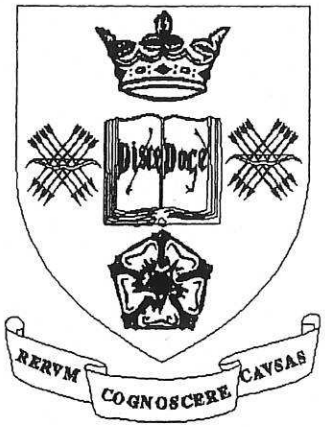
Multiscale Radial Basis Function Networks

S. A. Billings, H. L. Wei and M. A. Balikhin



Research Report No. 894

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

May 2005

Multiscale Radial Basis Function Networks

Stephen A. Billings, Hua-Liang Wei and M. A. Balikhin
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK
S.Billings@Sheffield.ac.uk, W.Hualiang@Sheffield.ac.uk

May 10, 2005

Abstract: A novel modelling framework is proposed for constructing parsimonious and flexible radial basis function network (RBF) models. Unlike a conventional standard Gaussian kernel based RBF network, where all the basis functions have the same scale (kernel width), or each basis function has a single individual scale, the new network construction approach adopts multiscale kernels (with multiple kernel widths for each selected centre) as the basis functions to provide more flexible representations with better generalization properties for general nonlinear dynamical systems. A standard orthogonal least squares (OLS) algorithm is then applied to select significant model terms (basis functions) to obtain parsimonious models.

Keywords: dynamical modelling, model term selection, neural network, nonlinear system identification, orthogonal least squares, radial basis function, regression

1. Introduction

Radial basis function (RBF) networks, as a special class of single hidden-layer feedforward neural networks, have been proved to be universal approximators [1-3]. One advantage of RBF networks compared with multi-layer perceptrons (MLP) is that the linearly weighted structure of RBF networks, where parameters in the units of the hidden layer can often be pre-fixed, can be easily trained with a fast speed without involving nonlinear optimization. Another advantage of RBF networks, compared with other basis function networks, is that each basis function in the hidden units is a nonlinear mapping which maps a multivariable input to a scalar value, and thus the total number of candidate basis functions involved in a RBF network model is not very large and does not increase when the number of input variables increases. With these attractive properties, RBF networks are an important and popular network model for function approximation [4, 5], classification and pattern recognition [6-10], dynamical modelling and control [11-15].

The training of RBF networks involves the optimization of three parameters: kernel centres, kernel widths and the connecting weights between these kernels (neurons). These parameters can be determined by performing either separate or combined procedures [10, 16]. While several efficient algorithms have been introduced to determine kernel centres [10, 16-20], few algorithms are available to determine kernel widths [21]. The kernel widths are thus often determined on the basis of some heuristics [17, 22]. Recently, Benoudjit and Verleysen [21] investigated the kernel width selection problem and proposed a one-dimensional search algorithm as a compromise between an exhaustive search on all basis function widths and a non-optimal a priori choice. Wang et al. [23] studied the problem of constructing generalized Gaussian kernel models for nonlinear dynamical modelling and proposed a repeatedly guided random search algorithm for calculating the individual diagonal covariance matrices based on boosting optimization.

The present study mainly concerns the construction and training of a novel class of RBF networks for nonlinear dynamical system modelling and identification. Unlike a conventional RBF, where all the basis functions have a common single scale (kernel width), or each basis function has a single individual scale, the new RBF network uses a set of multiscale basis functions, where each basis function has multiple scale parameters(kernel widths). The new network will be referred to as the *multiscale RBF network*. The positions (centres) of the basis functions can often be pre-clustered or pre-selected using some unsupervised clustering algorithms [10, 16-20], and for each selected centre, the associated multiple widths can often be restricted to a fixed grid, thus multiscale RBF networks can be easily converted into a linear-in-the-parameters model form. The standard orthogonal least squares algorithms [12, 24-26] can then be used to train a multiscale RBF network.

2. The Nonlinear Dynamical Modelling Problem

Consider the identification problem for nonlinear systems given N pairs of input-output observations, $\{u(t), y(t)\}_{t=1}^{N}$. Under some mild conditions a discrete-time nonlinear system can be described by the following NARMAX model [27]

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

where $u(t)$, $y(t)$ and $e(t)$ are the system input, output and noise variables, n_u, n_y and n_e are the maximum lags in input, output and noise, respectively, and f is some vector-valued and in general unknown nonlinear mapping. In practice it is usually assumed that $e(t)$ is an independent noise sequence. Model (1) relates the inputs and outputs and takes into account the combined effects of measurement noise, modelling errors and unmeasured disturbances represented by the noise variable $e(t)$ .One of the reasons that the moving average terms are included in the NARMAX model (1) is to ensure unbiased estimation.

In practice, the unknown nonlinear function f in model (1) often consists of two parts: the deterministic (noise independent) and the stochastic (noise correlated) submodels shown as below

$$y(t) = f_{yu}(y^{[t-1,n_y]},u^{[t-1,n_u]}) + f_{yue}(y^{[t-1,n_y]},u^{[t-1,n_u]},e^{[t-1,n_e]}) + e(t) \tag{2}$$

where the vector $z^{[t-1,n]}$ is defined as $z^{[t-1,n]} = [z(t-1),\cdots,z(t-n)]^T$. Note that each term of the submodel f_{yue} is dependent on at least one of the noise signals $e(t-1), e(t-2),\cdots,e(t-n_e)$. For a linear-in-the-parameters basis function network, model (2) can be expressed as

$$y(t) = \Phi_{yu}(t)\Theta_{yu} + \Phi_{yue}(t)\Theta_{yue} + e(t) \tag{3}$$

where $\Phi_{yu}(t)$ and $\Phi_{yue}(t)$ are regression matrices, and Θ_{yu} and Θ_{yue} are unknown parameter vectors.

The central objective for any identification problem is to find approximators $\hat{f}_{yu}$ and $\hat{f}_{yue}$ for the unknown functions f_{yu} and f_{yue} in (2) so that $\hat{f}_{yu}$ and $\hat{f}_{yue}$ can be used for system analysis, prediction and control. In practice, several types of basis functions including radial basis functions have been employed to construct the approximator $\hat{f}_{yu}$. In the present study, however, a new multiscale RBF network model with Gaussian kernels

will be used to construct the approximator $\hat{f}_{yu}$, and power-form polynomial models will be used to construct the noise related approximator $\hat{f}_{yue}$.

3. Multiscale RBF Networks

Prior to presenting the new multiscale RBF networks, conventional RBF networks are briefly reviewed and the drawbacks associated with conventional RBF networks for dynamical modelling are discussed.

3.1 Conventional RBF networks

Let $d = n_y + n_u$ and $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]^T$ with

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y + 1 \le k \le n_y + n_u \end{cases} \tag{4}$$

A traditionally adopted RBF network with Gaussian kernels for approximating the nonlinear function f_{yu} in model (2) is given below

$$\hat{f}_{yu}(\mathbf{x}(t)) = \sum_{i=1}^{M} \theta_i \varphi_i(\mathbf{x}(t); \boldsymbol{\sigma}_i, \mathbf{c}_i) \tag{5}$$

where $\varphi_i : \mathbf{R}^d \mapsto \mathbf{R}$ is the standard Gaussian kernel

$$\varphi_i(\mathbf{x}(t); \boldsymbol{\sigma}_i, \mathbf{c}_i) = \exp\left[-\frac{1}{2} (\mathbf{x}(t) - \mathbf{c}_i)^T \Lambda_i^{-1} (\mathbf{x}(t) - \mathbf{c}_i) \right] \tag{6}$$

where $\boldsymbol{\sigma}_i = [\sigma_{i,1}, \cdots, \sigma_{i,d}]^T$ are the scale (or dilation) parameters to determine the kernel widths, $\mathbf{c}_i = [c_{i,1}, \cdots, c_{i,d}]^T$ are the location (or translation) parameters to determine the kernel positions (centres), and $\Lambda_i = diag[\sigma_{i,1}^2, \cdots, \sigma_{i,d}^2]$ are referred to as the covariance matrices. The network model (5) can then be written as

$$\hat{f}_{yu}(\mathbf{x}(t)) = \sum_{i=1}^{M} \theta_i e^{-\frac{1}{2}(\mathbf{x}(t)-\mathbf{c}_i)^T \Lambda_i^{-1} (\mathbf{x}(t)-\mathbf{c}_i)} \tag{7}$$

In practice, the location (or translation) parameters $\mathbf{c}_i$ can often be pre-selected, for example all given observational data points can be considered as candidate kernel centres providing that the observational data set is not very long, or alternatively a set of clustered points can be chosen as the candidate kernel centres if the observational data set is long. As for the determination of the scale (or dilation) parameters $\boldsymbol{\sigma}_i$, two cases are often considered [21]:

i) All the diagonal covariance matrices are set to the identical diagonal form, $\Lambda_i = diag[\sigma^2, \cdots, \sigma^2]$.

ii) A general case where $\Lambda_i = diag[\sigma_{i,1}^2, \cdots, \sigma_{i,d}^2]$.

