_i(\cdot)$ for i=1,2, …, d. It is assumed that these *d* functional components are square-integrable over the domain of interest for given data sets. Also, the constant term f_0 can be set to zero and thus can be omitted. If the constant term is different from zero for a given system, it can then be assimilated by one or more of the *d* functional components, which are approximated using multiscale wavelet decompositions.

While the wavelet decomposition (8), where the dilation and translation parameters a_m and b_n are predetermined and constricted to a grid *G*, can be applied to identify a generalised model (9) for some low dimensional identification problems (*d* is small), for very high dimensional identification problems (*d* is large), the employment of (8) may be undesirable and the resultant model may become intractable, because now the model may involve a large number of candidate wavelet basis functions, and data arranging and data storage for such a situation may become prohibited.

An alternative to overcome the above difficulty is to approximate each of the *d* functional components, using an adaptive wavelet decomposition, where the dilation and translation parameters are estimated by means of nonlinear optimisation. For simplicity of description, the balance factor $a^{-1/2}$ in (6) will be dropped, and the family of wavelets of the form

$$\psi^{(a,b)}(x) = \psi(x; a, b) = \psi(ax - b) \tag{10}$$

will be used as the elementary building blocks to approximate the ith functional component $f_i(\cdot)$ as

$$f_i(x_i(t)) = \sum_{k=1}^{m_i} c_{i,k} \psi(x_i(t); a_{i,k}, b_{i,k}) \tag{11}$$

where the parameters $a_{i,k}, b_{i,k}, c_{i,k}$ are estimated, from given data, using some nonlinear optimisation algorithm. The model (9) can now be written as

$$\begin{aligned} y(t) &= f(x_1(t), x_2(t), \mathrm{L}\ , x_d(t)) \\ &= \sum_{i=1}^{d} f_i(x_i(t)) = \sum_{i=1}^{d} \sum_{k=1}^{m_i} c_{i,k} \psi(x_i(t); a_{i,k}, b_{i,k}) \end{aligned} \tag{12}$$

The generalised additive multiscale wavelet model (GAMWM) given by (12) will be used, in this study, to represent high dimensional STE systems. The remaining issue is how to construct, from given data, such an additive wavelet model.

4. Constructing the New GAMWM

Inspired by the successful applications of the projection pursuit regression (PPR) (Friedman and Stuetzle 1981) and other constructive learning algorithms (Fahlman and Lebiere 1990, Jones 1992, Hwang et al. 1994, Kwok and Yeung 1997a, 1997b, Reed and Marks 1999), this study proposes a simple orthogonal projection pursuit (OPP) learning scheme, implemented by a particle swarm optimisation (PSO) algorithm. Similar to other constructive algorithms, models produced by the OPP algorithm may, however, be redundant. To remove or reduce redundancy, a forward orthogonal regression (FOR) learning algorithm (Billings and Wei 2007a, Wei and Billings 2007), implemented using a mutual information estimation method, is applied to refine and improve the initially generated model by the OPP algorithm.

Note that in the following, the inner product is defined for sampled vectors in *N*-dimensional Euclidian space, for example, the inner product of the two vectors $\mathbf{u} = [u(1), u(2), \mathrm{L}\ , u(N)]^T$ and $\mathbf{v} = [v(1), v(2), \mathrm{L}\ , v(N)]^T$ is defined as $< \mathbf{u}, \mathbf{v} >= \mathbf{u}^T \mathbf{v} = \sum_{k=1}^{N} u(k) v(k)$; this is different from that defined in (7), where the inner product is imposed to functions in $L^2(\mathrm{R})$.

4.1 The OPP algorithm for coarse model identification

The basic idea of the OPP algorithm for coarse model identification is to successively approximate the function *f* by progressively minimising approximation errors. At each step, the wavelet transform (10) is performed on each of the *d* involved model variables $\{x_1(t), x_2(t), \mathrm{L}\ , x_d(t)\}$, and the resultant wavelets are then used to approximate the same relevant "fake desired target signal", by minimising the approximation error using a PSO algorithm. A total of *d* individual wavelets, with optimised parameters, are involved at each step. But only one wavelet function, which produces the minimum approximation error, is included in the coarse model. In other words, only the most competent and competitive variable, whose wavelet transform, with optimised parameters, produces the best approximation is considered at each step.

Let $\mathbf{y} = [y(1), y(2),..., y(N)]^T \in \mathrm{R}^N$ be the vector of given observations of the output signal, $\mathbf{x}_k = [x_k(1), x_k(2),\mathrm{L}\ , x_k(N)]^T$ the vector of the observations for the *k*th input variable, with *k*=1,2, …, *d*. For any given $\boldsymbol{\theta} = [a, b, c]^T$, let $\boldsymbol{\psi}_k = \psi(\mathbf{x}_k; a, b)$ and $\mathbf{g}(\mathbf{x}_k; \boldsymbol{\theta}) = c\boldsymbol{\psi}_k$.

The OPP algorithm is implemented in a stepwise fashion; at each step a construction vector that minimises the projection error will be determined. Starting with $\mathbf{r}_0 = \mathbf{y}$, let $\Theta_{0,k} = \arg\min_{\boldsymbol{\theta}}\{\| \mathbf{r}_0 - \mathbf{g}(\mathbf{x}_k; \boldsymbol{\theta}) \|^2\}$ and $J_{0,k} = \| \mathbf{r}_0 - \mathbf{g}(\mathbf{x}_k; \Theta_{0,k}) \|^2$. The first construction vector can then be chosen as $\mathbf{g}_1 = \mathbf{g}(\mathbf{x}_{1_1}; \boldsymbol{\theta}_{1_1})$, where $1_1 = \arg\min_k\{J_{0,k}\}$ and $\boldsymbol{\theta}_{1_k} = \Theta_{0,1_k}$. The residual vector, which can be used as the "fake desired target signal" to produce the second construction vector $\mathbf{g}_2$, is defined as $\mathbf{r}_1 = \mathbf{r}_0 - \alpha_1\mathbf{g}_1$, where $\alpha_1 = < \mathbf{r}_0, \mathbf{g}_1 > / \| \mathbf{g}_1 \|^2$. It can be shown that the residual vector $\mathbf{r}_1$ is orthogonal with the relevant construction vector $\mathbf{g}_1$.

Assume that at the (*n*-1)th step, a total of (*n*-1) construction vectors $\mathbf{g}_j = \mathbf{g}(\mathbf{x}_{1_j}; \boldsymbol{\theta}_{1_j})$, with *j*=1,2, …, *n*-1, have been obtained. Let $\mathbf{r}_{n-1}$ be the associated residual vector, $\Theta_{n,k} = \arg\min_{\boldsymbol{\theta}}\{\| \mathbf{r}_{n-1} - \mathbf{g}(\mathbf{x}_k; \boldsymbol{\theta}) \|^2\}$ and $J_{n,k} = \| \mathbf{r}_{n-1} - \mathbf{g}(\mathbf{x}_k; \Theta_{n,k}) \|^2$. The *n*th construction vector can then be given by $\mathbf{g}_n = \mathbf{g}(\mathbf{x}_{1_n}; \boldsymbol{\theta}_{1_n})$, where, $1_n = \arg\min_k\{J_{n,k}\}$ and $\boldsymbol{\theta}_{1_k} = \Theta_{n,1_k}$.

