Selective Ensemble of Multiple Local Model Learning for Nonlinear and Nonstationary Systems

Tong Liu[a,b,*], Sheng Chen[c,d,*], Shan Liang[a,b], Chris J. Harris[c]

[a] *Key Laboratory of Complex System Safety and Control (Ministry of Education), Chongqing University, Chongqing, 400044, China*
[b] *College of Automation, Chongqing University, Chongqing 400044, China*
[c] *School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*
[d] *King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Abstract

This paper proposes a selective ensemble of multiple local model learning for modeling and identification of nonlinear and nonstationary systems, in which the set of local linear models are self adapted to capture the newly emerging process characteristics and the prediction of the process output is also self adapted based on an optimally selected ensemble of subset linear local models. Specifically, our selective ensemble of multiple local model learning approach performs the model adaptation at two levels. At the level of local model adaptation, a newly emerging process state in the incoming data is automatically identified and a new local linear model is fitted to this newly emerged process state. At the level of online prediction, a subset of candidate local linear models are optimally selected and the prediction of the process output is computed as an optimal linear combiner of the selected subset local linear models. Two case studies involving chaotic time series prediction and modeling of a real-world industrial microwave heating process are used to demonstrate the effectiveness of our proposed approach, in comparison with other existing methods for modeling and identification of nonlinear and time-varying systems.

Key words: Nonlinear and time-varying system, online modeling and

*Corresponding author
Email addresses: liutong42@cqu.edu.cn (Tong Liu), sqc@ecs.soton.ac.uk (Sheng Chen), lightsun@cqu.edu.cn (Shan Liang), chrisharris57@msn.com (Chris J. Harris)

prediction, local model learning, selective ensemble

1. Introduction

Real-world systems often exhibit both nonlinear and nonstationary characteristics [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. For these time-varying nonlinear systems, batch global nonlinear modeling approaches [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26] become ineffective. Adaptive global nonlinear modeling of nonstationary processes is a challenging task, since both the model parameter values and the model structure must be adapted sufficiently fast in order to timely capture the changing nonlinear characteristics of the underlying process. However, most of the existing adaptive nonlinear modeling approaches do not perform online nonlinear model structure updating and they only use some recursive estimators, such as the recursive least squares (RLS) algorithm, to adapt the model parameter values [27, 28, 29, 30, 31, 32, 33, 34]. In particular, if the system's input space or operating region is known a priori, by covering the input space with sufficiently dense fixed nodes, the extreme learning machine (ELM) for single-hidden-layer neural networks [30, 31, 32] only needs to sequentially update the model weights using the RLS algorithm. Because the size of the nonlinear model has to be very large for ELM, online adaptation of the model weights is computationally costly and, moreover, it takes time to sufficiently change the model weights to match the changing nonlinear characteristics of the underlying process. Therefore, the online sequential ELM (OS-ELM) only works well for relatively slow time varying nonlinear processes with the known operating regions. In an attempt to improve the performance of OS-ELM in nonstationary environment, the work [35] proposed a time-varying OS-ELM (OS-ELM-TV), whose weights are function of time. Specifically, each weight of the OS-ELM-TV is a linear combination of a set of basis functions.

However, during the online operation of a time-varying industrial process, the process dynamics can vary significantly and the process may enter a new operating region which is completely outside the initial modeling space. This

will degrade the performance of the fixed-structure nonlinear modeling methods, such as the OS-ELM and OS-ELM-TV. In order to capture the newly emerging process's dynamics, the classical resource-allocating network (RAN) technique [36, 37] adapts the nonlinear model structure by growing the set of radial basis function (RBF) nodes, starting from an empty set of RBF nodes. By contrast, starting from an initial set of RBF nodes, the fast tunable RBF method [38] adjust RBF nodes as well as the model weights online to adaptively model nonstationary systems. The experimental results of [38] show that this fast tunable RBF method typically outperforms the RAN.

A well-known alternative to nonlinear modeling with a single global model is to adopt the multiple local models, which are capable of capture severe nonlinearity too [39, 40, 41, 42, 43]. The essence of multiple local modeling is to 'partition' the model input space into multiple 'regions', each covered by a local model. With a sufficiently fine partitioning, the characteristics of the process in each local region can be accurately modeled with a simple linear model. Moreover, in order to capture the newly emerging nonlinear characteristics of a time-varying system, an adaptive local modeling method must be able to grow its local models. The multiple local modeling framework of [42, 43] however does not have this capability, as it employs a fixed set of local RBF models. In the online soft sensor design, this capability of adaptively growing the set of local models has been demonstrated to be vital to achieve excellent performance in online soft sensing [44, 45]. This motivates our current work.

Against the above background, in this paper, we propose a selective ensemble of multiple local model learning approach for nonlinear and time-varying systems, in which the set of local linear models are self adapted to capture the newly emerging process state, and the prediction of the process output is also adapted based on an optimally selected ensemble of subset linear local models. Similar to the works of [44, 45], which consider a very different application of soft sensor design, our proposed selective ensemble of multiple local model learning approach performs the model adaptation at two levels. At the level of local model adaptation, a newly emerging process state in the incoming data

is automatically identified and a new local linear model is fitted to this newly emerged process state. At the level of online prediction or modeling, a subset of candidate local linear models are optimally selected and the prediction of the process output is computed as an optimal linear combiner of the selected subset local linear models. Noted that different from the work [43] which employs a fixed set of sub-models and only computes the weights of the ensemble online, the proposed method continuously learns newly emerging process states and identifies new sub-models accordingly. To our best knowledge, this work is the first to apply selective ensemble of multiple local linear model learning for modeling and identification of nonlinear and nonstationary systems.

The remainder of the paper is organized as follows. Section 2 is entirely dedicated to our selective ensemble of multiple local model learning approach, which includes adaptation of the local linear models and selective ensemble of local linear models for online prediction. Extensive experimental results are presented in Section 3, which includes the two case studies of online time series prediction involving Lorenz chaotic time series [46] and modeling of a real-world industrial microwave heating process (MHP) [47, 48, 49]. Our conclusions and future research directions are provided in Section 4.

2. Selective Ensemble of Multiple Local Model Learning Approach

To effectively model nonlinear and nonstationary systems, the proposed selective ensemble of multiple local model learning carries out two levels of model adaptation: (1) adaptation of the local linear model set, and (2) online adaptation of model prediction. We now detail these two components.

2.1. Adaptation of local linear model set

Consider the data sample set $\{\boldsymbol{x}(t), y(t)\}_{t=1}^{N}$ drawn from a process, where $\boldsymbol{x}(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}$ are the system's input vector and output, respectively. Assume that the nonlinear characteristics of the system over $\{\boldsymbol{x}(t), y(t)\}_{t=1}^{N}$ can be represented by the L local process states. Then the task is to automatically

construct the local linear models $\{f_l\}_{l=1}^{L}$ that are valid in their corresponding process states represented by their respective sub-datasets $\{\boldsymbol{X}_l, \boldsymbol{y}_l\}_{l=1}^{L}$, where each $\boldsymbol{X}_l$ contains W consecutive time samples of the input $\boldsymbol{x}(t)$ and $\boldsymbol{y}_l$ consists of the corresponding output samples.