Case i) is in the class of the most commonly used Gaussian kernel based RBF networks. Although it has been proved that the simple cases i) can often provide a universal approximation [3], these may not be a good choice for nonlinear dynamical modelling. This can be explained by observing the operating range features of the elements in the input vector $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]^T$ with $x_k(t)$ defined by (4). An implicit assumption on the input vector is that all the elements should play an equal role in constructing the network. This assumption is reasonable for black-box modelling, where no prior knowledge is known about which elements in the input vector $\mathbf{x}(t)$ are more significant and which elements are insignificant compared with others. Assume that the system input $u(t)$ and the output $y(t)$ are bounded in $[\underline{u}, \overline{u}]$ and $[\underline{y}, \overline{y}]$, respectively, and let $r_u = \overline{u} - \underline{u}$ and $r_y = \overline{y} - \underline{y}$. Consider two commonly encountered cases: (a) $r_y >> r_u$ and (b) $r_y << r_u$. Expanding the function φ_i in (6) as

$$\varphi_i(\mathbf{x}(t); \boldsymbol{\sigma}_i, \mathbf{c}_i) = \exp\left\{-\frac{1}{2}\left[\left(\frac{y(t-1)-c_{i,1}}{\sigma_i}\right)^2 + \left(\frac{y(t-n_y)-c_{i,n_y}}{\sigma_i}\right)^2 + \cdots \right.\right.$$
$$\left.\left. + \left(\frac{u(t-1)-c_{i,n_y+1}}{\sigma_i}\right)^2 + \cdots + \left(\frac{u(t-n_u)-c_{i,d}}{\sigma_i}\right)^2\right]\right\} \tag{8}$$

For case (a), the roles of the lagged output variables $y(t-p)$ may be exaggerated or the roles of the lagged input variables $u(t-q)$ may be downplayed (especially for large σ_i). Similarly, for case (b), the roles of the lagged input variables $u(t-q)$ may be exaggerated or the roles of the lagged output variables $y(t-p)$ may be downplayed. To overcome this dilemma, a reasonable solution is to set different scale parameters for the input and output related elements in the basis function (8).

A good choice for the scale parameters $\boldsymbol{\sigma}_i$ to balance the roles of the lagged output and input variables is to use the generalized RBF network, where the scale parameters $\boldsymbol{\sigma}_i$ are determined adaptively. But the determination of the scale parameters $\boldsymbol{\sigma}_i$ in an adaptive way usually involves global optimization procedures, which are much complex when the number of candidate model terms is large in the network. Motivated by the observations above, the present study proposes a new class of Gaussian kernel based RNF networks, which provide a trade-off between cases i) and ii), and which can easily be trained using standard linear learning algorithms for instance the well known OLS algorithms [12, 24-26].

3.2 Multiscale RBF networks

Inspired by the successful applications of multiresoution wavelet decompositions [28, 29], which have excellent local properties both in the time and the frequency domain, and which provide a remarkable tool for data processing in a hierarchical multiscale way, a multiscale modelling framework will be introduced into RBF networks, where a set of scale parameters (kernel widths) will be assigned to each selected kernel centre.

A multiscale RBF network possesses the following structure

$$\hat{f}_{yu}(\mathbf{x}(t)) = \sum_{i=0}^{I}\sum_{j=0}^{J}\sum_{m=1}^{N_c}\theta_{i,j,m}\varphi_{i,j,m}(\mathbf{x}(t);\boldsymbol{\sigma}_m^{(i,j)},\mathbf{c}_m) \tag{9}$$

where the basis functions $\varphi_{j,m}(\mathbf{x}(t);\boldsymbol{\sigma}_{j,m},\mathbf{c}_m)$ are defined as

$$\varphi_{i,j,m}(\mathbf{x}(t);\boldsymbol{\sigma}_m^{(i,j)},\mathbf{c}_m) = \exp\left[-\sum_{k=1}^{d}\left(\frac{x_k(t)-c_{m,k}}{\sigma_{m,k}^{(i,j)}}\right)^2\right] \tag{10}$$

Note that the factor -1/2 which appears in (6)-(8) is accommodated into $\sigma_{m,k}^{(i,j)}$ in (10). In practice, all given observations can be considered as candidate kernel centres $\mathbf{c}_m$ providing that the observational data set is not very long. For a long data set, both self-organized and supervised learning approaches [16, 17, 20] including the well-known *k*-means clustering algorithm [30] can be used to locate the centres of the basis functions in only those regimens of the input space where significant data are presented. The determination of the kernel widths $\boldsymbol{\sigma}_m^{(i,j)}$ is discussed below.

For given N pairs of input-output observations, $\{u(t),y(t)\}_{t=1}^{N}$, let $\mathbf{x}(t)=[x_1(t),\cdots,x_d(t)]^T$ be defined as in (4). Let $\sigma_y = \beta[\max_t\{y(t)\}-\min_t\{y(t)\}]$ and $\sigma_u = \beta[\max_t\{u(t)\}-\min_t\{u(t)\}]$, where β is a positive constant which can typically be set to between 1 and 10. In the multiscale RBF networks, the diagonal covariance matrices are chosen as below:

$$\Lambda_m^{(i,j)} = diag[(\sigma_{y,m,1}^{(i)})^2,\cdots,(\sigma_{y,m,n_y}^{(i)})^2,(\sigma_{u,m,1}^{(j)})^2,\cdots,(\sigma_{u,m,n_u}^{(j)})^2] \tag{11}$$

where $\sigma_{y,m,p}^{(i)} = \alpha^{-i}\sigma_y$, $\sigma_{u,m,q}^{(j)} = \alpha^{-j}\sigma_u$, $\alpha>1$ is a constant, and i=0, …, I; j=0, 1, …, J; m=1,2, …, M; p=1,2, …, n_y and q=1,2, …, n_u. A typical choice for the constant α is set $\alpha=2$. The basis functions $\varphi_{i,j,m}$ defined by (10) thus reduces to

$$\varphi_{i,j,m}(\mathbf{x}(t);\boldsymbol{\sigma}_m^{(i,j)},\mathbf{c}_m) = \exp\left[-\sum_{k=1}^{n_y}\left(\frac{x_k(t)-c_{m,k}}{\sigma_y^{(i)}}\right)^2 - \sum_{k=n_y+1}^{n_y+n_u}\left(\frac{x_k(t)-c_{m,k}}{\sigma_u^{(j)}}\right)^2\right] \tag{12}$$

where $\sigma_y^{(i)} = \alpha^{-i}\sigma_y$ and $\sigma_u^{(j)} = \alpha^{-j}\sigma_u$. Let $D_3 = \{\varphi_{i,j,m}(\cdot;\boldsymbol{\sigma}_m^{(i,j)},\mathbf{c}_m): i=0,\cdots,I; j=0,\cdots,J; m=1,\cdots,N_c\}$. The triple indexed set D_3 is referred to as the dictionary associated with the new multiscale networks. For convenience of description, rearrange the elements of D_3 so that the double index (i,j,m) can be indicated by a single index m=1,2,…, M, where $M=(I+1)(J+1)N_c$, to form a single indexed dictionary $D_1 = \{\phi_m(\cdot): \phi_m \in D_2, m=1,\cdots,M\}$. This study will not distinguish the two type of dictionaries D_1 and D_3, instead, a uniform symbol D will be used to indicate both of them. The network (9) can then be expressed as

$$\hat{f}_{yu}(\mathbf{x}(t)) = \sum_{m=1}^{M}\theta_m\phi_m(\mathbf{x}(t)) \tag{13}$$

4. Model Structure Detection

Model structure detection, or model subset selection, is a key step in any identification procedure and consists of detecting and selecting significant model terms from a redundant candidate model term set to determine a parsimonious final model. As will be seen in the examples later, the multiscale RBF network (13) may involve a great number of candidate model terms (basis functions) when the parameters I, J, and N_c are large. Many of these candidate model terms, however, may be redundant and only a subset of these model terms is significant. Including redundant model terms might lead to a large number of free parameters in the model, and as a consequence the model may become oversensitive to the training data and is likely to exhibit poor generalisation properties. Therefore, it is important to determine which terms should be included in the model. In the present study, the OLS algorithm [12, 24-26] will be used to solve the model structure detection problem for the multiscale RBF network models.