The associated residual vector can be defined as

$$\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{g}_n \tag{13}$$

where

$$\alpha_n = \frac{< \mathbf{r}_{n-1}, \mathbf{g}_n >}{\| \mathbf{g}_n \|^2} \tag{14}$$

From (14),

$$\| \mathbf{r}_n \|^2 = \| \mathbf{r}_{n-1} \|^2 - \frac{< \mathbf{r}_{n-1}, \mathbf{g}_n >^2}{\| \mathbf{g}_n \|^2} \tag{15}$$

By respectively summing (13) and (15) for *n* from 2 to *m*+1, yields

$$\mathbf{y} = \sum_{n=1}^{m} \frac{< \mathbf{r}_{n-1}, \mathbf{g}_n >}{\| \mathbf{g}_n \|^2} \mathbf{g}_n + \mathbf{r}_m = \sum_{n=1}^{m} \alpha_n \mathbf{g}_n + \mathbf{r}_m \tag{16}$$

$$\| \mathbf{r}_m \|^2 = \| \mathbf{y} \|^2 - \sum_{n=1}^{m} \frac{< \mathbf{r}_{n-1}, \mathbf{g}_n >^2}{\| \mathbf{g}_n \|^2} \tag{17}$$

The residual sum of squares, also called the sum of squared error, $\| \mathbf{r}_n \|^2$, can be used to form a criterion to stop the growing procedure. For example, the criterion can be chosen as the *error-to-signal*

ratio: ESR $=\| \mathbf{r}_n \|^2 / \| \mathbf{y} \|^2$; when ESR becomes smaller than a pre-specified threshold value, the growing procedure can then be terminated.

Now the OPP algorithm can briefly be summarised as follows.

The OPP algorithm:

Initialisation: $\mathbf{r}_0 = \mathbf{y}$; ESR=0;

while { ESR $\geq \eta$ or $n \leq m$PEM }; //{η is a pre-specified very small threshold value.}//

//{mOPP is the maximum number of construction functions // permitted to be included in the network} //

for n=1 to mOPP

for k=1 to d

//{Starting from some random (but reasonable) value for the // parameter vector $\boldsymbol{\theta}$, optimise the following function using // the PSO algorithm.}//

$$\Theta_{n,k} = \arg\min_{\boldsymbol{\theta}} \left\{ \| \mathbf{r}_{n-1} - \mathbf{g}(\mathbf{x}_k; \boldsymbol{\theta}) \|^2 \right\};$$

$$J_{n,k} = \| \mathbf{r}_{n-1} - \mathbf{g}(\mathbf{x}_k; \Theta_{n,k}) \|^2;$$

end for (k)

$$1_n = \arg\min_k \{ J_{n,k} \};$$

$$\boldsymbol{\theta}_{1_n} = \Theta_{n,1_n};$$

$$\mathbf{g}_n = \mathbf{g}(\mathbf{x}_{1_n}; \boldsymbol{\theta}_{1_n});$$ // The involved wavelet is $\psi(x_{1_n}; a_{1_n}, b_{1_n}) = \psi(a_{1_n} x_{1_n} - b_{1_n})$ //

$$\alpha_n = \frac{< \mathbf{r}_{n-1}, \mathbf{g}_n >}{\| \mathbf{g}_n \|^2};$$

$$\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{g}_n;$$

ESR $=\| \mathbf{r}_n \|^2 / \| \mathbf{y} \|^2$;

end for (n)

end while

It is clear from (15) that the sequence $\| \mathbf{r}_n \|^2$ is strictly decreasing and positive; thus, by following the method given in Kwok and Yeung (1997b) and Huang et al. (2006), it can easily be proved that the residual $\mathbf{r}_n$ is a Cauchy sequence, and as a consequence, the residual $\mathbf{r}_n$ converges to zero. The algorithm is thus convergent. The above OPP algorithm is in structure similar to the projection pursuit regression (Friedman and Stuetzle 1981) and other constructive learning algorithms (Mallat and Zhang 1993, Hwang et al. 1994, Kwok and Yeung 1997a, 1997b), however the implementation of the OPP algorithm is totally different from these existing algorithms. For example, in the projection pursuit regression method, the construction functions are nonparametric and in general unknown before hand; in the OPP algorithm, however, the construction functions are formed by a family of wavelets. In the matching pursuit method, the construction functions are restricted to a specified dictionary, where relevant adjustable parameters of individual candidates are permitted to vary on a given grid, while in the OPP algorithm no such limits are imposed on construction functions. Moreover, in the OPP algorithm, the elementary building blocks are some wavelets, where unknown parameters are optimised by using some PSO algorithm that does not need any information on the gradients of the object functions, this enables the PSO to be very suitable for nonlinear optimisation problems where

the relevant object functions are not differentiable or the gradients are computationally expensive or difficult to obtain. However, like the projection pursuit regression and the matching pursuit algorithms, the OPP algorithm may produce redundant models. To refine and improve the OPP produced network models, the forward orthogonal regression (FOR) learning algorithm, assisted by a mutual information method (Billings and Wei 2007a), is then applied to remove any severe redundancy.

4.2 The PSO algorithm for parameter optimisation

Particle swarm optimisation (PSO), originally inspired by some sociological behaviour associated with, for example, bird flocking (Kennedy et al. 2001), is a population-based stochastic optimisation algorithm that was first proposed by Kennedy and Eberhart in 1995 (Kennedy and Eberhart 1995, Eberhart and Kennedy 1995). In PSO, the population is referred to as a *swarm*, while the individuals are referred to as *particles*; each particle moves, in the search space, with some random *velocity*, and remembers and retains the *best position* it has ever been. The mechanism of PSO can succinctly be explained as follows. The position of each particle can be viewed as a possible solution to a given optimization problem. In each iteration (one step move), each particle accelerates its move toward a new potential position, by adaptively using information about its own *personal best position* obtained so far, as well as the information of the *global best position* achieved so far by any other particles in the swarm. Thus, if any promising new position is discovered by any individual particle, then all the other particles will move closer towards it. In this way, PSO will finally find, in an iterative manner, a best solution to the given optimisation problem (Parsopoulos and Vrahatis 2004, van den Bergh and Engelbrecht 2004).

Now consider an s dimensional optimisation problem, where the relevant parameter vector to be optimised is denoted by $\boldsymbol{\theta} = [\theta_1, \theta_2, \mathrm{L}\ , \theta_s]^T \in \Theta \subset \mathrm{R}^s$. Assume that a total of L particles are involved in the relevant swarm. Denote the position of the ith particle at the present time t by $\boldsymbol{\theta}_i(t)$, the relative velocity by $\mathbf{v}_i(t)$, the personal best position by $\mathbf{p}_i(t)$, and the global best position obtained so far by $\mathbf{p}_g(t)$. Following Kennedy et al. (2001), Shi and Eberhart (1998a, 1998b), Clerc and Kennedy (2002), PSO can be implemented using the iterative equations below

$$\mathbf{v}_i(t+1) = \chi\{\mathbf{v}_i(t) + c_1 r_1[\mathbf{p}_i(t) - \boldsymbol{\theta}_i(t)] + c_2 r_2[\mathbf{p}_g(t) - \boldsymbol{\theta}_i(t)]\} \tag{18a}$$

$$\boldsymbol{\theta}_i(t+1) = \boldsymbol{\theta}_i(t) + \mathbf{v}_i(t+1) \tag{18b}$$

where i=1,2, …, L; c_1 and c_2 are the acceleration coefficients, also referred to as the cognitive and social parameters; $\chi = 2/|2 - \phi - \sqrt{\phi^2 - 4\phi}|$, with $\phi = c_1 + c_2 > 4$, is a constriction factor used to obtain good convergence performance by controlling explosive particle movements; r_1 and r_2 are random numbers that are uniformly distributed in [0,1]. Typical choices for c_1 and c_2 are to set $c_1 = c_2 = 2$ (Kennedy and Eberhart 1995, Eberhart and Kennedy 1995).

Let $\pi(\boldsymbol{\theta})$ be the function that needs to be minimised, then the personal best position of each particle can be updated as below (van den Bergh and Engelbrecht 2004)

$$\mathbf{p}_i(t+1) = \begin{cases} \mathbf{p}_i(t), & \text{if } \pi(\boldsymbol{\theta}_i(t+1)) \geq \pi(\mathbf{p}_i(t)) \\ \boldsymbol{\theta}_i(t+1), & \text{if } \pi(\boldsymbol{\theta}_i(t+1)) < \pi(\mathbf{p}_i(t)) \end{cases} \tag{19}$$

While the global best position achieved by any particle during all previous iterations is defined as

$$\mathbf{p}_g(t+1) = \arg\min_{\mathbf{p}_i} \pi(\mathbf{p}_i(t+1)), \qquad 1 \leq i \leq L. \tag{20}$$

In the OPP algorithm discussed in the previous section, the objective function is defined as

$$\pi_{n-1}(\boldsymbol{\theta}) = \| \mathbf{r}_{n-1} - \mathbf{g}(\mathbf{x}_k; \boldsymbol{\theta}) \|^2 = \sum_{t=1}^{N} [r_{n-1}(t) - g(x_k(t); \boldsymbol{\theta})]^2 \tag{21}$$

where $g(x_k(t); \boldsymbol{\theta}) = \theta_3 \psi(\theta_1 x(t) - \theta_2)$ and N is the number of training samples.