Without loss of generality, let a data window $\mathcal{W}_{\text{ini}} = \left\{\boldsymbol{X}_{\text{ini}} \in \mathbb{R}^{W\times m}, \boldsymbol{y}_{\text{ini}} \in \mathbb{R}^{W}\right\}$ with W consecutive time samples $\{\boldsymbol{x}(t), y(t)\}_{t=t_{\text{ini}}}^{t_{\text{ini}}+W}$ be initially set, and a local linear model f_{ini} is built on it as

$$\widehat{\boldsymbol{y}}_{\text{ini}} = f_{\text{ini}}\big(\boldsymbol{X}_{\text{ini}}\big) = \boldsymbol{\Phi}\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\Phi} = \big[\mathbf{1}_W \ \boldsymbol{X}_{\text{ini}}\big] \in \mathbb{R}^{W\times(1+m)}$ and $\mathbf{1}_W$ denotes the W-dimensional vector whose elements are all one, while the model parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{(1+m)}$ is readily given by the least square (LS) estimate as

$$\boldsymbol{\beta} = \big(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\big)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}_{\text{ini}}. \tag{2}$$

The prediction error or residual vector of this local model over $\mathcal{W}_{\text{ini}}$ is given by

$$\boldsymbol{e}_{\text{ini}} = \boldsymbol{y}_{\text{ini}} - f_{\text{ini}}\big(\boldsymbol{X}_{\text{ini}}\big) \in \mathbb{R}^{W}. \tag{3}$$

After an initial local model f_{ini} is built, a shifted window $\mathcal{W}_{\text{sft}} = \left\{\boldsymbol{X}_{\text{sft}}, \boldsymbol{y}_{\text{sft}}\right\}$ is sequentially obtained by moving the window one step ahead, that is, $\mathcal{W}_{\text{sft}}$ contains the samples $\{\boldsymbol{x}(t), y(t)\}_{t=t_{\text{ini}}+1}^{t_{\text{ini}}+1+W}$. If the two local regions $\mathcal{W}_{\text{ini}}$ and $\mathcal{W}_{\text{sft}}$ are not significantly different, it can be considered that the process data within $\mathcal{W}_{\text{sft}}$ follow the same distribution as in $\mathcal{W}_{\text{ini}}$ and the window is continued to be shifted forward. Otherwise, $\mathcal{W}_{\text{sft}}$ is considered to represent a new operating mode different from the previous mode, and a new local linear model f_{new} should be developed based on $\mathcal{W}_{\text{sft}}$. Determining whether $\mathcal{W}_{\text{ini}}$ and $\mathcal{W}_{\text{sft}}$ are significantly different or not can be naturally casted as statistical hypothesis testing [51].

Specifically, let the estimation error vector produced by f_{ini} on $\mathcal{W}_{\text{sft}}$ be de-

fined as

$$\boldsymbol{e}_{\text{sft}} = \boldsymbol{y}_{\text{sft}} - f_{\text{ini}}\big(\boldsymbol{X}_{\text{sft}}\big). \tag{4}$$

Then whether $\mathcal{W}_{\text{ini}}$ and $\mathcal{W}_{\text{sft}}$ are similar or not can then be turned into the equivalent hypothesis testing that tests whether $\boldsymbol{e}_{\text{ini}}$ and $\boldsymbol{e}_{\text{sft}}$ are significantly different or not. Since f_{ini} is a linear model, $\boldsymbol{e}_{\text{ini}}$ and $\boldsymbol{e}_{\text{sft}}$ are considered not significantly different when both their means, μ_{ini} and μ_{sft}, as well as variances, σ^2_{ini} and σ^2_{sft}, are the same. Therefore, the two null hypotheses can be set to

$$H_0^{\mu} : \mu_{\text{ini}} = \mu_{\text{sft}}, \tag{5}$$

$$H_0^{\sigma^2} : \sigma^2_{\text{sft}} = \sigma^2_{\text{ini}}. \tag{6}$$

The mean μ_{ini} and variance σ^2_{ini} are estimated based on $\boldsymbol{e}_{\text{ini}}$, while μ_{sft} and σ^2_{sft} are estimated based on $\boldsymbol{e}_{\text{sft}}$. Since f_{ini} is an unbiased estimator, we have $\mu_{\text{ini}} = 0$ and $\sigma^2_{\text{ini}} = \frac{1}{W-1}\boldsymbol{e}^{\text{T}}_{\text{ini}}\boldsymbol{e}_{\text{ini}}$. Assuming that $\boldsymbol{e}_{\text{ini}}$ and $\boldsymbol{e}_{\text{sft}}$ follow normal distribution, the T and χ^2 statistics can be constructed as [51]

$$T = \sqrt{W}\big(\mu_{\text{sft}} - \mu_{\text{ini}}\big)/\sigma_{\text{sft}}, \tag{7}$$

$$\chi^2 = (W-1)\sigma^2_{\text{sft}}/\sigma^2_{\text{ini}}. \tag{8}$$

According to the statistical theory [51], if the hypotheses H_0^{μ} and $H_0^{\sigma^2}$ are both valid, the T statistic (7) and χ^2 statistic (8) follow the t distribution and χ^2 distribution with the degree of freedom $W-1$, respectively. Thus, the t-test and χ^2-test can be utilized to test the above two hypotheses. Specifically, the conditions of accepting H_0^{μ} and $H_0^{\sigma^2}$ are

$$|T| < \lambda_t \text{ and } \chi^2 < \lambda_{\chi}, \tag{9}$$

where λ_t is the threshold of the T statistic for the given significance level α_t which satisfies $\Pr\{|T| < \lambda_t\} = 1-\alpha_t$, while λ_{χ} is the threshold of the χ^2 statistic for the given significance level α_{χ}, which satisfies $\Pr\{\chi^2 < \lambda_{\chi}\} = 1-\alpha_{\chi}$.

Let the local model set contain $L > 1$ independent local linear models $\{f_l\}_{l=1}^{L}$, and $f_{\text{ini}} = f_L$. When one or both conditions of (9) are violated, $\mathcal{W}_{\text{ini}}$ and $\mathcal{W}_{\text{sft}}$ are significantly different, and the new local linear model $f_{\text{new}} = f_{\text{sft}}$ is identified, which is different from f_L. We need to test whether f_{new} is different from the other models f_l for $1 \leq l \leq L-1$. This task can also be fulfilled based on the statistical hypothesis testing. Let the predicted errors of $\{\boldsymbol{X}_{\text{sft}}, \boldsymbol{y}_{\text{sft}}\}$ based on f_{new} and f_l be defined respectively by

$$\boldsymbol{e}_{\text{new}} = \boldsymbol{y}_{\text{sft}} - f_{\text{new}}(\boldsymbol{X}_{\text{sft}}), \tag{10}$$

$$\boldsymbol{e}_l = \boldsymbol{y}_{\text{sft}} - f_l(\boldsymbol{X}_{\text{sft}}),\ 1 \leq l \leq L-1. \tag{11}$$

With the assumption that $\boldsymbol{e}_{\text{new}}$ and $\boldsymbol{e}_l$ follow normal distribution, the T and χ^2 statistics are constructed according to

$$T_l = \sqrt{W}(\mu_l - \mu_{\text{new}})/\sigma_l, \tag{12}$$

$$\chi_l^2 = (W-1)\sigma_l^2/\sigma_{\text{new}}^2, \tag{13}$$

where μ_{new} and σ_{new}^2 are the mean and variance of $\boldsymbol{e}_{\text{new}}$, which can be estimated using $\boldsymbol{e}_{\text{new}}$, while μ_l and σ_l^2 are the mean and variance of $\boldsymbol{e}_l$, which can be estimated in the same way. Based on the statistical theory [51], if the null hypotheses

$$H_l^{\mu} :\ \mu_l = \mu_{\text{new}}, \tag{14}$$

$$H_l^{\sigma^2} :\ \sigma_l^2 = \sigma_{\text{new}}^2, \tag{15}$$

are both valid, the T_l statistic in (12) and χ_l^2 statistic in (13) follow the t distribution and χ^2 distribution with the degree of freedom $W-1$, respectively. Therefore, if there exist an $l \in \{1, 2 \ldots L-1\}$ such that

$$|T_l| < \lambda_t \text{ and } \chi_l^2 < \lambda_\chi, \tag{16}$$

the hypotheses (14) and (15) are both valid, and $\boldsymbol{e}_{\text{new}}$ and $\boldsymbol{e}_l$ are regarded to be identical. Consequently, f_{new} and f_l are the same model, and one of them should be removed. Since f_l is 'older' than f_{new}, we keep the local model f_{new} and delete f_l. On the other hand, if one or both conditions are violated $\forall l \in \{1, 2 \dots L-1\}$, f_{new} is different from f_l for $1 \leq l \leq L$. Thus, we have identified a new process state, and we add f_{new} to the local model set by setting $L = L + 1$ and $f_L = f_{\text{new}}$.