Let $\mathbf{y} = [y(1), \cdots, y(N)]^T$ be a vector of measured outputs at N time instants, and $\alpha_m = [\phi_m(1), \cdots, \phi_m(N)]^T$ be a vector associated with the mth candidate model term, where $\phi_m \in D$ for m=1,2, ..., M, and D is a dictionary produced by lagged outputs, inputs and the noise terms involved. From the viewpoint of practical modelling and identification, the finite dimensional set $S = \{\alpha_1, \cdots, \alpha_M\}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $S_n = \{\beta_1, \cdots, \beta_n\} = \{\alpha_{i_1}, \cdots, \alpha_{i_n}\} \subseteq S$, where $\beta_k = \alpha_{i_k}$, $i_m \in \{1,2,\cdots,M\}$ and m=1,2, ..., n, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\beta_1, \cdots, \beta_n$ as below

$$\mathbf{y} = \theta_1\beta_1 + \cdots + \theta_n\beta_n + \mathbf{e} \tag{14}$$

or in a compact matrix form

$$\mathbf{y} = P\theta + \mathbf{e} \tag{15}$$

where the matrix $P = [\beta_1, \cdots, \beta_n]$ is of full column rank, $\theta = [\theta_1, \cdots, \theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is an approximation error. From matrix theory, the full rank matrix P can be orthogonally decomposed as

$$P = QR \tag{16}$$

where R is an $n \times n$ unit upper triangular matrix and Q is an $n \times n$ matrix with orthogonal columns $q_1, q_2, \cdots, q_n$. Substituting (16) into (15), yields

$$\mathbf{y} = (PR^{-1})(R\theta) + \mathbf{e} = Qg + \mathbf{e} \tag{17}$$

where $g = [g_1, \cdots, g_n]^T = R\theta$ is an auxiliary parameter vector. Using the orthogonal property of Q, g_i can be directly calculated from $\mathbf{y}$ and Q as $g_i = (\mathbf{y}^T q_i)/(q_i^T q_i)$ for i=1,2, ..., n. The unknown parameter vector θ can then be easily calculated from g and R by substitution using the special structure of R. Several orthogonal transforms including Gram-Schmidt, modified Gram-Schmidt and Householder transformations can be applied to implement the orthogonal decomposition [12, 24-26].

Assume that the error $\mathbf{e}$ in model (17) is uncorrelated with vectors β_j for j=1,2, …, n, the total sum of squares of the output from the origin can then be expressed as

$$\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{n} g_i^2 q_i^T q_i + \mathbf{e}^T\mathbf{e} \tag{18}$$

Note that the total sum of squares $\mathbf{y}^T\mathbf{y}$ consists of two parts, the desired output $\sum_{i=1}^{n} g_i^2 q_i^T q_i$, which can be explained by the selected regressors (model terms), and the part $\mathbf{e}^T\mathbf{e}$, which represents the residual sum of squares. Thus, $g_i^2 q_i^T q_i$ is the increment to the desired total sum of squares of the output brought by q_i. The ith error reduction ratio (ERR) introduced by q_i (or equally by including β_i), is defined as

$$\mathrm{ERR}[i] = \frac{g_i^2 (q_i^T q_i)}{\mathbf{y}^T\mathbf{y}} \times 100\% = \frac{(\mathbf{y}^T q_i)^2}{(\mathbf{y}^T\mathbf{y})(q_i^T q_i)} \times 100\%, \quad i=1,2, \ldots, n, \tag{19}$$

This ratio provides a simple but an effective index to indicate the significance of adding the ith term into the model.

The orthogonalization procedure for model term selection is usually implemented in a stepwise way, one term at a time. The *sum of error reduction ratio* (SERR) and the *error-to-signal ratio* (ESR) at step j due to $q_1, \cdots, q_j$ (or equally due to $\beta_1, \cdots, \beta_j$) are defined as

$$\mathrm{SERR}[j] = \sum_{i=1}^{j} \mathrm{ERR}[i] \tag{20}$$

$$\mathrm{ESR}[j] = \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{y}} = 1 - \sum_{i=1}^{j} \frac{g_i^2 q_i^T q_i}{\mathbf{y}^T\mathbf{y}} = 1 - \sum_{i=1}^{j} \mathrm{ERR}[i] = 1 - \mathrm{SERR}[j] \tag{21}$$

The selection procedure will be terminated when ESR of an identified model satisfies some specified conditions. In the present study, the following modified version of the generalized cross-validation (GCV) is considered as the criterion for model size determination

$$V(j) = \left(1 + \frac{2\lambda j}{N}\right) \mathrm{ESR}[j] \tag{22}$$

where $\lambda = \max\{1, \rho N\}$ and $0.002 \le \rho \le 0.01$. The selection procedure will be terminated at the step where the index function $V(j)$ is minimized.

5. A Three-Stage Modelling Procedure

As mentioned in Section 2, a typical NARMAX model often consists of two parts: the deterministic (noise independent) and the stochastic (noise correlated) submodels as shown by (2) and (3). In the present study, a three-stage modelling approach is proposed to construct multiscale RBF based NARMX models. The main idea of the three-stage modelling scheme is as follows:

- Initially, construct a multiscale based NARX model.

- The effects of correlated noise and unmeasured disturbances must be accommodated using in the model residuals (errors) from the identified NARX model. Viewing the modelling error $\varepsilon(t)$ as the output and treating the lagged system outputs $y(t-i)$ and inputs $u(t-j)$, coupled with the lagged error $\varepsilon(t-k)$, as the inputs, fit a polynomial model for the error $\varepsilon(t)$.
- Combine the identified error model with the network NARX model, and re-estimate all the model parameters recursively. An unbiased model should then be obtained.

Stage 1. Implementation of the NARX model using Multiscale RBF networks

For a given identification problem, the objective is to build a multiscale network to identify the unknown nonlinear mapping f_{yu} in (2). Assuming that N input-output data points, $\{u(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$ have been observed, let $\mathbf{x}(t)=[x_1(t),\cdots,x_d(t)]^T$ with $x_k(t)$ being defined by (4). The nonlinear function $f_{yu}(\mathbf{x}(t))$ can be approximated using a multiscale network model (13). Assume that a total of m_{yu} significant basis functions are selected so that the nonlinear function $f_{yu}(\mathbf{x}(t)) = f_{yu}(\mathbf{y}^{[t-1,n_y]},\mathbf{u}^{[t-1,n_u]})$ can then be approximated by

$$\hat{f}_{yu}(\mathbf{x}(t))=\sum_{m=1}^{m_{yu}}\hat{\theta}_{k_m}\phi_{k_m}(\mathbf{x}(t)) \tag{23}$$

Stage 2. Noise modelling

In many cases the noise terms in the NARMAX model (3) will form a correlated or coloured noise sequence. This is likely to be the case for most real data sets. In this case the approximation (23) is likely to fail to give a sufficient description due to the bias in the parameter estimates. The effects of correlated noise and unmeasured disturbances must then be characterized by modelling the residuals associated with the identified NARX model. The NARX modelling error is defined as

$$\varepsilon(t)=y(t)-\hat{f}_{yu}(\mathbf{x}(t)) \tag{24}$$

The residual signal $\varepsilon(t)$ can be related to the input $u(t)$ and the out $y(t)$ by a nonlinear model. In the present study, the following polynomial model of degree ℓ is applied to model the residual sequence as below

$$\begin{aligned}\varepsilon(t) &= f_{yue}(y^{[t-1,n_y]},u^{[t-1,n_u]},\varepsilon^{[t-1,n_e]})\\ &=\sum_{i_1=1}^{d}\gamma_{i_1}x_{i_1}(t)+\sum_{i_1=1}^{d}\sum_{i_2=i_1}^{d}\gamma_{i_1i_1}x_{i_1}(t)x_{i_2}(t))+\cdots\\ &+\sum_{i_1=1}^{d}\cdots\sum_{i_\ell=i_{\ell-1}}^{d}\gamma_{i_1i_2\cdots i_\ell}x_{i_1}(t)x_{i_2}(t)\cdots x_{i_\ell}(t)+\varepsilon_1(t)\end{aligned} \tag{25}$$