With regard to the termination of the optimisation procedure, the criterion can be chosen as follows. Let 'mPSO' be the maximum number of permitted iterations. The optimization procedure can then be terminated when either the iteration index exceeds 'mPSO', or when the parameter to be optimized becomes stable, that is, when $\| \boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t) \|^2 \leq \delta$, where δ is a pre-specified small number, say $\delta \leq 10^{-5}$.

4.3 The FOR algorithm for model refinement

Assume that a total of m_j wavelets of the form $\psi_{j,k}(x_i) = \psi(x_j; a_{j,k}, b_{j,k}) = \psi(a_{j,k} x_i - b_{j,k})$, with k=1,2, …, m_j, are involved for the jth 'input' variable x_j, after having performed the PSO based OPP procedure on the given data set. The number of involved wavelets for all the d variables is then $M = m_1 + m_2 + \mathrm{L} \ + m_d$. Denote the set of these M wavelets by

$$\Omega = \{\psi_{j,k} : \psi_{j,k}(x_j) = \psi(x_j; a_{j,k}, b_{j,k}), (j,k) \in \Gamma\} \tag{22}$$

where $\Gamma = \{(j,k) : j = 1,2,\mathrm{L}\ , d; k = 1,2,\mathrm{L}\ , m_j\}$. Note that all the parameters $a_{j,k}$ and $b_{j,k}$ have already been estimated at the coarse model identification procedure. Experience shows that the set Ω may be redundant, and a refinement procedure thus needs to be performed to produce a parsimonious model.

The objective of this refinement stage is to reselect the most significant wavelet functions from the set Ω, to form a more compact model for a given nonlinear identification problem. Let $\mathbf{y}$ and $\mathbf{x}_k$ be defined as in previous sections, and let $\boldsymbol{\psi}_{j,k} = \psi(\mathbf{x}_j; a_{j,k}, b_{j,k})$, where $(j,k) \in \Gamma$. Also, let D be a set that exactly consisting of the M wavelet vectors $\boldsymbol{\psi}_{j,k}$ in Ω, with $(j,k) \in \Gamma$, that is,

$$D = \{\boldsymbol{\varphi}_i : \boldsymbol{\varphi}_i \in \Omega, i = 1,2,\mathrm{L}\ , M\} \tag{23}$$

The model refinement problem amounts to finding, from the vector dictionary D, a full dimensional subset $D_m = \{\mathbf{p}_1, \mathrm{L}\ , \mathbf{p}_m\} = \{\boldsymbol{\varphi}_{i_1}, \mathrm{L}\ , \boldsymbol{\varphi}_{i_m}\}$, where $\boldsymbol{\alpha}_k = \boldsymbol{\varphi}_{i_k}$, $i_k \in \{1,2,\mathrm{L}\ , M\}$ and k=1,2, …, m (generally $m << M$), so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\mathbf{p}_1, \mathbf{p}_2, \mathrm{L}\ , \mathbf{p}_m$ as below

$$\mathbf{y} = \beta_1 \mathbf{p}_1 + \mathrm{L}\ + \beta_m \mathbf{p}_m + \mathbf{e}_m \tag{24}$$

where $\mathbf{e}_m$ is the associated model residual vector.

The orthogonal least squares (OLS) type algorithms (Billings et al. 1989, Chen et al. 1989) can be used to determine model basis functions (model terms). In this study, however, a variation of the OLS algorithm, called the forward orthogonal regression (FOR) algorithm, implemented using a mutual information method (Billings and Wei 2007a, Wei and Billings 2007), is employed for the model refinement. Assume that $\mathbf{x}$ and $\mathbf{y}$ are two random discrete variables, with alphabet $\mathtt{X}$ and $\mathtt{Y}$, respectively, and with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(\mathbf{x}, \mathbf{y})$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, given as (Cover and Thomas 1991)

$$I(\mathbf{x}, \mathbf{y}) = \sum_{x \in \mathtt{X}} \sum_{y \in \mathtt{Y}} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{25}$$

The mutual information $I(\mathbf{x}, \mathbf{y})$ is the reduction in the uncertainty of $\mathbf{y}$ due to the knowledge of $\mathbf{x}$, and vice versa. Mutual information provides a measure of the amount of information that one variable shares with another one. If $\mathbf{y}$ is chosen to be the system output (the response), and $\mathbf{x}$ is one regressor in a linear model, $I(\mathbf{x}, \mathbf{y})$ can then be used to measure the coherence of $\mathbf{x}$ with $\mathbf{y}$ in the model. Several algorithms have been proposed to estimate mutual information from observed data, see for example Moddemeijer (1989, 1999), Darbellay and Vajda (1999), and Paninski (2003) and the references therein.

Detailed discussions on the utility of the mutual information for model term selection can be found in Billings and Wei (2007a) and Wei and Billings (2007). Now, let $\mathbf{p}_1, \mathbf{p}_2, \mathrm{L}\ , \mathbf{p}_n$ be the n selected linearly independent basis vectors after the nth step search, and let $\mathbf{q}_1, \mathbf{q}_2, \mathrm{L}\ , \mathbf{q}_n$ be a group of orthogonal vectors, generated from the vectors $\mathbf{p}_1, \mathbf{p}_2, \mathrm{L}\ , \mathbf{p}_n$, by means of some orthogonal transformation. Following Billings et al. (1989), Chen et al. (1989), the error reduction ratio (ERR), produced by including the nth basis vector $\mathbf{q}_n$, or equivalently by including $\mathbf{p}_n$, is defined as

$$\mathrm{ERR}_n = \frac{\gamma_n^2 \| \mathbf{q}_n \|^2}{\| \mathbf{y} \|^2} \tag{26}$$

where $\gamma_n = < \mathbf{y}, \mathbf{q}_n > / \| \mathbf{q}_n \|^2$. ERR can be used to measure the significance of individual model terms in that it provides an index indicating the contribution made by each selected individual model term to explain the total variance in the desired output signal.

Let $\mathbf{e}_n$ be the residual vector produced at the nth search step. Similar to in the OPP algorithm, the model residual vector $\mathbf{e}_n$ can be used to form a criterion to terminate the search procedure. Following the suggestion in Billings and Wei (2007b), the following adjustable prediction error sum of squares (APRESS), also referred to as the adjustable generalised cross-validation (AGCV), will be used to monitor the regressor search procedure

$$\text{APRESS}_n = (1 - \lambda n / N)^2 \text{MSE}(n) \tag{27}$$

where $\text{MSE}(n) = \| \mathbf{e}_n \|^2 / N$ is the mean-square-error that is associated to the model of n model terms. The number of regressors (wavelet functions) will be chosen as the value where APRESS arrives it minimum. Billings and Wei (2007b) suggest that the adjustable parameter λ be chosen between 5 and 10.

Following Billings and Wei (2007a) and Wei and Billings (2007), the mutual information based forward orthogonal regression (FOR) algorithm, is briefly summarised below.