Remark 1. *Although the aforementioned procedure seems to describe offline training, this local learning strategy can readily operate online. Specifically, during online operation, when the newest data sample $\{\boldsymbol{x}(t), y(t)\}$ is available, the data window $\mathcal{W}_{\text{sft}}$ can be shifted one sample ahead, and the corresponding learning procedure can then be carried out.*

The proposed online adaptive local model set development procedure is summarized in Algorithm 1. The significance levels in the statistical testings are typically set to $\alpha_t = 0.05$ and $\alpha_\chi = 0.05$ [51]. The window size is a key algorithmic parameter of Algorithm 1. A small W may lead to large number of local models, which will increase online operating time, but it may result in better nonstationary adaptation capability. A large W has the opposite efforts. The effects of the window size W to the achievable performance will be further investigated in the simulation study.

2.2. Adaptation of model prediction

After the online operation at time sample t, Algorithm 1 produces the local model set of $\{f_l\}_{l=1}^{L}$. At the next time sample of $t_{\text{next}} = t+1$, the task of online modeling update is to produce the model prediction $\widehat{y}(t_{\text{next}})$ for the process's true output $y(t_{\text{next}})$, given the process input $\boldsymbol{x}(t_{\text{next}})$ and the available local model set $\{f_l\}_{l=1}^{L}$. One way of generating this online prediction is to produce a mixture of experts by combining all the local linear models [52, 53, 54]. However, there exist evidence in literature that combining part of the ensemble models rather than all of them may achieve better performance [55, 56]. Therefore,

Algorithm 1 Adaptation of local linear model set

Initialization
Collect $\mathcal{W}_{\text{ini}}$ with W consecutive samples from historical data, and construct the LS linear model f_{ini} on $\mathcal{W}_{\text{ini}}$.
Calculate $\boldsymbol{e}_{\text{ini}}$, and estimate μ_{ini} and σ^2_{ini}.
Set $L = 1$, $\{\mathcal{W}_L, f_L\} = \{\mathcal{W}_{\text{ini}}, f_{\text{ini}}\}$ and $\mathcal{W}_{\text{sft}} = \mathcal{W}_L$.
Step 1: New local model detection
When a new data sample is available, shift $\mathcal{W}_{\text{sft}}$ one sample ahead.
Calculate $\boldsymbol{e}_{\text{sft}}$, and estimate μ_{sft} and σ^2_{sft}.
Construct T and χ^2 statistics using (7) and (8).
If both conditions of (9) are satisfied
  Go to **Step 1**.
End if
Construct the LS linear model f_{sft} on $\mathcal{W}_{\text{sft}}$.
Set $\mathcal{W}_{\text{new}} = \mathcal{W}_{\text{sft}}$ and $f_{\text{new}} = f_{\text{sft}}$.
Calculate $\boldsymbol{e}_{\text{new}}$, and estimate μ_{new} and σ^2_{new}.
Step 2: Redundant local model deletion
For $l = 1, 2, \ldots, L-1$
  Compute $\boldsymbol{e}_l$, and estimate μ_l and σ^2_l.
  Construct T_l and χ^2_l statistics using (12) and (13).
  If both conditions of (16) are satisfied
    Delete f_l, set $f_i = f_{i+1}$ for $i = l, l+1, \cdots, L-1$,
    set $L = L - 1$, then go to **Step 3**.
  End if
End for
Step 3: Add new local model
Set $L = L + 1$, $\mathcal{W}_L = \mathcal{W}_{\text{new}}$ and $f_L = f_{\text{new}}$.
Return to **Step 1**.

we adopt a selective ensemble of local linear models from the local model set $\{f_l\}_{l=1}^{L}$ and compute the selective-ensemble based online prediction using the $p(>1)$ latest labeled data $\{\boldsymbol{x}(t-i), y(t-i)\}_{i=0}^{p-1}$.

Let $\boldsymbol{e}_l(t) = [e_l(t)\ e_l(t-1) \cdots e_l(t-p+1)]^{\mathrm{T}}$ be the modeling error vector of the lth local linear model f_l on the available data set $\{\boldsymbol{x}(t-i), y(t-i)\}_{i=0}^{p-1}$, which is given by

$$e_l(t-i) = y(t-i) - f_l(\boldsymbol{x}(t-i)),\ 0 \le i \le p-1. \tag{17}$$

The performance metric of the lth local model is defined as

$$J_l(t) = \|\boldsymbol{e}_l(t)\|^2. \tag{18}$$

By further defining

$$J_{l_{\max}}(t) = \max_{1 \le l \le L} J_l(t), \tag{19}$$

we can normalize the performance metrics of (18) to

$$\bar{J}_l(t) = \frac{J_l(t)}{J_{l_{\max}}(t)},\ 1 \le l \le L. \tag{20}$$

Obviously, $\bar{J}_l(t) \in (0,\ 1]$. Clearly, the best local model, whose index $l_1 = l_{\min}$ is given by

$$l_{\min} = \arg \min_{1 \le l \le L} \bar{J}_l(t), \tag{21}$$

should be selected. Moreover, other local models whose performance metrics (20) are below a given threshold $0 < \varepsilon \le 1$ are also selected. Note that if we set $\varepsilon = 1$, all the L local models are selected, while if the threshold is chosen to be $\varepsilon \le J_{l_{\min}}(t)$, only the best local model f_{l_1} is selected.

Assume that $M(\ge 1)$ local models are selected at time t for predicting the system output at t_{next}, and the indexes of the selected local models are given in

the index set Γ as

$$\Gamma = \{l_1, l_m | 2 \leq m \leq M, J_{l_m}(t) \leq \varepsilon, 1 \leq l_m \leq L\}. \tag{22}$$

This selection procedure yields the M local model outputs

$$\widehat{y}_{l_m}(t-i) = f_{l_m}(\boldsymbol{x}(t-i)),\ 1 \leq m \leq M, \tag{23}$$

for $0 \leq i \leq p-1$. The estimate $\widehat{y}(t-i)$ of the process output $y(t-i)$ is given as the weighted sum of the M selected subset models, which is computed by

$$\widehat{y}(t-i) = \sum\nolimits_{m=1}^{M} \theta_m(t) \widehat{y}_{l_m}(t-i),\ 0 \leq i \leq p-1, \tag{24}$$

where nonnegative $\theta_m(t)$ is the combining coefficient for the mth selected local model, and the combining coefficients must satisfy the constraint of

$$\sum\nolimits_{m=1}^{M} \theta_m(t) = 1. \tag{25}$$

The estimation errors

$$e(t-i) = y(t-i) - \widehat{y}(t-i),\ 0 \leq i \leq p-1, \tag{26}$$

are utilized to determine the combining coefficients.