This form of model is used because if the system is nonlinear it is also highly likely that the noise will involve nonlinear cross product terms with both system input and the output. Assume that a total of m_{yue} significant

basis functions are selected for the noise model, the nonlinear function $f_{yue}(\tilde{\mathbf{x}}(t))$ $= f_{yue}(y^{[t-1,n_y]}, u^{[t-1,n_u]}, \varepsilon^{[t-1,n_e]})$ in (3) can then be approximated by

$$\hat{f}_{yue}(\tilde{\mathbf{x}}(t)) = \sum_{m=1}^{m_{yue}} \hat{\gamma}_{k_m} p_{k_m}(\tilde{\mathbf{x}}(t)) \tag{26}$$

where $p_m(\cdot)$ are selected model terms of the form $z_1^{i_1}(t)\cdots z_\ell^{i_\ell}(t)$, where $z_j^{i_j}(t) \in \{y(t-1),\cdots, y(t-n_y), u(t-1),\cdots, u(t-n_u), \varepsilon(t-1),\cdots, \varepsilon(t-n_e)\}$ for j=1, ..., ℓ, with $0 \le i_j \le \ell$ and $0 \le i_1 + \cdots + i_\ell \le \ell$. Note that at least one $z_j^{i_j}(t)$ is related to $\varepsilon(t-k)$ for $k = 1,2,\cdots,n_e$. The elements of the extended regression vector $\tilde{\mathbf{x}}(t)$ is defined as

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y+1 \le k \le n_y+n_u \\ \varepsilon(t-(k-n_y-n_u)) & n_y+n_u+1 \le k \le n_y+n_u+n_e \end{cases} \tag{27}$$

Stage 3. Parameter re-estimation

In order to obtain an unbiased network model, the identified model $\hat{f}_{yu}$ and $\hat{f}_{yue}$ should be combined as a whole and the model parameters should then be re-calculated. Let $\phi_1(t),\cdots,\phi_{m_{yu}}(t)$ be the m_{yu} selected model terms in (23) with $\phi_m(t) = \phi_m(\mathbf{x}(t))$, and let $\Phi_{yu}(t) = [\phi_1(t),\cdots,\phi_{m_{yu}}(t)]$ and $\Phi_{yue}(t) = [p_1(t),\cdots,p_{m_{yue}}(t)]$. An unbiased model can often be obtained by re-estimating the model parameters in a recursive way as described below.

(a) Calculate the model parameter estimate $[\hat{\theta}_{yu}, \hat{\theta}_{yue}]^T$ of the model

$$y(t) = \Phi_{yu}(t)\theta_{yu} + \Phi_{yue}(t)\theta_{yue} \tag{28}$$

Let

$$\varepsilon_1(t) = y(t) - \hat{y}(\tilde{\mathbf{x}}(t)) = y(t) - [\Phi_{yu}(t)\hat{\theta}_{yu} + \Phi_{yue}(t)\hat{\theta}_{yue}] \tag{29}$$

(b) If $\|\boldsymbol{\varepsilon}_1\|/\|\boldsymbol{\varepsilon}\| \approx 1$, stop the parameter re-estimation procedure; otherwise, go to (c).

(c) Set $\{\varepsilon(t)\}_{t=1}^N = \{\varepsilon_1(t)\}\}_{t=1}^N$, repeat (a).

Note that the residual signal defined by (24) and (29) is in fact the one-step-ahead prediction error, which is different from the often used model prediction error defined as

$$\hat{\varepsilon}(t) = y(t) - \hat{y}_{mpo}(t) \tag{30}$$

where $\hat{y}_{mpo}(t)$ is recursively calculated from an identified estimator $\hat{f}_{yu}$ from some given initial values in the sense that

$$\hat{y}(t) = \hat{f}_{yu}(\hat{y}_{mpo}(t-1), \cdots, \hat{y}_{mpo}(t-n_y), u(t-1), \cdots, u(t-n_u)) \tag{31}$$

A key step following the above three-stage modelling is model validation. A commonly used approach to check the validity of the identified model is to use higher order statistical correlation analysis [31]. An alternative for model validity tests is to check both the short and the long term predictive ability of the model.

6. Applications in Nonlinear Dynamical Modelling

Three examples, which are all related to real data sets, are described to illustrate the applicability and effectiveness of the new multiscale RBF network for the identification of nonlinear dynamical systems. In all the three examples, significant model terms were selected and the model size was determined using the orthogonal least squares algorithm given in Section 4. The performance of the identified networks will be measured by both the mean-squared-errors (MSE) and the normalized root-mean-squared-errors (NRMSE) defined as below:

$$\text{MSE} = \frac{1}{N_{test}} \sum_{t=1}^{N_{test}} [y(t) - \hat{y}(t)]^2 \tag{32}$$

$$\text{NRMSE} = \sqrt{\frac{\text{MSE}}{\text{SST}}} \tag{33}$$

where N_{test} is the length of a test data set, $y(t)$ and $\hat{y}(t)$ are the measurements and associated predictions, respectively, over the test data set, and $\text{SST} = (1/N_{test}) \sum_{t=1}^{N_{test}} [y(t) - \bar{y}]^2$ with $\bar{y} = (1/N_{test}) \sum_{t=1}^{N_{test}} y(t)$.

6.1 Example 1—a thermoplastic auxetic foam

An experimental data set relating to the testing and nonlinear design of a polyurethane foam has been obtained. This data set contained 1000 input-output data points and was used in this example to identify a mathematical model for this foam. The 1000 observations are shown in Fig.1, where the input, $u(t)$, indicates the displacement (unit: *mm*) of the servo hydraulic actuator of a machine, and the output, $y(t)$, indicates the force (unit: *k*N) due to the foam response. The 1000 input-output data points were split in two parts: the first 500 input-output points were used for model estimation, and the remaining points were used for model testing.

To construct a multiscale RBF network model on the basis of the estimation data set, the input vector for the network was chosen as $\mathbf{x}(t) = [u(t), u(t-1), u(t-2), u(t-3)]^T$. To save time for training the network, the 500 points in the estimation data set were classified into 100 groups using the *k*-means clustering algorithm. These clustered points, symbolized by $\mathbf{c}(m) = [c_1(m), c_4(m), c_3(m), c_4(m)]^T$ for m=1,2, ..., 100, were used as the candidate kernel centres. The multiscale parameters $\sigma_{m,k}^{(j)}$ in the network (9) were chosen as follows:

i) $\sigma_{0,u} = \{\max(u(t)) - \min(u(t))\}_{t=1}^{500} \approx 5.0$.

ii) $\sigma_u = 5\sigma_{0,u}$.

iii) $\sigma_{m,k}^{(j)} = 2^{-j}\sigma_u$, where m=1, 2, ..., 100; k=1,2,3,4; j=0,1,..., 6.

TABLE I
THE SELECTED CENTERS AND WIDTHS, ESTIMATED PARAMETERS AND THE ESR VALUES IN THE MULTISCALE RBF NETWORK MODEL FOR THE FOAM DATA SET IN EXAMPLE 1

Step j	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$	$c_{j,4}$	σ_j	θ_j	ESR[j]
1	-6.9515	-6.9345	-6.1589	-5.5324	1.5625	176.1613	0.2569
2	-6.3540	-4.5590	-4.0830	-3.9000	1.5625	-81.7875	0.2202
3	-7.3586	-6.9763	-6.6937	-6.4911	1.5625	-477.3835	0.1746
4	-6.6366	-6.9672	-6.8309	-6.1688	1.5625	-14.0225	0.1016
5	-6.9284	-6.1570	-5.5160	-5.4496	1.5625	-266.1140	0.0603
6	-5.3587	-5.7007	-5.0783	-4.4187	1.5625	50.2678	0.0522
7	-7.0130	-7.1230	-6.3540	-4.5590	3.1250	-215.3419	0.0391
8	-6.8717	-7.0758	-6.7815	-6.8248	1.5625	243.7460	0.0318
9	-5.4429	-5.8990	-5.6366	-5.9418	1.5625	125.3605	0.0202
10	-6.8116	-6.5477	-6.6869	-6.9347	1.5625	-81.3019	0.0148
11	-5.4473	-5.5985	-5.3282	-5.1270	1.5625	-115.6347	0.0125
12	-5.1766	-5.4931	-6.4767	-5.9013	3.1250	58.1285	0.0114
13	-6.4967	-6.3046	-5.6053	-5.1156	1.5625	163.4745	0.0108
14	-6.8020	-7.1775	-7.0770	-7.9285	1.5625	60.7656	0.0098
15	-5.4837	-4.8660	-5.3282	-5.1590	1.5625	-26.9015	0.0094
16	-6.9158	-7.3115	-7.1299	-6.5904	1.5625	327.4159	0.0090
17	-6.8101	-6.8680	-7.2617	-7.0924	1.5625	-200.8561	0.0084
18	-5.1088	-5.0690	-5.5297	-6.5767	1.5625	-57.2841	0.0079
19	-5.9052	-6.5587	-6.7583	-6.4984	1.5625	-63.3750	0.0075
20	-5.7650	-6.5678	-6.1002	-5.7891	1.5625	-87.3694	0.0071
21	-5.7310	-5.0598	-4.4220	-4.6662	1.5625	49.6644	0.0068
22	-7.3586	-6.9763	-6.6937	-6.4911	6.2500	-199.2928	0.0066
23	-6.0068	-6.4545	-7.0995	-7.1941	1.5625	319.4106	0.0062
24	-6.0565	-6.6102	-6.7566	-7.0200	1.5625	-333.0986	0.0058
25	-5.3976	-6.2704	-5.8128	-5.5299	3.1250	253.5972	0.0056
26	-6.6125	-6.6334	-6.2769	-6.5832	1.5625	237.6314	0.0055