The FOR-MI algorithm:

Step 1**:** Set $U_1 = \{1, 2, \text{L}\ , M\}$;

for j=1 to M

$\mathbf{q}_j^{(1)} = \boldsymbol{\varphi}_j$;

$I^{(1)}[j] = MI(\mathbf{r}_0, \mathbf{q}_j^{(1)})$; // Calculate the mutual information for all // candidate basis vectors.//

end for

$l_1 = \arg\max_{i \in U_1} \{I^{(1)}[i]\}$; $V_1 = \{l_1\}$;

$\mathbf{p}_1 = \boldsymbol{\varphi}_{l_1}$; $\mathbf{q}_1 = \mathbf{p}_1$; $\gamma_1 = \dfrac{< \mathbf{y}, \mathbf{q}_1 >}{\| \mathbf{q}_1 \|^2}$; $\mathbf{r}_1 = \mathbf{r}_0 - \gamma_0 \mathbf{q}_1$;

$\text{ERR}[1] = \dfrac{\gamma_1^2 \| \mathbf{q}_1 \|^2}{\| \mathbf{y} \|^2}$; $\text{APRESS}[1] = \dfrac{1}{(1 - \lambda / N)^2} \dfrac{\| \mathbf{r}_1 \|^2}{N}$;

Step n, $n \geq 2$ **:**

For n=2 to M

$U_n = U_{n-1} \setminus V_{n-1}$;

for $j \in U_n$

$\mathbf{q}_j^{(n)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{n-1} \dfrac{< \boldsymbol{\varphi}_j, \mathbf{q}_k >}{\| \mathbf{q}_k \|^2} \mathbf{q}_k$;

$I^{(n)}[j] = MI(\mathbf{r}_{n-1}, \mathbf{q}_j^{(n)})$; //Calculate the mutual information for all // for all candidate basis vectors.// //{if $\| \mathbf{q}_j^{(n)} \|^2 \leq \varepsilon$, set $I^{(n)}[j] = 0$ }//

end for (end loop for j)

$$1_n = \arg\max_{j \in U_n}\{I^{(n)}[j]\};\ V_n = \{1_n\} \cup \{\arg_{j \in U_n}(\| \mathbf{q}_j^{(n)} \|^2 < \varepsilon)\};$$

$$\mathbf{p}_n = \boldsymbol{\varphi}_{1_n};\quad \mathbf{q}_n = \mathbf{q}_{1_n}^{(n)};\quad \gamma_n = \frac{<\mathbf{y}, \mathbf{q}_n>}{\| \mathbf{q}_n \|^2};\ \mathbf{r}_n = \mathbf{r}_{n-1} - \gamma_n \mathbf{q}_n;$$

$$\mathrm{ERR}[n] = \frac{\gamma_n^2 \| \mathbf{q}_n \|^2}{\| \mathbf{y} \|^2};\ \mathrm{APRESS}[n] = \frac{1}{(1-\lambda n/N)^2}\frac{\| \mathbf{r}_n \|^2}{N};$$

for k=1 to n

$$r_{k,n} = \frac{<\mathbf{p}_n, \mathbf{q}_k>}{\| \mathbf{q}_k \|^2},\ \text{for } k < n;\ r_{k,n} = 1,\ \text{for } k = n;$$

end for (end loop for k)

end for (end loop for n)

The FOR algorithm provides an effective tool for successively selecting significant model terms (basis functions) in supervised learning problems. Terms are selected step by step, one term at a time. The inclusion of redundant bases, which are linearly dependent on the previous selected bases, can be efficiently excluded by eliminating the candidate basis vectors for which $\| \mathbf{q}_j^{(n)} \|^2$ are less than a predetermined threshold ε, say $\varepsilon \le 10^{-10}$. Assume that a total of m significant vectors are selected, then the unknown parameter $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_m]^T$, relative to the model (24), can easily be calculated from the triangular equation $\mathbf{R}\boldsymbol{\beta} = \boldsymbol{\gamma}$, where R is an upper triangular matrix and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_m]^T$ with $\gamma_i = <\mathbf{y}, \mathbf{q}_i> / \| \mathbf{q}_i \|^2$ for i=1,2,…, m.

5. Application in Chemical Reaction Modelling

The new wavelet based additive modelling framework can be applied to identify some SPE phenomena, where the true models are unknown and the initial candidate models may thus involve a great number of 'input' or 'independent' variables. To illustrate the application of the new modelling procedure, the Belousov-Zhabotinsky (Belousov 1959, Zhabotinsky 1964, Winfree 1972, Kuramoto 1984) reaction was considered here as an example.

The BZ reaction, as an excitable medium, is an important class of chemical reactions exhibiting a spatio-temporal oscillatory behaviour. As a classical example of nonequilibrium thermodynamics, the BZ reaction provides an interesting chemical model of nonequilibrium biological phenomena, and the modelling and identification of these type of reactions is of extreme interest for theoretical analysis of relevant phenomena.

By adopting the recipe given by Winfree (1972), an experiment resulting in a thin layer BZ reaction was carried out in the laboratory, and a set of images were captured (sampled) with equal time intervals during the experiment, using a digital video camera that is connected to a PC via a USB socket. The sampled images were processed and saved as patterns with a resolution of 360 by 500 pixels. Some of these patterns are shown in Fig. 1.

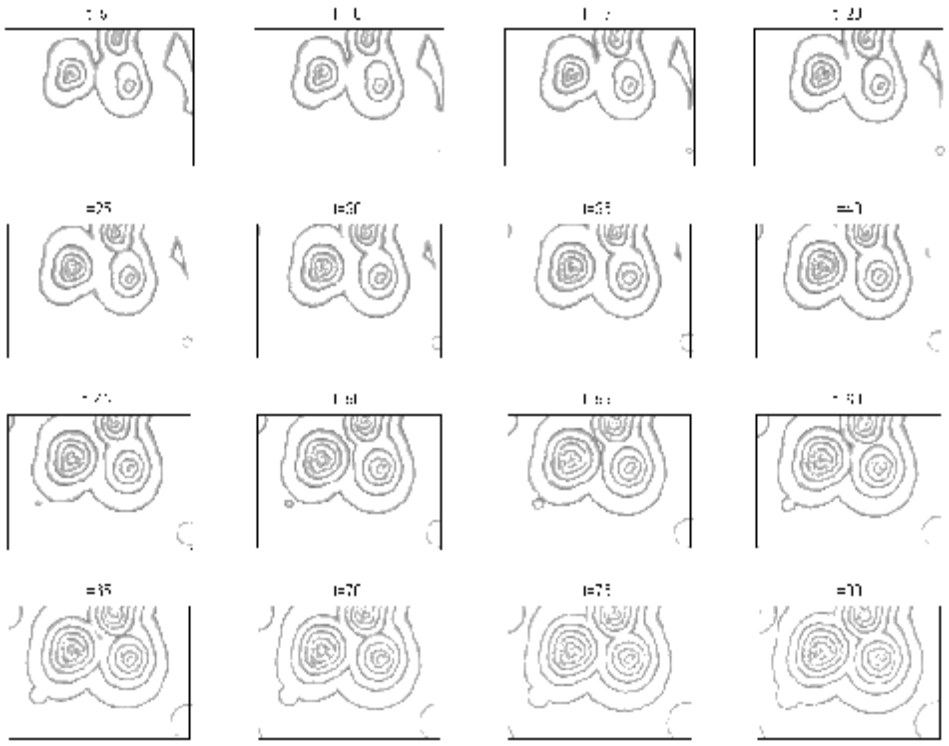

Fig. 1 Some snapshots for the BZ reaction at different time instants. The size of each template is 360×500 (360 pixels in the vertical direction and 500 pixels in the horizontal direction).

The proposed GAMWM modelling framework was applied to these sampled images, and the objective was to identify a mathematical model for the BZ reaction. Details of the identification procedure are given below.

5.1 The initial models and the training data

Consider the model of the form (2), where the number of total model variables is determined by two factors: the radius of the neighbourhood, r, and the time lag, n_{lag}. Two cases were considered here: r=1, n_{lag}=4, and r=2, n_{lag}=4. The initial models, corresponding to these two cases, thus involve a total of $(2\times1+1)^2\times4=36$ and $(2\times2+1)^2\times4=100$ model variables, respectively. For the later case, most existing identification approaches may be prohibited and thus are not applicable. The proposed GAMWM modelling framework, however, can effectively solve the identification problem associated with such a situation.