Specifically, the optimal combining coefficients can be obtained by minimizing the LS cost function

$$V(t) = \frac{1}{2} \sum\nolimits_{i=0}^{p-1} e^2(t-i), \tag{27}$$

subject to the constraint (25). Because of $\sum_{m=1}^{M}\theta_m(t)=1$,

$$\begin{aligned}V(t)=&\frac{1}{2}\sum\nolimits_{i=0}^{p-1}\left(y(t-i)-\sum\nolimits_{m=1}^{M}\theta_m(t)\widehat{y}_{l_m}(t-i)\right)^2\\=&\frac{1}{2}\sum\nolimits_{i=0}^{p-1}\left(\sum\nolimits_{m=1}^{M}\theta_m(t)y(t-i)-\sum\nolimits_{m=1}^{M}\theta_m(t)\widehat{y}_{l_m}(t-i)\right)^2\\=&\frac{1}{2}\sum\nolimits_{i=0}^{p-1}\left(\sum\nolimits_{m=1}^{M}\theta_m(t)e_{l_m}(t-i)\right)^2=\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t),\end{aligned}\tag{28}$$

where $\boldsymbol{\theta}(t)=\left[\theta_1(t)\cdots\boldsymbol{\theta}_M(t)\right]^{\mathrm{T}}$ and $\bar{\boldsymbol{E}}(t)$ is the estimated error covariance matrix which is given as

$$\bar{\boldsymbol{E}}(t)=\sum_{i=0}^{p-1}\begin{bmatrix}e_{l_1}^2(t-i) & \cdots & e_{l_1}(t-i)e_{l_M}(t-i)\\ \vdots & \ddots & \vdots\\ e_{l_1}(t-i)e_{L_M}(t-i) & \cdots & e_{l_M}^2(t-i)\end{bmatrix}.\tag{29}$$

The problem of determining the optimal $\boldsymbol{\theta}(t)$ can then be formulated as the following optimization

$$\begin{array}{ll}\min\limits_{\boldsymbol{\theta}} & \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t),\\ \text{s.t.} & \sum_{m=1}^{M}\theta_m(t)=1.\end{array}\tag{30}$$

The Lagrangian function for the optimization (30) is given by

$$L\big(\boldsymbol{\theta}(t);\gamma\big)=\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t)+\gamma\big(\mathbf{1}_M^{\mathrm{T}}\boldsymbol{\theta}(t)-1\big),\tag{31}$$

where $\gamma>0$ is Lagrange multiplier. Letting $\frac{\partial}{\partial\boldsymbol{\theta}(t)}L=\mathbf{0}_M$ yields

$$\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t)+\gamma\mathbf{1}_M=\mathbf{0}_M,\tag{32}$$

where $\mathbf{0}_M=[0\cdots0]^{\mathrm{T}}\in\mathbb{R}^M$. This suggests that the optimal combining vector $\widehat{\boldsymbol{\theta}}$ can be obtained as follows. First, calculate

$$\widetilde{\boldsymbol{\theta}}(t)=\bar{\boldsymbol{E}}^{-1}(t)\mathbf{1}_M,\tag{33}$$

which is followed by the normalization

$$\widehat{\theta}_m(t) = \frac{1}{\sum_{j=1}^{M} \widetilde{\theta}_j(t)} \widetilde{\theta}_m(t),\ 1 \le m \le M. \tag{34}$$

The prediction $\widehat{y}(t_{\text{next}})$ for the process's true output $y(t_{\text{next}})$ is produced as the selected ensemble

$$\widehat{y}(t_{\text{next}}) = \sum\nolimits_{m=1}^{M} \widehat{\theta}_m(t) f_{l_m}\left(\boldsymbol{x}(t_{\text{next}})\right) \tag{35}$$

Algorithm 2 summarizes the online prediction and adaptive modeling operations. The choice of p trades off the computational complexity and the robustness against noise. The threshold ε is another algorithmic parameter of Algorithm 2 that trades off performance with computational complexity. How p and ε influence the achievable performance will be further investigated in the simulation study.

Algorithm 2 Online prediction and adaptive modeling

1: **Initialization**
2: At the beginning of online operation, the local model set $\{\mathcal{W}_l, f_l\}_{l=1}^{L}$ has been constructed.
3: Set $\{\mathcal{W}_L, f_L\} = \{\mathcal{W}_{\text{ini}}, f_{\text{ini}}\}$ and $\mathcal{W}_{\text{sft}} = \mathcal{W}_L$.
4: **Step 1: Online prediction**
5: Give input $\boldsymbol{x}(t_{\text{next}})$ at new sample time $t_{\text{next}} = t + 1$.
6: Calculate the performance metrics $\bar{J}_l(t)$ using (20) for $1 \le l \le L$ on past p data points.
7: Select the subset models with the index set Γ of (22).
8: Calculate the error covariance matrix $\bar{\boldsymbol{E}}(t)$ using (29).
9: Calculate the optimal combining coefficients $\widehat{\boldsymbol{\theta}}(t)$ using (33) and (34).
10: Predict true process output $y(t_{\text{next}})$ with the selective ensemble (35).
11: Carry out other unrelated online operations.
12: **Step 2: Online model adaptation**
13: When the observation $y(t_{\text{next}})$ is available, add $\{\boldsymbol{x}(t_{\text{next}}), y(t_{\text{next}})\}$ to the dataset with $t = t + 1$.
14: Shift $\mathcal{W}_{\text{sft}}$ one sample ahead, and perform relavent local model set adaptation.
15: Set $t_{\text{next}} = t_{\text{next}} + 1$, and go to **Step 1**.

3. Two Case Studies

Two case studies involving chaotic time series prediction and modeling of a real-world industrial MHP are used to evaluate the proposed selective ensemble of multiple local model learning approach. The well-known online modeling algorithms, the RAN [37], the OS-ELM with sigmoid nodes (OS-ELM (sigmoid)), the OS-ELM with RBF nodes (OS-ELM (RBF)) [30, 31, 32], and the OS-ELM-TV [35] as well as the fast tunable RBF [38], are employed as the benchmarks. The two performance indexes, the mean square error (MSE)

$$\mathrm{MSE}(t) = \frac{1}{t}\sum\nolimits_{i=1}^{t}\big(y(i)-\widehat{y}(i)\big)^2, \tag{36}$$

and the mean absolute error (MAE)

$$\mathrm{MAE}(t) = \frac{1}{t}\sum\nolimits_{i=1}^{t}\big|y(i)-\widehat{y}(i)\big|, \tag{37}$$

are utilized to evaluate the online prediction performance, where $\widehat{y}(i)$ denotes the model prediction for $y(i)$. The computational complexity of an online modeling method is measured by its online average computational time per sample (ACTpS), which is defined as $\mathrm{ACTpS} = \frac{\text{Total time}}{\text{Total number of samples}}$. The experiments are carried out on Matlab 2017a, running on a PC with i7-3770 3.40 GHz processor of 4 cores and 16GB of RAM.

3.1. Online time series prediction

We first consider the prediction of Lorenz chaotic time series. Lorzen chaotic time series [46] is governed by the three differential equations as

$$\begin{cases} \frac{\mathrm{d}\,x(t)}{\mathrm{d}\,t} &= a(y(t)-x(t)), \\ \frac{\mathrm{d}\,y(t)}{\mathrm{d}\,t} &= cx(t)-x(t)z(t)-y(t), \\ \frac{\mathrm{d}\,z(t)}{\mathrm{d}\,t} &= x(t)y(t)-bz(t), \end{cases} \tag{38}$$

where a, b and c are the parameters that control the behaviour of Lorzen system. The fourth-order Runge-Kutta method with a step size of 0.01 is used

to generate the samples, and only Y-dimension samples $\{y(t)\}$ are used for the time-series prediction. The 60-steps ahead prediction is considered, which predicts $y(t)$ with the past samples

$$\boldsymbol{x}(t) = \big[y(t-60)\ y(t-66)\ y(t-72)\ y(t-78)\big]^{\mathrm{T}}. \tag{39}$$

In all the simulations, after removing a large number of initial samples, 4,000 data samples are generated. The first 1,000 samples are employed for initial model training and the last 3,000 samples are used for online prediction and adaptive modeling. Note that our proposed learning approach does not really need a large number of training samples, as it can start online operation with just $L = 1$ local linear model. But the OS-ELM needs a large number of training samples, as the ELM model must contain a large number of hidden nodes.