Although a total of 700 candidate model terms were involved in the initial multiscale network model, only 26 significant model terms were selected using the orthogonal least squares algorithm. To eliminate the bias on the estimated parameters, the three-stage modelling approach was performed by setting an extended input vector as $\tilde{\mathbf{x}}(t) = [u(t), \cdots, u(t-3),\ \varepsilon(t-1), \cdots, \varepsilon(t-5)]^T$, and choosing a full noise model as (25) with a nonlinear degree ℓ =2. The finally identified multiscale RBF network model contained 25 process model terms and 9 noise related model terms. The kernel centres and widths, estimated model parameters and the ESR values defined by (21) are shown in Table 1, where the noise related model terms have been omitted. The MSE and NRMSE values with respect to model predicted outputs produced by the identified model over the estimation data set were 32.2881 and 0.1129, respectively, and the values were 33.8371 and 0.1136, respectively, over the test data set (points from 501 to 1000). The model predicted outputs and prediction errors are shown in Fig. 2.

Simulation studies showed that a standard Gaussian RBF network with a common single kernel width could also be fitted to the foam data set, but as expected more model terms were required in the final identified standard network model to achieve the same approximation accuracy as that produced by the new multiscale network model.

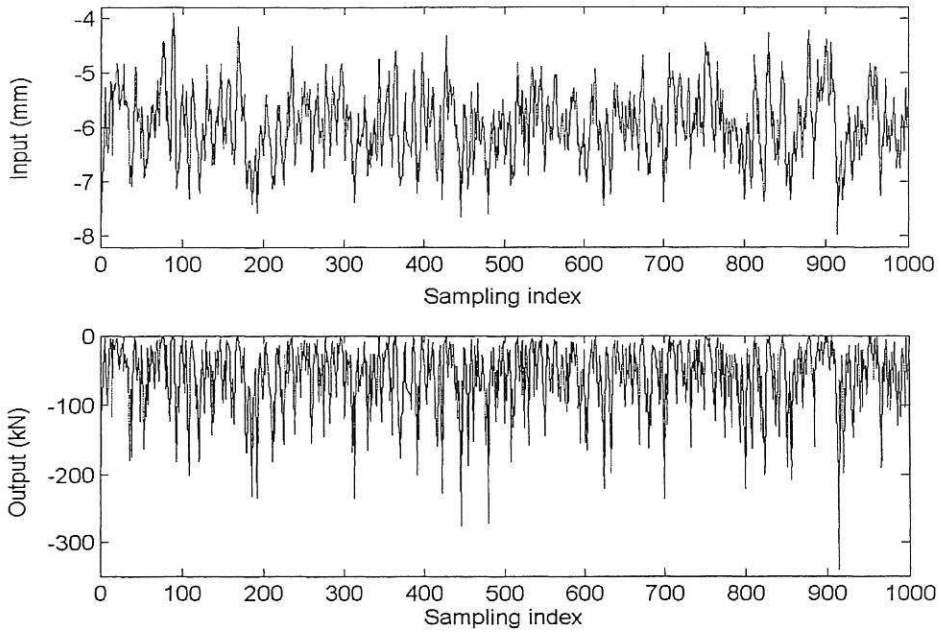


Fig. 1 The input and output for the thermoplastic auxetic foam described in Example 1.

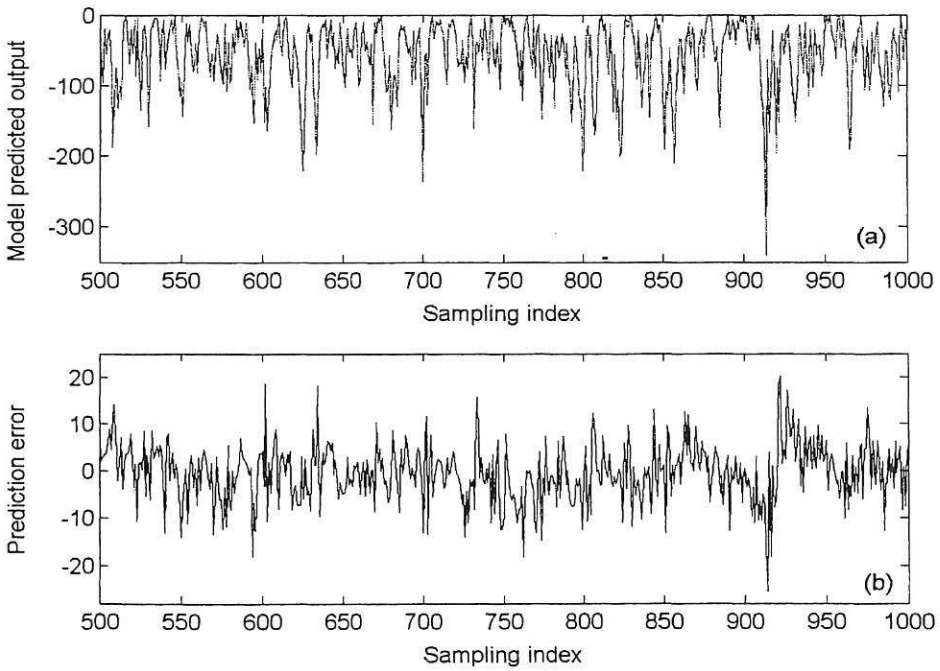


Fig. 2 The model predicted output and prediction error for the thermoplastic auxetic foam described in Example 1. The solid line in (a) indicates the measurements, and the dashed line indicates the model predicted outputs.

6.2 Example 2—a gas furnace system

Figure 3 shows 296 input-output data points to indicate the variation of the input gas feed rate and corresponding output CO_2 concentration measured as the percentage of the outlet gas from a gas furnace [32]. Previous results have shown that it was difficult for existing state-of-art regression techniques to train a standard Gaussian kernel based RBF network for this data set using a single common kernel width [23, 33]. In the following, however, it was shown that this data set can be well described using a multiscale RBF network.

All the 296 input-output data points were used for network training and the input vector was chosen as $\mathbf{x}(t) = [y(t-1), y(t-2), u(t-1), \cdots, u(t-4)]^T$, where u and y indicate the system input and output, respectively. The 296 points were classified into 50 groups using the k-means clustering algorithm. These clustered points, symbolized by $\mathbf{c}(m) = [c_1(m), \cdots, c_6(m)]^T$ for m=1,2, ..., 50, were used as the candidate kernel centres. The basis functions in the multiscale network model were of the form (12), and the scale parameters $\sigma_{y,m,k}^{(i)}$ and $\sigma_{u,m,k}^{(j)}$ were chosen as follows:

i) $\sigma_{0,y} = \{\max(y(t)) - \min(y(t))\}_{t=1}^{296} \approx 14.90$, and $\sigma_{0,u} = \{\max(u(t)) - \min(u(t))\}_{t=1}^{296} \approx 5.55$.

ii) $\sigma_y = 4\sigma_{0,y}$, and $\sigma_u = 4\sigma_{0,u}$.

iii) $\sigma_{y,m,k}^{(i)} = 2^{-i}\sigma_y$, and $\sigma_{u,m,k}^{(j)} = 2^{-j}\sigma_u$ with m=1, ..., 50; i=0,..., 4; j=0,1,..., 5; k=1,...,6.