The state variable $s_{i,j}(t)$, at the present time instant t, was initially assumed to be associated with state variables in the past four adjacent neighbourhoods at the previous time instants t-1, t-2, t-3, and t-4. Any five patterns, at the abutting time instants t, t-1, t-2, t-3, t-4, are called an adjacent pattern group. For arbitrary time instant, the data pair, $\{\mathbf{x}(t), y(t)\}$, where $\mathbf{x}(t)$ and $y(t)$ are defined by (4) and (5), is called a data pair. Notice that $\mathbf{x}(t)$ and $y(t)$ are also implicitly associated with the spatial location indices i and j (see Table 1). As a consequence, for any given time instant t, there may be a large number of data pairs.

Training data sets, for the two initial models, corresponding to the two cases r=1, n_{lag} =4, and r=2, n_{lag} =4, were independently formed as follows. Firstly, 8 adjacent pattern groups were randomly chosen from the first 80 sampled patterns. Secondly, 500 data pairs were randomly chosen in each of these 8 adjacent pattern groups. The resultant two data sets, consisting of a total of N=4000 data pairs, $\{\mathbf{x}(k), y(k)\}_{k=1,2,\ldots,N}$, were then used for model identification, where $y(k)$ represents the value of the relevant central cell at the present time instant, and $\mathbf{x}(k) = [x_1(k), x_2(k), \mathrm{L}\ , x_d(k)]^T$ represent the values of the d involved cells at a squared lattice, at the previous time instants. Note that for the first case, d=36, and for the second case, d=100.

5.2 Identification results

The construction function was chosen to be the cardinal B-spline function of order 2, which is defined as

$$\psi(x) = \begin{cases} 1+x, & \text{for } -1 \le x \le 0 \\ 1-x, & \text{for } \ 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Note that the function given by (28) is not a 'wavelet' in the strict sense. This function, as well as other cardinal B-splines, however, as an elementary component with several key wavelet properties, can be used to form a class of biorthogonal wavelet systems, which can be applied for signal representation including multiresolution analysis (MRA) (Unser et al. 1993, Chui 1992, Unser 1999).

The function (28) was used as the elementary building blocks for constructing the generalised additive multiscale wavelet models. All the experiment conditions involved in the modelling procedure are shown in Table 2. The error-to-signal ratio (ESR), produced by the OPP+PSO algorithm, for the two modelling cases, that is, for r=1, n_{lag} =4 and r=2, n_{lag} =4, is shown in Fig. 2. A total of 200 construction functions of the form (10) were involved, for both of the two cases, after having performed the optimisation procedure on the associated data sets.

Significant individual wavelets were then selected, using the FOR-MI algorithm, from the dictionary of the form (23), which contains 200 individual candidate wavelets of the

form $\psi_{j,k} = \psi(x_j; a_{j,k}, b_{j,k})$, with j=1,2, …, d (d=36 or 100) and k=1,2, …, m_j. Note that both the dilation and translation parameters of the wavelets have already been optimised. The adjustable prediction error sum of squares (APRESS), defined by (27), suggests that a total of 26 and 22 wavelets should be included in the final models, respectively, for the two cases. The structure of the final models is of the form (12), and the parameters of the associated wavelets are shown in Tables 3 and 4, respectively, where the basis functions were ranked according to the order they have entered into the model. Notice that the symbols $s(i, j; o)$, in the first column of Tables 3 and 4, have the same meanings as those of $s_{i,j}(o)$, representing the stave variables associated with the cells at the locations (i, j).

Table 2. Some conditions involved in the modelling for the identification of the BZ reaction.

Size of the arrays of cells	360×500
Number of model variables	For the first case: 36. For the second case: 100.
mOPP in the OPP algorithm	200
η in the OPP algorithm	10^{-4}
Swarm's size in the PSO algorithm	20
c_1, c_2 in the PSO algorithm	$c_1 = c_2 = 2.05$
χ in the PSO algorithm	0.7298
mPSO in the PSO algorithm	500
δ in the PSO algorithm	10^{-5}
ε in the FOR algorithm	10^{-10}
λ in the FOR algorithm	10

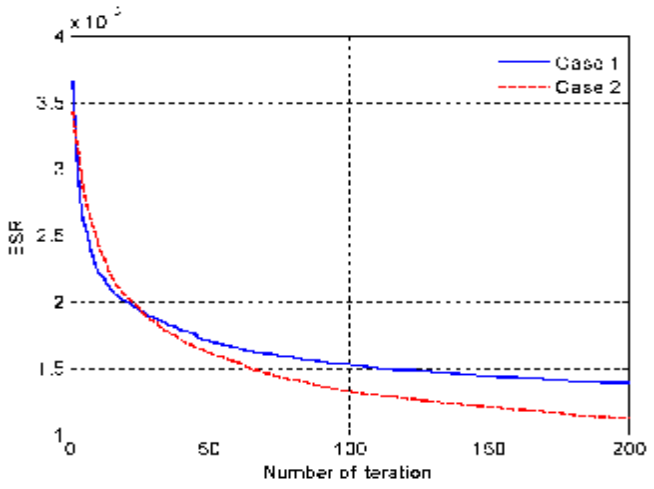

Fig. 2 The ESR index produced by the PSO based OPP algorithm. The solid line is for the first case r=1, n_{lag} =4, and the dashed line is for the second case r=2, n_{lag} =4.

Table 3. Model parameters and the associated mutual information and ERR values for the first case modeling, where r=1and n_{lag} =4.

Cell	Parameter			Mutual Information	ERR (%)
	c	a	b		
$s(i, j; t\text{-}1)$	-0.1113613234	3.8408459941	2.8450837979	0.4225	11.4484
$s(i, j; t\text{-}1)$	0.0557852403	4.2996948717	3.2497767282	0.6439	1.6516
$s(i, j\text{-}1; t\text{-}1)$	1.7919585090	0.5078399143	0.9507093126	0.6496	86.5475
$s(i, j\text{-}1; t\text{-}1)$	0.1394020814	4.0337783748	3.1490571756	0.2555	0.0330
$s(i, j\text{-}1; t\text{-}1)$	0.0534835351	2.4148153204	1.4352719087	0.2852	0.0320
$s(i, j\text{-}1; t\text{-}1)$	0.2087648842	5.0744523011	2.7837964511	0.3465	0.0003
$s(i, j; t\text{-}4)$	-0.0250904775	0.9499966382	-0.3344248874	0.3010	0.0101
$s(i, j; t\text{-}3)$	0.1290364751	3.3279581627	2.3296782012	0.2933	0.0088
$s(i\text{+}1, j; t\text{-}1)$	-0.1724189094	1.1313358888	0.1327239467	0.2849	0.0227
$s(i\text{-}1, j; t\text{-}1)$	-0.1181300168	1.1650008763	0.1647982765	0.2492	0.0473
$s(i, j; t\text{-}3)$	-0.1059315572	3.1479703660	2.3208915380	0.3304	0.0034
$s(i\text{-}1, j\text{+}1; t\text{-}1)$	-0.9535664802	1.1016170640	-0.1513421027	0.3027	0.0131
$s(i, j; t\text{-}2)$	0.2282715580	1.0233317306	0.0325486342	0.2610	0.0017
$s(i\text{-}1, j\text{+}1; t\text{-}1)$	-0.1118315067	1.2314969328	0.2314266980	0.2680	0.0033
$s(i\text{+}1, j\text{-}1; t\text{-}2)$	-0.0978967141	3.7570593203	2.571605679	0.2415	0.0004
$s(i\text{+}1, j; t\text{-}1)$	-0.0180375003	13.1181416431	10.1035711418	0.2390	0.0001
$s(i, j; t\text{-}4)$	0.1478468955	1.1819682761	0.1847641348	0.2313	0.0026
$s(i\text{+}1, j\text{-}1; t\text{-}2)$	0.0585486243	4.5782269976	3.1362071320	0.2557	0.0003
$s(i\text{-}1, j\text{+}1; t\text{-}1)$	1.0128540478	0.8015637372	-0.3732090912	0.2574	0.0001
$s(i\text{+}1, j\text{+}1; t\text{-}2)$	-0.0489460422	2.8986512223	1.9034078809	0.2378	0.0025
$s(i, j\text{+}1; t\text{-}3)$	0.0069154423	4.2272195966	2.8592631148	0.2060	0.0007
$s(i, j\text{-}1; t\text{-}3)$	0.0203976108	9.9551936231	6.8318100170	0.1986	0.0004
$s(i\text{-}1, j\text{-}1; t\text{-}2)$	-0.0534139564	2.7418648560	1.6758961781	0.2039	0.0036
$s(i, j\text{+}1; t\text{-}3)$	0.0498998614	4.6110908548	2.8916749940	0.2002	0.0002
$s(i\text{-}1, j\text{+}1; t\text{-}2)$	-0.0746461110	2.2972781915	1.3314358378	0.2075	0.0035
$s(i\text{+}1, j\text{+}1; t\text{-}1)$	-0.2650697089	1.1203267338	-0.1289945838	0.2046	0.0026
Sum of ERR					99.8403%