The OS-ELM-TV employs the polynomial function $p_f(x) = x^2 + x$ as the hidden layer activation function and the 3-order Legendre function as the output weight basis function [35], while the training data are used for weight initialization. For the fast tunable RBF, the training is done by the orthogonal least squares algorithm [16] to construct an initial small RBF model. During online operation, the node replacement threshold and the number of data points for weight adaptation are empirically chosen as 10^{-6} and 5, respectively. The step size and the maximum iterations are empirically set to 0.01 and 5, respectively.

3.1.1. Fixed Parameters a, b and c

First Lorzen time series parameters are fixed to $a = 10$, $b = 8/3$ and $c = 28$. We start by investigating the impact of window size W on adaptive local modeling as well as the influence of W, the number of the latest data samples p and the threshold ε on selective ensemble. The dashed curve in Fig. 1 (a) shows the number of local linear models obtained as the function of W on the training dataset. As expected, small W leads to large number of local models identified, and vice versa. Starting with the set of local models identified in training, Algorithm 1 is also applied to the testing dataset, and the number of

local linear models obtained as the function of W is depicted in Fig. 1 (a) as the solid curve. Obviously, the number of local models increases during online adaptation, as the newly emerging process states in the testing data have been identified. Given the threshold $\varepsilon = 0.01$, Fig. 1 (b) depicts the prediction MSE as the function of W with $p = 5$ and $p = 10$, respectively. In general, small W yields better prediction accuracy at the expense of high ACTpS and vice versa. Taking into account both prediction accuracy and computational complexity, it can be seen from Fig. 1 (b) that $W = 37$ to 39 are appropriate. With $\varepsilon = 0.01$, Fig. 1 (c) shows the impact of the number of latest labeled data samples p on the achievable prediction performance for $W = 37$ and $W = 39$, respectively. It can be seen that the test MSE first decreases as p increases. After reaching the minimum value, the test MSE begins increasing as p increases further. In this case, $p = 7$ to 10 are appropriate in terms of prediction accuracy. Fig. 1 (d) demonstrates how the threshold ε impacts on online prediction and adaptive modeling, in terms of trade off between prediction accuracy and computational complexity, given $p = 10$ and two values of $W = 37$ and 39. It can be seen that when ε is smaller than certain value, only the single best local linear model is selected, which results in the lowest computational complexity but the poorest test MSE. Beyond this certain value, increasing ε improves the test MSE while increasing the number of local linear models selected in ensemble prediction. Fig. 1 (d) indicates that the best prediction MSEs can be achieved with $\varepsilon = 0.010$ for $W = 37$ and $\varepsilon = 1$ for $W = 39$, respectively.

Table 1 compares the online prediction and adaptive modeling performance of the proposed selective ensemble of multiple local model learning with those achieved by the OS-ELM, the OS-ELM-TV, the RAN and the fast tunable RBF. Not surprisingly, the OS-ELM has very poor online prediction accuracy with the highest ACTpS. This agrees with the experimental results of [38]. Observe that the OS-ELM-TV can attain the lowest ACTpS but its online prediction accuracy is the worst. It can also be seen that adding more hidden nodes to the OS-ELM-TV may degrade its online performance. The RAN is significantly better than the OS-ELM, in terms of both achievable MSE and ACTpS. The MSE of the

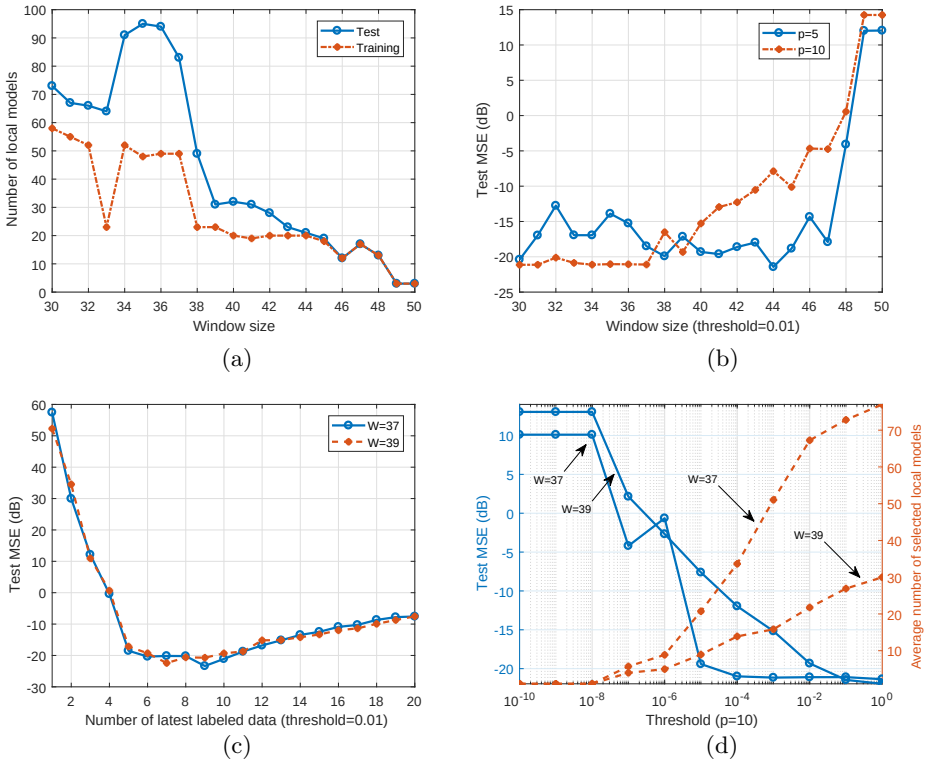


Figure 1: Lorenz time series with fixed parameters: (a) influence of window size W on number of local models, (b) influence of W on prediction accuracy given $\varepsilon = 0.01$ and two values of p, (c) influence of number of latest labeled data p on prediction accuracy given $\varepsilon = 0.01$ and two values of W, and (d) influence of threshold ε on prediction accuracy and average number of local models in selective ensemble given $p = 10$ and two values of W.

Table 1: Lorenz time series with fixed parameters: comparison of online prediction and adaptive modeling performance for the OS-ELM, OS-ELM-TV, RAN, fast tunable RBF and the proposed selective ensemble of multiple local model learning

Model	MSE (dB)	MAE	Online ACTpS (ms)	Models/Nodes	
				Initial	Final
OS-ELM (Sigmoid)	16.7853	4.8137	6.35	500	500
	15.7036	4.1841	41.75	1000	1000
OS-ELM (RBF)	16.9318	4.5486	6.11	500	500
	17.3510	4.7206	35.36	1000	1000
OS-ELM-TV	19.8955	7.7182	0.15	10	10
	20.1744	7.9257	0.42	50	50
RAN	5.0932	0.8375	1.49	0	142
Tunable RBF	-5.2476	0.0557	0.19	10	10
	-20.2228	0.0437	0.37	30	30
Proposed ($W = 39$, $\varepsilon = 1$, $p = 7$)	**-22.4413**	**0.0181**	**0.62**	23	31
Proposed ($W = 37$, $\varepsilon = 0.001$, $p = 9$)	**-23.2790**	**0.0179**	**1.20**	49	83

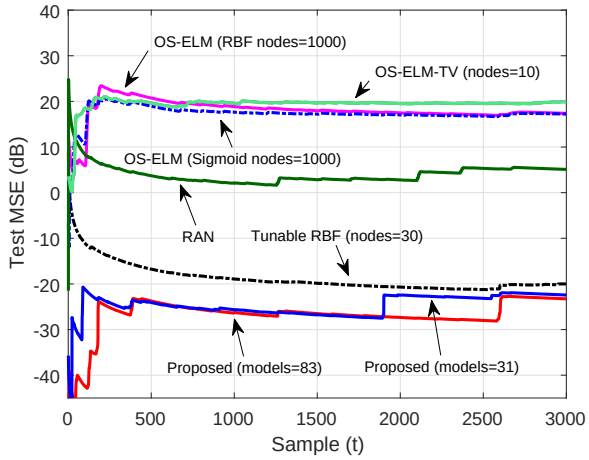


Figure 2: Lorenz time series with fixed parameters: test MSE learning curves for various modeling methods.

tunable RBF is much better than that of the RAN, and it has the second lowest ACTpS. Our proposed approach achieves the best online prediction accuracy and its ACTpS is significantly lower than that of the OS-ELM. Fig. 2 depicts the online MSE learning curves of various models, which again demonstrates the superior prediction accuracy performance of our proposed approach.