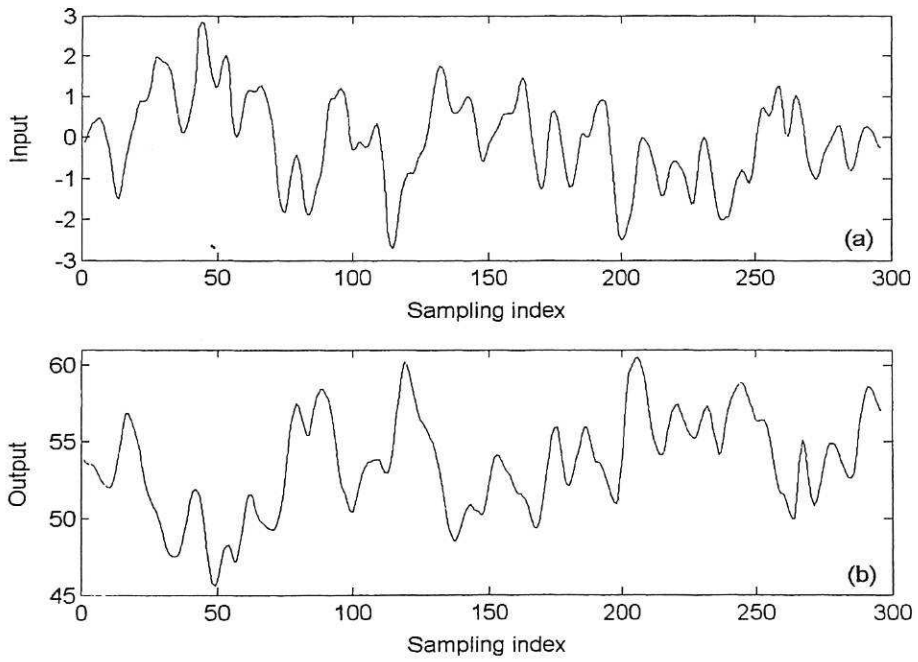


Fig. 3 The input (a) and output (b) signals for the gas furnace system described in Example 2.

TABLE 2
THE SELECTED CENTERS AND WIDTHS, ESTIMATED PARAMETERS AND THE ESR VALUES IN THE MULTISCALE RBF NETWORK MODEL FOR THE GAS FURNACE DATA SET IN EXAMPLE 2

Step j	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$	$c_{j,4}$	$c_{j,5}$	$c_{j,6}$	$\sigma_y^{(j)}$	$\sigma_u^{(j)}$	θ_j	ESR[j] $\times 10^4$
1	58.1556	57.0000	-1.6936	-2.0033	-2.1969	-2.2186	59.6	22.2	-0.1953	7.5522
2	50.0286	49.7143	-0.9813	-0.5891	-0.0743	0.4024	14.9	5.55	-5.5682	1.4321
3	53.6308	53.9231	0.3075	0.2727	0.1496	0.0266	59.6	2.775	2.9811	0.9386
4	51.6000	50.8000	-1.5203	-1.2613	-0.8615	-0.3533	7.45	5.55	7.0282	0.8339
5	54.6000	55.2750	0.6238	0.6906	0.7045	0.5946	7.45	22.2	-68.7932	0.7289
6	54.6900	54.1400	-1.7681	-1.7000	-1.4630	-1.0162	7.45	22.2	27.3027	0.6923
7	59.3375	59.3875	-0.7873	-1.0107	-1.3084	-1.6531	29.8	22.2	47.7286	0.5927
8	47.3333	47.5333	0.1637	0.4823	1.2050	1.8153	7.45	2.775	-2.3036	0.5595
9	59.3375	59.3875	-0.7873	-1.0107	-1.3084	-1.6531	7.45	2.775	0.5695	0.4600
10	55.1214	54.9857	0.1898	-0.0194	-0.2816	-0.5118	7.45	11.1	54.6500	0.3995
11	53.9778	52.9667	-1.1330	-1.3351	-1.2371	-0.9826	7.45	5.55	-16.3038	0.3261
12	52.8500	52.1875	-1.2479	-1.1481	-0.9204	-0.5880	14.9	11.1	-9.5552	0.2885
13	54.6900	54.1400	-1.7681	-1.7000	-1.4630	-1.0162	7.45	2.775	3.7231	0.2604
14	54.6900	54.1400	-1.7681	-1.7000	-1.4630	-1.0162	59.6	2.775	-2.6528	0.2493
15	55.9800	56.3333	0.1495	0.0626	-0.0877	-0.3893	14.9	5.55	-19.2078	0.2413
16	51.4200	50.8800	-0.4012	-0.4754	-0.4016	-0.0212	59.6	11.1	34.2387	0.2346
17	52.8500	52.1875	-1.2479	-1.1481	-0.9204	-0.5880	7.45	2.775	1.6988	0.2258

Although a total of 1500 candidate model terms were involved in the initial multiscale network model, only 17 significant model terms were selected using the orthogonal least squares algorithm. To eliminate the bias on model estimation, the three-stage modelling approach was performed by setting an extended input vector as $\tilde{\mathbf{x}}(t) = [y(t-1), y(t-2), u(t-1), \cdots, u(t-4), \varepsilon(t-1), \cdots, \varepsilon(t-10)]^T$, and choosing a full noise model as (25) with a nonlinear degree ℓ =1. This was equivalent to setting a linear noise model. The kernel centres and widths, estimated model parameters and the ESR values defined by (21) are shown in Table 2, where the noise related model terms have been omitted. The MSE and NRMSE values for one-step-ahead predictions were 0.0588 and 0.0759, respectively, and the two values for model predicted outputs were 0.7078 and 0.2632, respectively. The one-step-ahead prediction and the model residual are shown in Fig. 4, and the model predicted output and the prediction error are shown in Fig. 5. The result produced by the identified multiscale network model was nearly equivalent to that in [23], where a weighted boosting optimization algorithm was used to train a generalized Gaussian kernel model, and where the MSE value for one-step-ahead predictions was 0.054 but the MSE value for model predicted output was not given.

6.3 Example 3—a terrestrial magnetosphere dynamical system

The *Dst* index is used to measure the disturbance of the geomagnetic field during geomagnetic storms. The forecasting of the *Dst* index is very important in helping to prevent the negative effects of geomagnetic storms. Fig. 6 shows 800 data points of measurements of the solar wind parameter VB_s (input, measured unit: mV/m) and the *Dst* index (output, measured unit: nT) with a sampling interval T=1hour. Inspection of the Fig. 6 shows that several strong magnetospheric storms (*Dst* < -100 nT) and substrong stroms (*Dst* < -50 nT) took place during the time period under investigation. This data set was separated into the estimation set consisting of 400 input-output data points and the validation set consisting of the remaining data points. The objective was to

identify input-output nonlinear model based on the estimation data set. This model was then used to predict the *Dst* index.

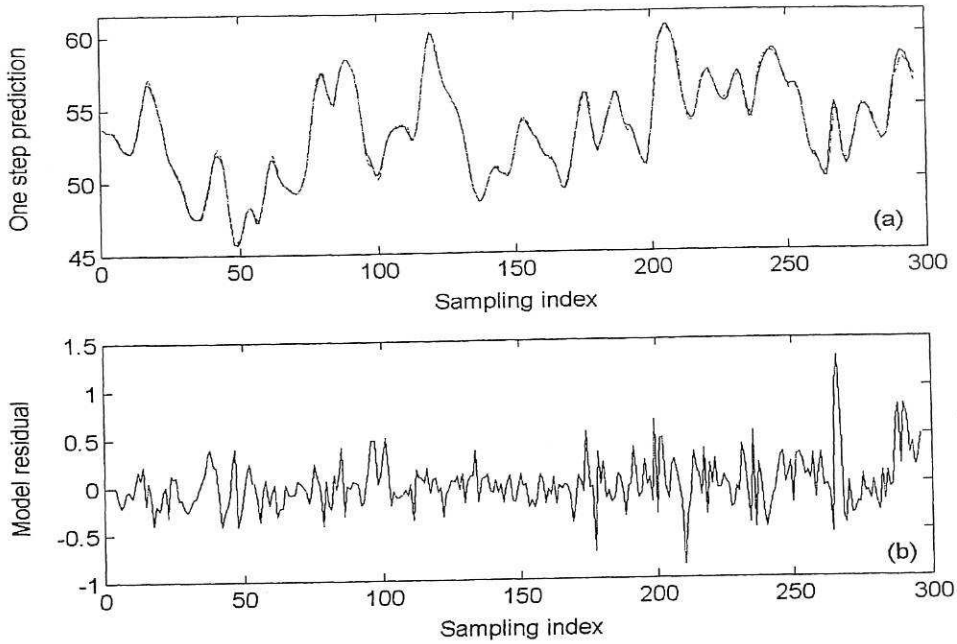


Fig. 4 The one-step-ahead prediction and model residual for the gas furnace data set described in Example 2. In (a), the solid line indicates the measurements, and the dashed line indicates one-step-ahead prediction.