5.3 Model performance evaluation

To evaluate the performance of the identified additive wavelet models, the short-term predictive capability of the models was inspected. Denote the observation of the image (pattern) measured at the time instant t by $X(t)$. The k-step-ahead prediction, denoted by $\hat{X}(t+k \mid X(t), X(t-1), X(t-2), X(t-3); f)$, where f represents the identified nonlinear function, is the iteratively produced result by the identified model, on the basis of $X(t)$, $X(t\text{-}1)$, $X(t\text{-}2)$ and $X(t\text{-}3)$, but without using information on observations for patterns at any other time instants. As an example, the measurements at the time instants t=81, 82, 83, 84, were considered and used to calculate the 1-, 4-, and 8-step-ahead predictions, and these are shown in Fig. 3.

To quantitatively measure the performance of the identified models, the 2-D normalised mean-square-error (NMSE), defined as below, was considered

Table 4. Model parameters and the associated mutual information and ERR values for the second case modeling, where r=2and n_{lag} =4.

Cell	Parameter			Mutual Information	ERR (%)
	c	a	b		
$s(i, j; t\text{-}1)$	1.0000920021	1.2292201279	1.1692873968	0.4362	94.6567
$s(i\text{+}1, j\text{-}1; t\text{-}1)$	1.0961117713	0.2252175522	1.1795704565	0.3454	1.4067
$s(i\text{-}1, j\text{+}2; t\text{-}2)$	-0.1443107728	2.3059178081	1.3076835666	0.2933	2.0007
$s(i, j; t\text{-}1)$	-0.1171817131	3.0572377230	2.3400923504	0.3349	1.6040
$s(i, j; t\text{-}2)$	0.0110158450	4.4819903536	3.48612375230	0.3345	0.0047
$s(i\text{+}1, j; t\text{-}3)$	0.2349323434	1.7830321993	0.9173262355	0.2930	0.0515
$s(i\text{+}1, j\text{-}1; t\text{-}2)$	0.0047590764	7.6437365155	5.8012090756	0.3106	0.0001
$s(i\text{+}1, j\text{-}1; t\text{-}2)$	-0.0077441796	8.1703845040	5.5222689564	0.3103	0.0028
$s(i\text{-}2, j\text{-}1; t\text{-}1)$	-0.0374089268	3.8082029627	2.5280894175	0.2739	0.0022
$s(i\text{+}2, j\text{+}1; t\text{-}1)$	-0.0886680186	2.8836245906	1.8770173153	0.2669	0.0167
$s(i, j\text{+}2; t\text{-}2)$	-0.0280476559	5.8975582677	4.8570011044	0.2667	0.0016
$s(i\text{+}1, j; t\text{-}3)$	-0.0955759272	2.5469787758	1.2902705475	0.2501	0.0040
$s(i\text{+}2, j\text{-}1; t\text{-}2)$	-0.0654987697	2.3447767771	1.3473739411	0.2437	0.0034
$s(i\text{+}2, j\text{+}1; t\text{-}3)$	0.0123840401	2.4881153550	1.5489510183	0.2442	0.0029
$s(i, j; t\text{-}1)$	0.2500311636	2.5176213549	1.5514863164	0.2434	0.0227
$s(i, j\text{+}2; t\text{-}4)$	0.2463242832	0.4876748764	-0.5138555005	0.2276	0.0128
$s(i, j\text{+}2; t\text{-}2)$	0.0048319233	2.9458302335	1.9318118117	0.2323	0.0002
$s(i\text{-}1, j\text{+}2; t\text{-}2)$	0.0123733687	5.7636275247	4.7199405993	0.2391	0.0017
$s(i\text{-}1, j; t\text{-}4)$	0.0241436192	8.4533718197	5.5030465346	0.2319	0.0080
$s(i\text{-}1, j\text{-}2; t\text{-}2)$	-0.1667885198	1.3174794458	0.3064101933	0.2129	0.0116
$s(i\text{-}1, j; t\text{-}4)$	0.0787648605	3.3853091789	2.2114094953	0.2294	0.0067
$s(i\text{-}2, j\text{-}2; t\text{-}3)$	-0.0079349342	4.8920332394	3.8801885685	0.2204	0.0003
Sum of ERR					99.8224%

$$\mathrm{NMSE}(t) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} | s_{i,j}(t) - \hat{s}_{i,j}^{(k)}(t) |^2}{\sum_{i=1}^{I}\sum_{j=1}^{J} | s_{i,j}(t) |^2} \tag{29}$$

where $s_{i,j}(t)$ represent the observations at the time instant t, $\hat{s}_{i,j}^{(k)}(t)$ represent the corresponding k-step-ahead predicted values from a given model, and I and J define the size of the associated patterns. Starting with the observations at the time instants t=81, 82, 83, 84, the two identified models given in Tables 3 and 4 were used to calculate k-step-ahead predictions, for k=1,2, …, 10. The predicted values were then compared with the associated observations, corresponding to the time instants t=85, 86, …, 94, respectively. The associated normalised mean-square-errors, for the two models, are shown in Fig. 4.

From Fig. 3 and Fig. 4, it can be concluded that: i) the identified additive wavelet models can capture the main spatio-temporal evolution dynamics of the BZ reaction; ii) both models can provide very good short term, say 1-step-ahead, predictions; iii) the model given by Table 3 is superior to the model given by Table 4, this implies that the radius of the neighbourhood, r, can be chosen as 1 when a generalised additive multiscale model is applied to represent the BZ reaction.