3.1.2. Time-Varying Parameters b and c

In this simulation, we set $a = 10$, and let b and c vary with time according to

$$b = \frac{4 + 3(1 + \sin(0.1t))}{3}, \tag{40}$$

$$c = 25 + 3\big(1 + \cos\big(2^{0.001t}\big)\big). \tag{41}$$

Fig. 3 investigates the impacts of the key algorithmic parameters, W, p and ε, on online prediction and adaptive modeling for our proposed method. Similar to the case of fixed parameters, we can draw the same/similar conclusions from Fig. 3 (a) to Fig. 3 (d). Table 2 compares the online prediction and adaptive modeling performance of various modeling methods, while Fig. 4 illustrates their online MSE learning curves. It can be seen again that our proposed method achieves the best online prediction accuracy with relatively low ACTpS, and the

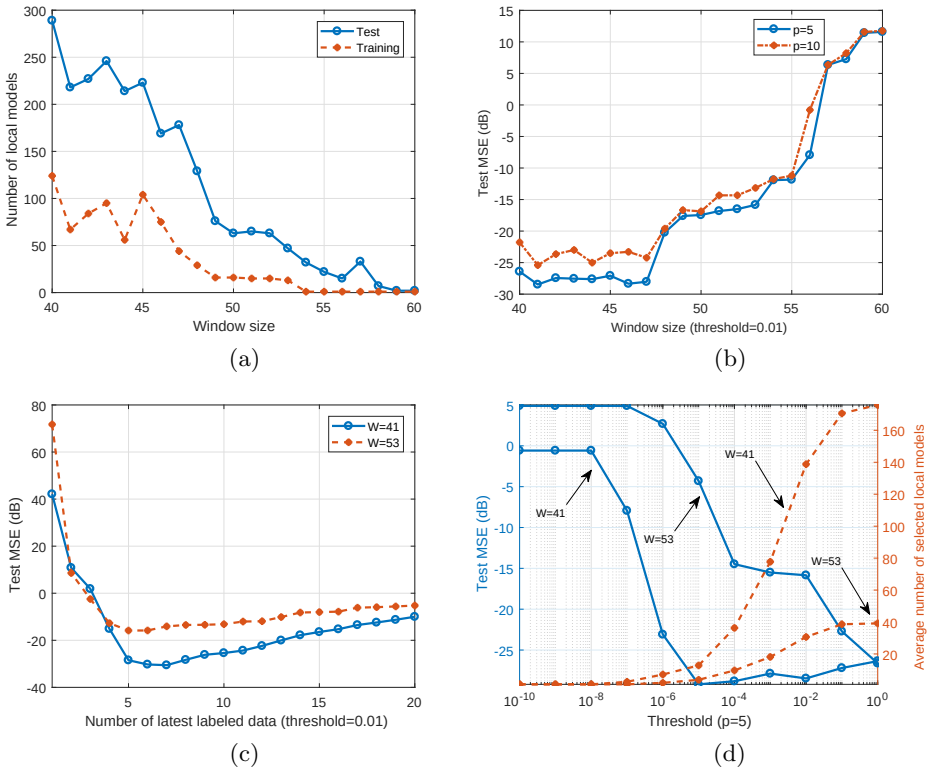


Figure 3: Lorenz time series with time-varying parameters: (a) influence of window size W on number of local models, (b) influence of W on prediction accuracy given $\varepsilon = 0.01$ and two values of p, (c) influence of number of latest labeled data p on prediction accuracy given $\varepsilon = 0.01$ and two values of W, and (d) influence of threshold ε on prediction accuracy and average number of local models in selective ensemble given $p = 5$ and two values of W.

fast tunable RBF has the second best MSE performance with the second lowest ACTpS, while the ELM has the worst performance. In particular, although the OS-ELM-TV can attain the lowest ACTpS, its online prediction accuracy is the worst, which is about 40 dB higher than our proposed approach.

3.1.3. Lorzen Time Series with Time-Based Drift

The parameters of Lorenz system are fixed to $a = 10$, $b = 8/3$ and $c = 28$ but the samples $\{y(t)\}$ are weighted by an exponential time-based drift to obtain the new series $\{\widetilde{y}(t)\}$ according to

$$\widetilde{y}(t)(t) = 1.1^{0.01t} y(t). \tag{42}$$

Table 2: Lorenz time series with time-varying parameters: comparison of online prediction and adaptive modeling performance for the OS-ELM, OS-ELM-TV, RAN, fast tunable RBF and the proposed selective ensemble of multiple local model learning

Model	MSE (dB)	MAE	Online ACTpS (ms)	Models/Nodes	
				Initial	Final
OS-ELM (Sigmoid)	11.0689	2.5390	6.65	500	500
	10.7813	2.4533	42.75	1000	1000
OS-ELM (RBF)	10.8456	2.4077	6.23	500	500
	10.4071	2.3092	35.55	1000	1000
OS-ELM-TV	12.8957	3.2010	0.15	10	10
	17.9352	6.0364	0.42	50	50
RAN	3.5611	0.9198	0.56	0	79
Tunable RBF	-20.9151	0.0440	0.18	10	10
	-22.7813	0.0409	0.38	30	30
Proposed ($W = 53$, $\varepsilon = 1$, $p = 6$)	**-26.6323**	**0.0132**	**0.72**	13	47
Proposed ($W = 41$, $\varepsilon = 0.01$, $p = 4$)	**-28.4732**	**0.0106**	**4.19**	67	218

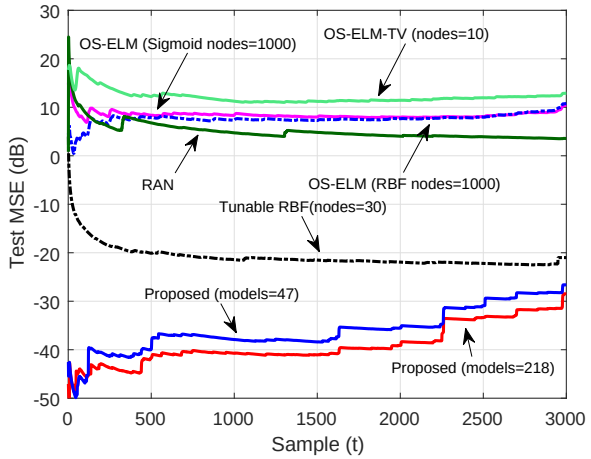


Figure 4: Lorenz time series with time-varying parameters: test MSE learning curves for various modeling methods.

The new time series $\{\widetilde{y}(t)\}$ is used for the time series prediction. In this case, $\{\widetilde{y}(t)\}$ is even more nonstationary than the dataset in the previous simulation with time-varying parameters. In particular, the dynamic range of $\widetilde{y}(t)$ changes from $[-20,\ 20]$ initially to $[-2000,\ 2000]$ in the end.