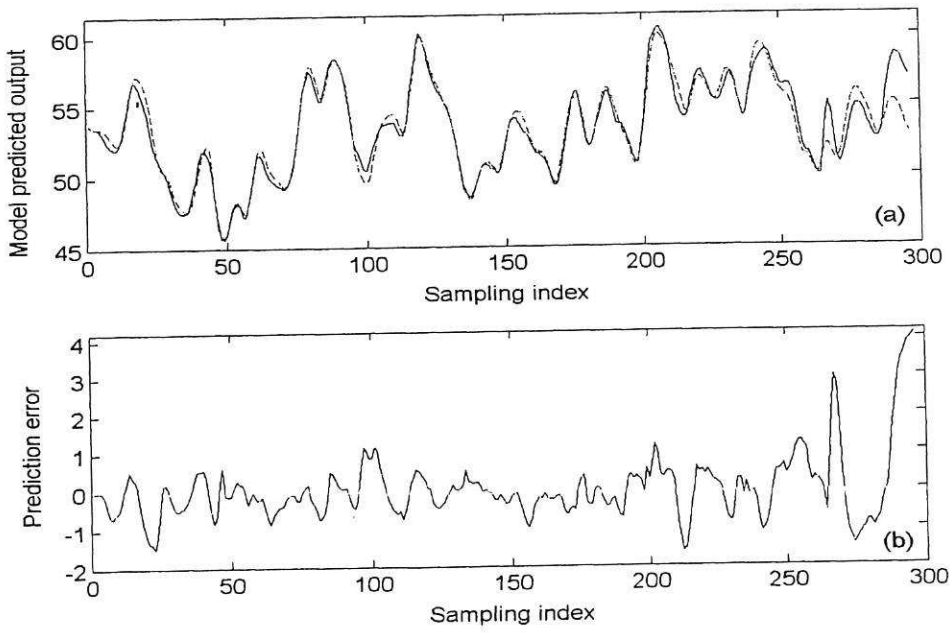


Fig. 5 The model predicted output and the prediction error for the gas furnace data set described in Example 2. In (a), the solid line indicates the measurements, and the dashed line indicates the model predicted output.

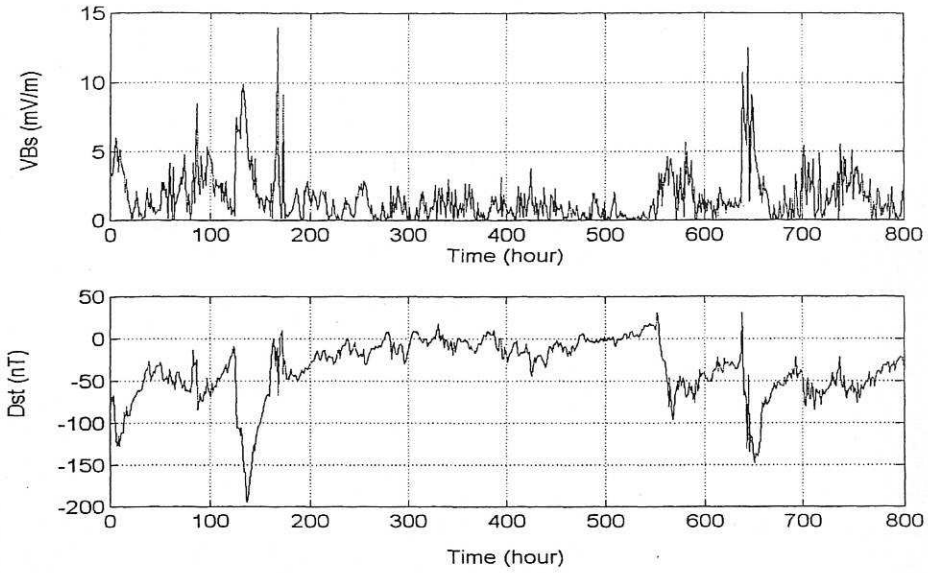


Fig. 6 The measurements for the input (VBs) and output (Dst) of the terrestrial magnetosphere dynamical system described in Example 3.

Previous studies have shown that the data set can be adequately fitted by choosing the input vector as $\mathbf{x}(t) = [y(t-1), \cdots, y(t-4), u(t-1), u(t-2)]^T$, where $y(\cdot)$ and $u(\cdot)$ indicate the measurements of the system output (*Dst*)and input (VB_s), respectively. For this data set, numerical experimental results show that it was difficult to train a standard Gaussian kernel based RBF network with only a single common kernel width. In fact, many different kernel widths have been tested to identify a standard network model using only a single common kernel width, but all the resulting models either involved too many model terms or produced poor generalization properties. The multiscale modelling framework, however, can be used to describe this data set.

The 400 points in the estimation data set were classified into 50 groups using the *k*-means clustering algorithm.These clustered points, symbolized by $\mathbf{c}(m) = [c_1(m), \cdots, c_6(m)]^T$ for m=1,2, ..., 50, were used as the candidate kernel centres. The basis functions in the multiscale network model were of the form (12), and the scale parameters $\sigma_{y,m,k}^{(i)}$ and $\sigma_{u,m,k}^{(j)}$ were chosen as follows:

i) $\sigma_{0,y} = \{\max(y(t)) - \min(y(t))\}_{t=1}^{400} \approx 200$, and $\sigma_{0,u} = \{\max(u(t)) - \min(u(t))\}_{t=1}^{400} \approx 15$.

ii) $\sigma_y = \sigma_{0,y}$, and $\sigma_u = \sigma_{0,u}$.

iii) $\sigma_{y,m,k}^{(i)} = 2^{-i}\sigma_y$, and $\sigma_{u,m,k}^{(j)} = 2^{-j}\sigma_u$ with m=1, ..., 50; i=0,..., 4; j=0,1,..., 4; k=1,...,6.

Although a total of 1250 candidate model terms were involved in the initial multiscale network model, only 16 significant model terms were selected using the orthogonal least algorithm. To eliminate the bias on the estimated parameters, the three-stage modelling approach was performed by setting an extended input vector as $\tilde{\mathbf{x}}(t) = [y(t-1), \cdots, y(t-4), u(t-1), u(t-2), \varepsilon(t-1), \cdots, \varepsilon(t-3)]^T$, and choosing a full noise model as (25) with a nonlinear degree ℓ =2. The kernel centres and widths, model parameters, estimated model parameters and the ESR values defined by (21) are shown in Table 3, where the noise related model terms were omitted. The two-hour-ahead prediction is shown in Fig. 7.

TABLE 3
THE SELECTED CENTERS AND WIDTHS, ESTIMATED PARAMETERS AND THE ESR VALUES IN THE MULTISCALE RBF NETWORK MODEL FOR THE MAGNEOSPHERE SYSTEM IN EXAMPLE 3

Step j	$c_{j,1}$	$c_{j,2}$	$c_{j,3}$	$c_{j,4}$	$c_{j,5}$	$c_{j,6}$	$\sigma_y^{(j)}$	$\sigma_u^{(j)}$	θ_j	ESR[j]
1	-159.6000	-145.8030	-126.1750	-113.0700	8.5046	9.4793	200	15	-181.2023	0.0807
2	-174.6194	-177.1959	-178.2351	-173.0579	4.7456	5.3740	50	15	-73.6629	0.0606
3	1.6930	2.2000	1.8676	0.1383	0.9820	0.8503	50	3.75	16.0137	0.0455
4	-63.1040	-49.3255	-66.6260	-66.9575	3.8423	4.6915	200	1.875	33.4176	0.0413
5	-129.5895	-124.1915	-105.9675	-102.3870	6.6749	7.1847	50	3.75	62.6576	0.0381
6	-21.1705	-47.6470	-60.9250	-64.6740	3.3863	0.9060	25	3.75	30.6993	0.0354
7	-0.7605	-14.4600	-402070	-68.3700	0.1095	1.9697	25	7.5	32.3289	0.0336
8	75.7863	-21.7183	-20.3560	-22.5823	5.9604	9.6060	50	15	-185.7594	0.0320
9	-10.5493	-20.1333	-50.0877	-35.1720	2.0550	2.6419	100	3.75	13.0834	0.0300
10	-75.7863	-21.7183	-20.3560	-22.5823	5.9604	9.6060	50	7.5	206.1002	0.0290
11	-47.1329	-43.5356	-28.5761	-26.6370	1.7431	2.0069	50	3.75	47.9005	0.0280
12	-39.5377	-38.1117	-47.1173	-55.1880	2.2130	2.2641	25	3.75	-19.8037	0.0273
13	-89.0360	-67.1750	-70.7150	-68.8700	4.9944	4.2966	100	1.875	-21.1147	0.0266
14	-75.7863	-21.7183	-20.3560	-22.5823	5.9604	9.6060	200	1.875	-79.2633	0.0261
15	-47.1329	-45.5356	-28.5761	-26.6370	1.7431	2.0069	100	1.875	-19.1104	0.0255
16	-43.3495	-51.3098	-54.4737	-55.2028	1.7041	1.6938	50	1.875	11.5482	0.0251