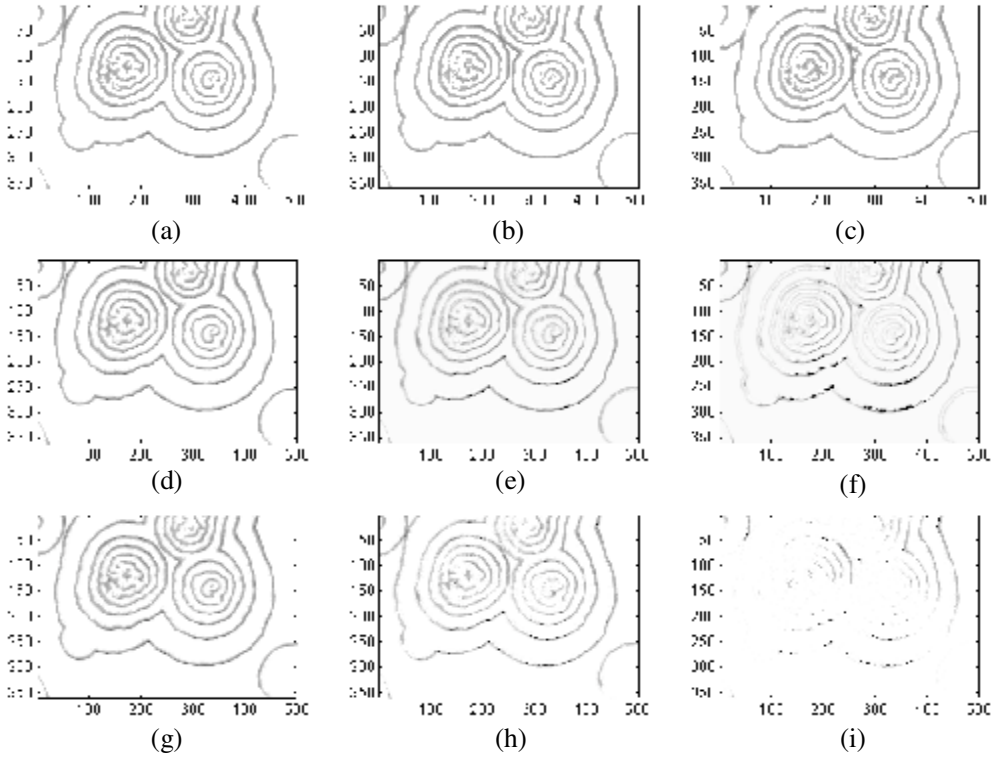

Fig. 3 The 1-, 4-, and 8-step-ahead predictions, based on the observations at the time instants t=81,82,83,84, for the BZ reaction. (a)-(c) True measurements at the time instants t=85, 89, and 94; (d)-(f) The 1-, 4-, and 8-step-ahead predicted results for (a), (b) and (c), respectively, using the model given in Table 3; (g)-(i) The 1-, 4-, and 8-step-ahead predicted results for (a), (b) and (c), respectively, using the model given in Table 4.

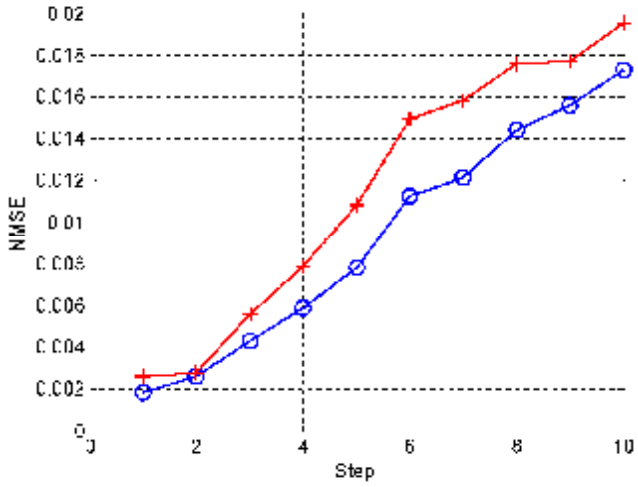

Fig. 4 The normalised mean-square-errors for k-step-ahead predictions, with k=1,2, …, 10. The circled-line is for the model given by Table 3, and the crossed-line is for the model given by Table 4.

6. Conclusions

Identification of spatio-temporal evolutionary (SEE) systems often involves a great number of 'input' variables, meaning that a very high dimensional modelling problem may have to be solved. To effectively solve such types of very high dimensional identification problems, a new family of wavelet based generalized additive multiscale models has been introduced. A novel two-stage hybrid learning method has been proposed for constructing such an additive multiscale model. The new learning algorithm produces a transparent model, where individual basis functions (model terms) are explicitly available.

By introducing the PSO algorithm, which is easy to implement, the calculation of gradients required by classical nonlinear optimisation algorithms can be avoided. This makes the new multiscale modelling framework very suitable for STE system identification, where relevant object functions may not be differentiable or relevant gradients are very difficult to obtain. By applying the mutual information based forward orthogonal regression (FOR) algorithm, the initially produced model by the PSO based orthogonal projection pursuit (OPP) learning algorithm, can significantly be refined and improved, and a parsimonious model containing only a small number of basis functions can then be obtained.

The proposed learning scheme and algorithms can also be applied in high dimensional pure temporal dynamical system identification including very high dimensional time series modelling and prediction. In addition, the proposed method may also be adapted to an alternative approach for neighbourhood detection of STE systems, as discussed in the example.

Acknowledgements

The authors gratefully acknowledge that this work was supported by Engineering and Physical Sciences Research Council (EPSRC), U.K. They gratefully acknowledge the help from Dr A. F. Routh who supervised and Dr Y. Zhao who conducted the B-Z experiments.

References

A. Adamatzky and V. Bronnikov, "Identification of additive cellular automata," *J. Comput. Syst. Sci.*, vol. 28, pp. 47–51, 1990.

A. Adamatzky, *Identification of Cellular Automata*. London: Taylor & Francis, 1994.

A. Adamatzky, "Automatic programming of cellular automata: Identification approach," *Kybernetes*, vol.26, pp. 126–133, 1997.

M. Aerts, G. Claeskens, and M. P. Wand MP, "Some theory for penalized additive models," *J. Statist. Plann. Infer.*, vol. 103, pp. 455-470, Apr. 2002

L. Astic, V. Pellier-Monnin, and F. Godinot, "Spatio-temporal patterns of ensheathing cell differentiation in the rat olfactory system during development," *Neuroscience*, vol. 84, no.1, pp. 295-307, May 1998.

J. Bascompte and R. V. Sole (ed.), 1998, *Modelling Spatiotemporal Dynamics in Ecology*. Berlin: Springer, 1998.

B.P. Belousov, "A periodic reaction and its mechanism," in Collection of Short Papers on Radiation Medicine (in Russian), Medgiz, Moscow, pp.145-152, 1959.

R. Berezney, K. S. Malyavantham, A. Pliss, S. Bhattacharya, and R. Acharya, " Spatio-temporal dynamics of genomic organization and function in the mammalian cell nucleus," *Advances In Enzyme Regulation*, vol. 45, pp.17-26, 2005.

S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO nonlinear systems suing a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, no.6, pp.2157-2189, Jun.1989.

S.A. Billings and D. Coca, ''Identification of coupled map lattice models of deterministic distributed parameter systems,'' *Int. J. Sys. Sci*., 33, pp. 623–634, 2002.

S. A. Billings, L. Z. Guo, and H. L. Wei, "Identification of coupled map lattice models for spatio-temporal patterns using wavelets," *Int. J. Control*, vol.14, No. 14, pp. 1021-1038, Nov 2005.

S. A. Billings and H. L. Wei, "A new class of wavelet networks for nonlinear system identification," *IEEE Trans. Neural Networks*, 16, no.4, pp. 862-874, July 2005.

S. A. Billings and H. L. Wei, "Sparse model identification using a forward orthogonal regression algorithm aided by mutual information," *IEEE Trans. Neural Networks*, vol.18, no.1, pp. 306-310, Jan. 2007a.

S. A. Billings and H. L. Wei, "An adaptive orthogonal search algorithm for model subset selection and nonlinear system identification," *Int. J. Control*, 2007b (in press).

S.A. Billings and Y. Y. Yang, ''Identification of the neighbourhood and CA Rules from Spatio-temporal CA patterns,'' *IEEE Trans. Syst. Man Cybern. B*, 33, pp. 332–339, 2003.

A. Brezger and S. Lang, "Generalized structured additive regression based on Bayesian P-splines," *J. Comput. Statist. Data Anal*., vol.50, no.4, pp. 967–991, Feb. 2006.

A. Buja, T. Hastle and R. Tibshirani, " Linear smoothers additive-models," *Ann. Stat.*, vol. 17, no.2, pp.453–510, June 1989.

Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 11, pp. 1409–1424, Nov. 2002.

S. Chen, S. A. Billings and W. Luo, "Orthogonal least squares methods and their application to nonlinear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.

L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 35, no. 12, pp. 1257–1272, Dec. 1988a.

L. O. Chua and L. Yang, "Cellular neural networks: Applications," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 35, no. 12, pp. 1273–1290, Dec. 1988b.

L. O. Chua and T. Roska, *Cellular Neural Networks and Visual Computing*. Cambridge: Cambridge University Press, 2002.

C. K. Chui, *An Introduction to Wavelets*. New York: Academic, 1992.

M. Clerc and J. Kennedy, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space," *IEEE Trans. Evol. Comput.*, vol. 6, no.1, pp. 58–73, Feb. 2002.

D. Coca and S. A. Billings, "Identification of coupled map lattice models of complex spatio-temporal patterns," *Phys. Lett.*, vol. A287, pp. 65–73, 2001.