How the algorithmic parameters, W, p and ε, influence the performance of our selective ensemble of multiple local model learning approach is illustrated in Fig. 5. Furthermore, Table. 3 compares the online prediction and adaptive modeling performance of various modeling methods, while Fig. 6 depicts their online MSE learning curves. Again, the same/similar observations as the pre-

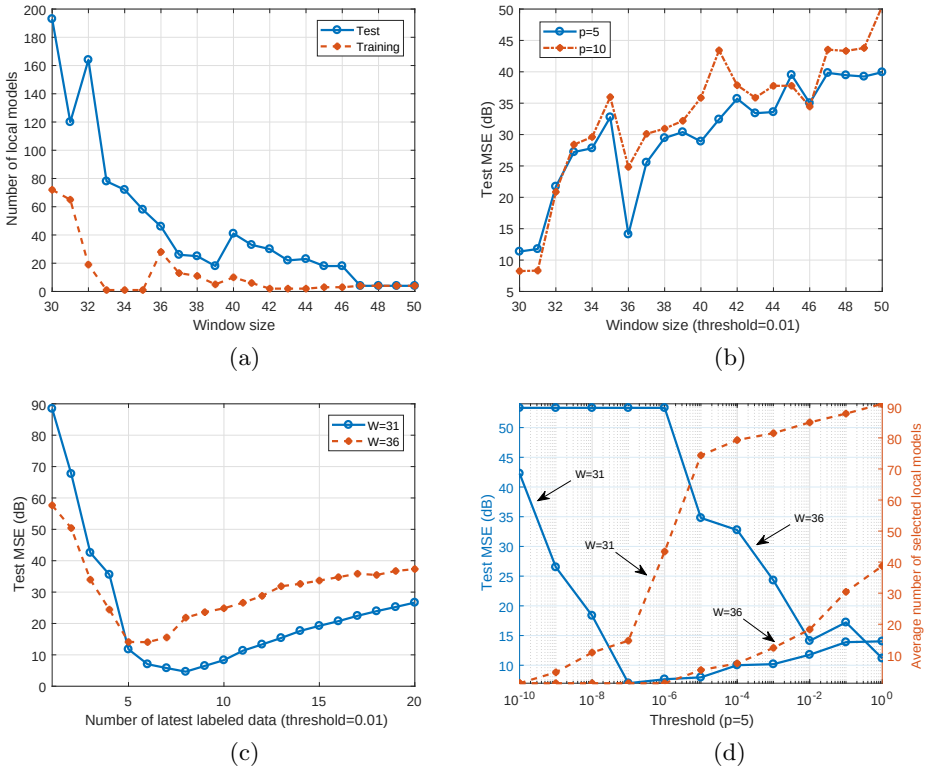


Figure 5: Lorenz time series with time-based drift: (a) influence of window size W on number of local models, (b) influence of W on prediction accuracy given $\varepsilon = 0.01$ and two values of p, (c) influence of number of latest labeled data p on prediction accuracy given $\varepsilon = 0.01$ and two values of W, and (d) influence of threshold ε on prediction accuracy and average number of local models in selective ensemble given $p = 5$ and two values of W.

Table 3: Lorenz time series with time-based drift: comparison of online prediction and adaptive modeling performance for the OS-ELM, OS-ELM-TV, RAN, fast tunable RBF and the proposed selective ensemble of multiple local model learning

Model	MSE (dB)	MAE	ACTpS (ms)	Models/Nodes	
				Initial	Final
OS-ELM (Sigmoid)	52.7165	262.6718	6.15	500	500
	52.7175	262.0425	41.75	1000	1000
OS-ELM (RBF)	52.7802	250.2566	5.68	500	500
	52.7402	248.4557	32.55	1000	1000
OS-ELM-TV	54.1309	268.0022	0.15	10	10
	73.1339	1993.3448	0.41	50	50
RAN	48.3101	29.8495	0.45	0	155
Tunable RBF	36.9557	45.2908	0.18	10	10
	36.6295	42.6192	0.28	30	30
Proposed ($W = 36$, $\varepsilon = 0.01$, $p = 5$)	**14.1114**	**1.3605**	**0.34**	28	46
Proposed ($W = 31$, $\varepsilon = 10^{-7}$, $p = 5$)	**6.9794**	**0.7076**	**0.51**	65	120

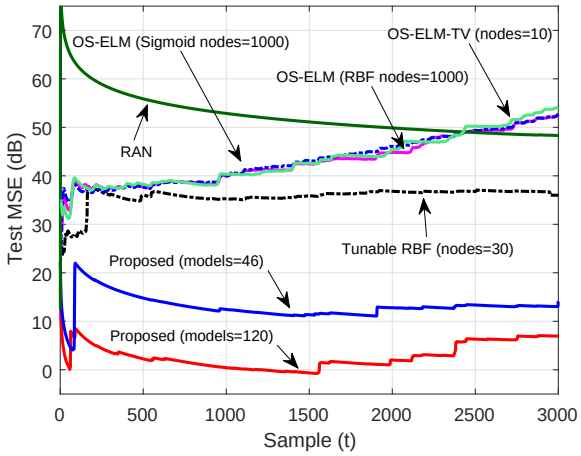


Figure 6: Lorenz series with time-based drift: test MSE learning curves for various modeling methods.

vious two cases can be observed. In particular, it can be seen from Fig. 6 that our proposed method is the most effective in tracking this highly nonstationary and nonlinear Lorenz time series.

3.2. Real-world industrial microwave heating system

Microwave heating technology has found wide-ranging applications in industry due to its many advantages over conventional heating methods, which include selective and volumetric heating, rapid heat transfer and pollution-free environment [11]. However, a major drawback associated with microwave heating is the temperature runaway, caused by properties of material and the inner electromagnetic field distribution [10], which may lead to unwanted combustion and destruction in industrial processes. To improve the safety and efficiency of microwave heating technology in industrial applications, an accurate model is required for the purpose of temperature prediction and control [47, 48, 49, 57, 50]. This is a challenging task, because MHP is a complex thermal process with nonlinear dynamics and nonstationary characteristics. Unlike conventional heat transfer and heat radiation, microwave heating not only involves thermal dynamic variation but also coupled with conversion of microwave energy [58]. Temperature of heated material is a crucial measurement during MHP, as thermal runaway often occurs due to the time-varying physicochem-

ical properties of material. With the increase of the material temperature, its dielectric loss increases dramatically, which conversely poses a positive feedback to temperature increase [59]. Therefore, accurate online temperature estimation is vital to detect thermal runaway in advance.

3.2.1. Description of System

A real-world industrial microwave heating system [49], as illustrated in Fig. 7, is used in this case study, which consists of five microwave generators and waveguides, temperature measurement sensors and the control system hosted in programmable logic controller (PLC). Microwave generated by each microwave generator is transmitted through the corresponding waveguide, fed into the cavity and absorbed by the heated material. Each microwave generator has a maximum power supply of 3 kW at 2.45 GHz. The material is continuously transported through cavity by the conveyor belt, whose speed can be adjusted by a motor driver. Three fiber optical sensors (FOSs), denoted as FOS1 to FOS3, are placed at three different locations using microwave transparent taps to online record multiple-points of temperature.