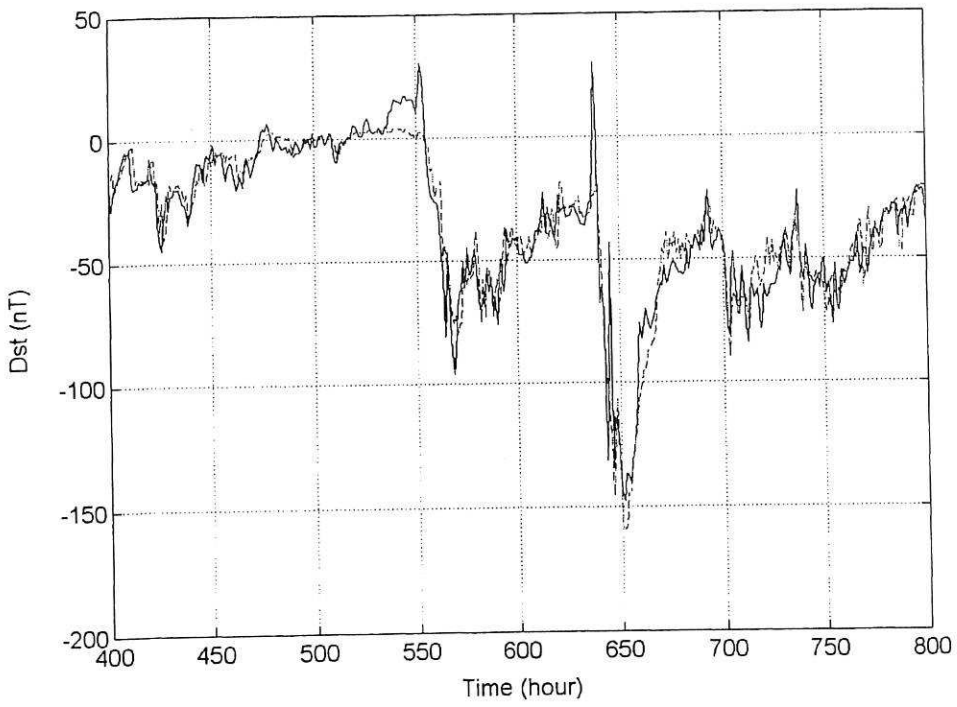


Fig. 7 A comparison between 2 hour ahead predictions and the measurements for the Dst index of the terrestrial magnetosphere dynamical system described in Example 3. The thin solid line indicates the measurements; the bold dashed lines indicate 2 hour ahead predictions.

7. Conclusions

There many cases where commonly used conventional Gaussian kernel based RBF networks may not work well to produce a reliable model with good generalization properties for nonlinear dynamical systems. This motivated the introduction of the multiscale concept to construct new RBF networks by observing the successful applications of multiresolution decompositions as in wavelet theory. Compared with commonly used conventional RBF networks, the new multiscale RBF networks are more flexible and can often provide parsimonious models for a wide class of nonlinear dynamical systems. Examples using three real data sets have clearly shown that the modelling capability of conventional RBF networks has been significantly enhanced by introducing the multiscale concept into RBF networks. In many cases the NARX model may fail to give a sufficient description due to effects of the noise signal, which may be a correlated or coloured noise sequence. The bias problem has been solved by performing the three-stage modelling procedure.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC(UK). We are grateful to Dr S. Sorrentino, Dr Z. Q. Lang and Professor G. Tomlinson for providing the thermoplastic auxetic foam data.

References

[1] E. J. Hartman, J. D. Keeler, and J. M. Kowalski, "Layered neural networks with Gaussian hidden units as universal approximations," *Neural Computation*, 1990, 2(2), pp. 210-215.

[2] T. Poggio and F. M. Girosi, "Regularization algorithms for learning that are equivalent to multiplayer networks," *Science*, 1990, 247(4945), pp. 978-982.

[3] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, 1991, 3(2), pp. 246-257.

[4] J. Park and I. W. Sandberg, "Approximation and radial basis function networks," *Neural Comput.*, 1993, 5(2), pp. 305-316.

[5] D. S. Broomhead and D. Lowe, " Multivariable functional interpolation and adaptive networks," *Complex Systems*, 1988, 2, pp. 321-355.

[6] T. Poggio and F. M. Girosi, "Networks for approximation and learning," *Proc. IEEE*, 1990, 78(9), pp. 1481-1497.

[7] T. Poggio and S. Edelman, "A network that learns to recognize three dimensional objects," *Nature*, 1990, 343, pp. 263-266.

[8] M. T. Musavi, W. Ahmed, K. H. Chan, K. B. Faris, and D. M. Hummels, "On the training of radial basis function classifiers," *Neural Networks*, 1992, 5(4), pp. 595-603.

[9] Y. S. Hwang and S. Y. Bang, " An efficient method to construct a radial basis function neural network classfier," Neural Networks, 1997, 10(8), pp. 1495-1503.

[10] F. Schwenker, H. A. Kestler, and G. Palm, "Three learning phase for radial-basis-function networks," *Neural Networks*, 2001, 14(4-5), pp. 439-458.

[11] S. Chen, S. A. Billings, C. F. N. Cowan, and P. M. Grant, "Practical identification of NARMAX models using radial basis functions," *Int. J. Control*, 1990, 52(6), pp. 1327-1350.

[12] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, 1991, 2(2), pp. 302-309.

[13] S. Chen and S. A. Billings, "Neural networks for nonlinear dynamic system modelling and identification," *Int. J. Control*, 1992, 56(2), pp. 319-346.

[14] S. A. Billings and C. F. Fung, "Recurrent radial basis function networks for adaptive noise cancellation," *Neural Networks*, 1995, 8(2), pp. 273-290.

[15] G. P. Liu, V. Kadirkamanathan, and S. A. Billings, "Varaible neural networks for adaptive control of nonlinear systems," *IEEE Trans. Syst. Man Cybernetic—Part C*, 1999, 29(1), pp. 34-43.

[16] S. Haykin, Neural Networks—A Comprehensive Foundation (2nd ed.). New Jersey: Prentice Hall, 1999.

[17] J. Moody and C. Darken, " Fast learning in networks of locally tuned processing units," *Neural Comput.*, 1989, 1, pp. 281-294.

[18] M. J. Orr, "Regularization in the selection of radial basis function centres," *Neural Comput.*, 1995, 7(3), pp. 606-623.

[19] D. Sanchez, "Second derivative dependent placement of RBF centres," *Neurocomputing*, 1995, 7(3), pp. 311-317.

[20] S. A. Billings and S. Chen, "The determination of multivariable nonlinear models for dynamic systems using neural networks," in *Neural Network Systems Techniques and Applications*, C.T. Leondes, Eds., San Diego: Academic Press, 1998, pp. 231-278.

[21] N. Benoudjit and M. Verleysen, " On the kernel widths in radial-basis function networks," *Neural Processing Letters*, 2003, 18(2), pp. 139-154.

[22] A. Saha and J. D. Keeler, " Algorithms for better representation and faster learning in radial basis function networks," *Advances in Neural Information Processing Systems*, 1989, 2, pp. 482-489.

[23] X. X. Wang, S. Chen and D. J. Brown, " An approach for constructing parsimonious generalized Gaussian kernel regression models," *Neurocomputing*, 2004, 62, pp. 441-457.

[24] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, 1988, 48(1), pp. 193-210.

[25] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, 1989, 49(6), pp.2157-2189.

[26] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, 1989, 50(5), pp.1873-1896.

[27] I. J. Leontaritis, and S. A. Billings, "Input-output parametric models for non-linear systems—part I: deterministic non-linear systems," *Int. J. Control*, 1985, 41(2), pp.303-328.

[28] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, 1989, 11(7), pp. 674-693.

[29] C. K. Chui, *An Introduction to Wavelets*. Boston: Academic Press, 1992.

[30] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Recognition*. (2nd ed.). New York: John Wiley & Sons, 2001.

[31] S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, 1986, 44(1), pp.235- 244.

[32] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*(3rd ed.). New Jersey: Prentice Hall, 1994.

[33] S. Chen, X. Hong, C. J. Harris, and P. M. Sharkey, "Sparse modelling using orthogonal forward regression with PRESS statistic and regulation," *IEEE Trans. Syst. Man Cybernetic—Part B*, 2004, 34(2), pp. 898-911.