T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

T. Czaran, *Spatiotemporal Models of Population and Community Dynamics*. London: Chapman & Hall, 1998.

G.A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol.45, no.4, pp.1315-1321, 1999.

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

D. S. Dimitrova and R. Berezney, "The spatio-temporal organization of DNA replication sites is identical in primary, immortalized and transformed mammalian cells," *J. Cell Science*, vol. 115, no. 21, pp.4037-4051, Nov. 2002.

Y. Dolak and C. Schmeiser, "Kinetic models for chemotaxis: Hydrodynamic limits and spatio-temporal mechanisms," *J. Math. Bio.*, vol.51, no.6, pp.595-615. Dec. 2005.

R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6 th Symp. Micro Mach. Human Sci.*, pp. 39-43, Nagoya, Japan, Oct. 4-6, 1995.

S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, pp. 524–532, 1990.

J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Statist. Assoc.*, vol. 76, no. 376, pp.817–823, Dec.1981.

T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. London: Chapman & Hall, 1990.

G. B. Huang, L. Chen, and C. K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

J. N. Hwang, S. R. Lay, M.Maechler, R. D. Martin, and J. Schimert, "Regression modeling in back-propagation and projection pursuit learning," *IEEE Trans. Neural Networks*, vol. 5, no. 3, pp. 342–353, May 1994.

A. Ilachinski, *Cellular Automata: A Discrete Universe*, New Jersey : World Scientific, 2001.

B. Jahne, *Spatio-Temporal Image Processing: Theory and Scientific Applications*. Berlin: Springer-Verlag, 1993.

L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992.

K. Kaneko, *Theory and Application of Coupled Map Lattices*. New York: Wiley, 1993.

J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proc.IEEE Int. Conf. Neural Networks*, vol. IV, pp.1942–1948, Perth, Australia, 1995.

J. Kennedy, R. C. Eberhart and Y. Shi, *Swarm Intelligence*, San Francisco: Morgan Kaufmann Publishers, 2001.

J. Kim and J. Woods, "Spatio-temporal adapive 3-D Kalman filter for video," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 414–424, Mar. 1997.

J. Kim and J. W. Woods, "3-D Kalman filter for image motion estimation," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 42–52, Jan. 1998.

Y. Kuramoto, *Chemical Oscillations*, *Waves*, *and Turbulence*. Berlin: Springer, 1984.

T. Y. Kwok and D. Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 630–645, May 1997a.

T. Y. Kwok and D. Y. Yeung, "Objective functions for training new hidden units in constructive neural networks," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 1131–1148, Sep. 1997b.

M. J. Lado, C. Cadarso-Suarez, J. Roca-Pardinas, and P. G. Tahoces, "Using generalized additive models for construction of nonlinear classifiers in computer-aided diagnosis systems," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 246–253, Apr. 2006.

S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

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

S. Mandelj, I. Grabec and E. Govekar, ''Statistical approach to modeling of spatiotemporal dynamics," *Int. J. Bifurcation and Chaos*, vol. 11, pp. 2731–2738, 2001.

R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no.3, pp. 233-246, 1989.

R. Moddemeijer, "A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations," *Signal Processing*, vol. 75, no. 1, pp 51-63, 1999.

C. W. Ngo, T. C. Pong, and H. J. Zhang, "Motion analysis and segmentation through spatio-temporal slices processing," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 341–355, Mar. 2003.

H. T. Nguyen, Q. Ji, and A. W. M. Smeulders, "Spatio-temporal context for robust multitarget tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 1, pp. 52–64, Jan. 2007.

L. Paninski, **"**Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp.1191-1253, 2003.

U. Parlitz and C. Merkwirth, ''Prediction of spatiotemporal time series based on reconstructed local states'', *Phys. Rev. Lett.*, vol.84, pp. 2820–2823, 2000.

K. E. Parsopoulos and M. N. Vrahatis, "On the computation of all global minimizes through particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no.3, pp. 211–224, June 2004.

R. D. Reed and R. J. Marks II, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Cambridge, MA: The MIT Press, 1999.

Y. Ricquebourg and P. Bouthemy, "Real-time tracking of moving persons by exploiting spatio-temporal image slices," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 797–808, 2000.

D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. Cambridge: Cambridge University Press, 2003.

F. J. Sanchez-Marin, Y. Srinivas, K. N. Jabri, and D. L. Wilson, "Quantitative image quality analysis of a nonlinear spatio-temporal filter," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 288-295, Feb. 2001.

Y. Shi and R. C. Eberhart, "Parameter selection in particle swarm optimization," in *Lecture Notes In Computer Science*, V. W. Porto, N. Saravanan, D. Waagen, and A. E. Eiben (Eds), vol. 1447, pp. 591-600, 1998a.

Y. Shi and R. C. Eberhart, "A modified particle swarm optimizer," in *Proc. IEEE Conf. Evolutionary Computation*, pp. 69-73, Anchorage, AK, USA, 4^{th} -9^{th} May, 1998b.

F. L. Silva, J. C. Principe, and L. B. Almeida (ed.), *Spatiotemporal Models in Biological and Artificial Systems*. Washington: IOS Press, 1997.

A. Sitz, J. Kurths, and H. U. Voss, "Identification of nonlinear spatio-temporal systems via partitioned filtering," *Physical Review E*, vol. 68, 016202, 2003.

H. Spors and A. Grinvald, "Spatio-temporal dynamics of odor representations in the mammalian olfactory bulb," *Neuron*, vol. 34, no.2, pp.301-315, Apr. 2002.

C. J. Stone, "Additive regression and other nonparametric models," *Ann. Stat.*, vol. 13, no.2, pp.689–705, 1985.

C. J. Stone, "The dimensionality reduction principle for generalized additive models," *Ann. Stat.*, vol. 14, no.2, pp.590–606, June 1986.

M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 1549-1560, Nov. 1995.

M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Process. Mag.*, vol. 16, no. 6, pp. 22–38, Nov. 1999.

M. Unser and A. Aldroubi, and M. Eden, "A family of polynomial spline wavelet transforms," Signal Processing, vol.30, no.2, pp.141-162, Jan. 1993.

D. Van De Ville, T. Blu, and M. Unser, "Integrated wavelet processing and spatial statistical testing of fMRI data," NeuroImage, vol. 23, no. 4, pp.1472-1485, Dec. 2004.

F. van den Bergh, "An analysis of particle swarm optimizers," Ph.D. Dissertation, Dept. Comput. Sci., Univ. Pretoria, Pretoria, South Africa, 2002.

F. van den Bergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE Trans. Evolutionary Computation*, vol. 8, no.3, pp.225-239, June 2004.

C.J. Veenman, M.J.T. Reinders, and E. Backer, "A Cellular Coevolutionary Algorithm for Image Segmentation," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 304-316, Mar. 2003.

H. L. Wei and S. A. Billings,"An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon," *Nonlin. Processes Geophys.*, vol.13, pp.577–584, 2006a.

H. L. Wei and S. A. Billings, "Long term prediction of nonlinear time series using multiresolution wavelet models," *Int. J. Control*, vol. 79, no.6, pp. 569-580, June 2006b.

H. L. Wei and S. A. Billings, "Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information," *Int.J. Modelling, Identification and Control*, 2007 (in press).

A. T. Winfree, "Spiral waves of chemical activity," Science, vol. 175, no.4022, pp. 634-636, 1972.

S. Wolfram, *Cellular Automata and Complexity*. New York: Addison-Wesley, 1994.

S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *J. Aerica. Statist.Assoc.*, vol. 99, no. 467, pp.673-686, Sep. 2004.

Y. S. Xia and H. Leung, "Nonlinear spatial-temporal prediction based on optimal fusion," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 975–988, July 2006.

Q. Yang and B. Parvin, "High-resolution reconstruction of sparse data from dense low-resolution spatio-temporal data," *IEEE Trans. Image Process.*, vol.12, no.6, pp.671-677, June 2003.

A.M. Zhabotinsky, Periodic liquid phase reactions. Proc. Acad. Sci. USSR, 157, pp.392-395, 1964.