During the real-time operation of this MHP, the control center receives the measured temperature values from the FOSs, and sends control commends,

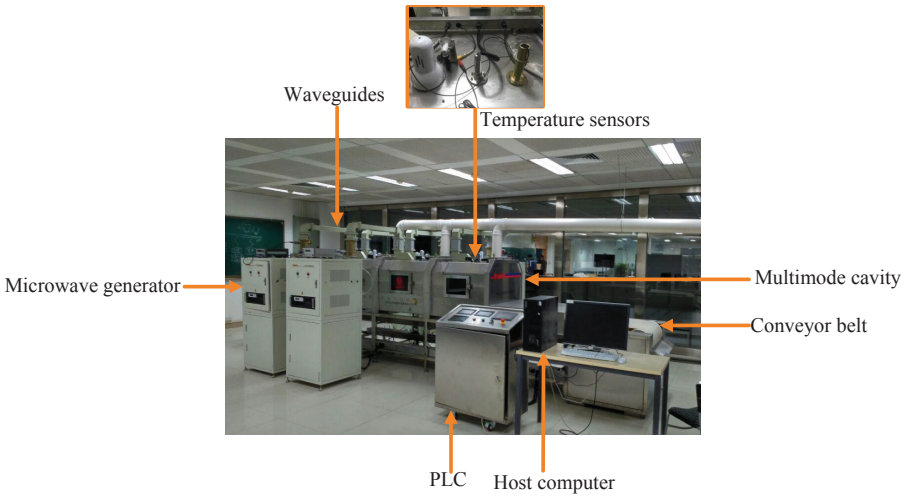


Figure 7: An industrial microwave heating system.

which include the five microwave powers $u_{p_i}(t)$, $1 \leq i \leq 5$, for the five microwave generators as well as the conveyor speed $v(t)$ to the cavity. Thus, the control inputs to this MHP are given by

$$\boldsymbol{u}(t) = \left[u_{p_1}(t)\ u_{p_2}(t)\ u_{p_3}(t)\ u_{p_4}(t)\ u_{p_5}(t)\ v(t)\right]^{\mathrm{T}}. \tag{43}$$

Each FOS measures the temperature, which is the MHP's output $y_{s_j}(t)$ at the FOS's location, where $1 \leq j \leq 3$. Because of near instantaneous response of MHP, the temperature $y_{s_j}(t)$ at the jth FOS's location can be adequately represented by [47, 49, 59]

$$y_{s_j}(t) = f_{\mathrm{nl-ns},j}(\boldsymbol{x}_j(t); t), \tag{44}$$

where $f_{\mathrm{nl-ns},j}(\cdot; t)$ represents the corresponding unknown nonlinear and time-varying system mapping with the input

$$\boldsymbol{x}_j(t) = \left[y_{s_j}(t-1)\ \boldsymbol{u}^{\mathrm{T}}(t-1)\right]^{\mathrm{T}} \in \mathbb{R}^7. \tag{45}$$

From large amount of data collected from this industrial microwave heating system [47, 49], we use three datasets from the three FOSs, and each data set

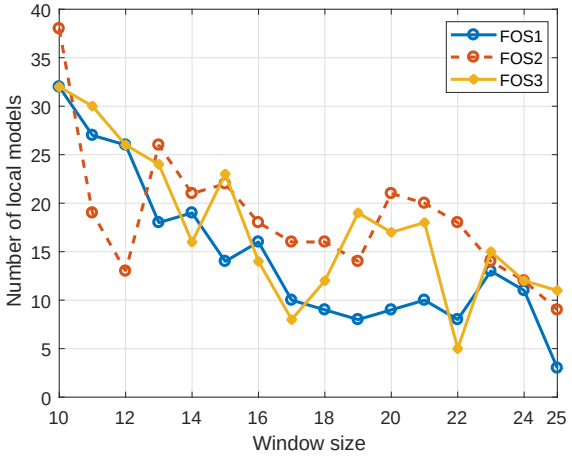


Figure 8: Influence of window size W on number of local models obtained for three training datasets of MHP.

contains 3,000 data samples. We first normalize the five microwave power inputs and the temperature measurements according to

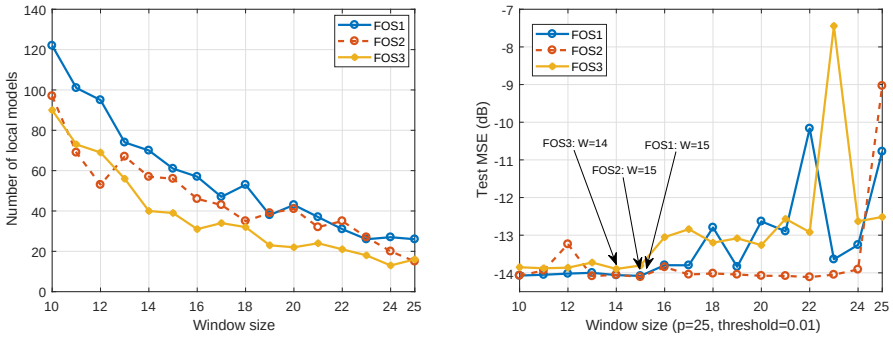
$$\bar{u}_{p_i}(t) = \frac{u_{p_i}(t)}{1000},\ 1 \le i \le 5, \tag{46}$$

$$\bar{y}_{s_j}(t) = \frac{y_{s_j}(t) - y_{\min,s_j}}{y_{\max,s_j} - y_{\min,s_j}},\ 1 \le j \le 3, \tag{47}$$

where $y_{\min,s_j}$ and $y_{\max,s_j}$ are the minimum and maximum temperature measurements of the jth FOS, respectively. For each FOS's dataset, we use the first 1,000 samples for model training, and the last 2,000 samples for online prediction and adaptive modeling.

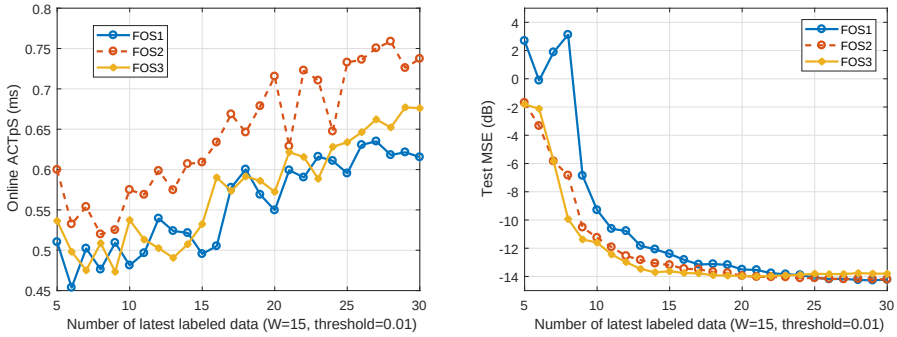
3.2.2. Experimental Results

We investigate the influence of the algorithmic parameters, the window size W, the number of latest data samples p and the threshold ε, on our selective ensemble learning approach. First we apply Algorithm 1 to the training datasets of the three FOSs, and Fig. 8 shows the numbers of local linear models obtained as the functions of W. As expected, small W leads to large number of local models identified and vice versa. With the initial local model sets identified in training, we then apply Algorithm 2 to the three testing datasets. Fig. 9 (a) shows the number of total local linear model identified as the function of W, while Fig. 9 (b) shows the influence of W on online prediction accuracy given $p =$



(a) (b)

Figure 9: Influence of window size W on: (a) number of total local models obtained, and (b) online prediction accuracy given $p = 25$ and $\varepsilon = 0.01$, for three testing datasets of MHP.

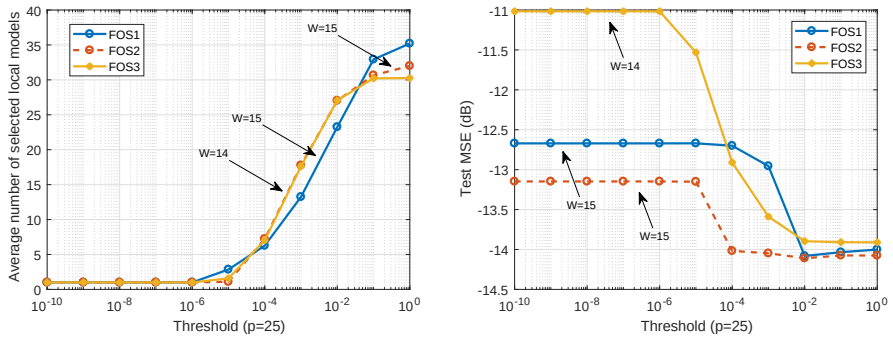


(a) (b)

Figure 10: Influence of number of latest labeled data samples p on: (a) online average computational time per sample, and (b) online prediction accuracy, both obtained given $W = 15$ and $\varepsilon = 0.01$ for three testing datasets of MHP.

25 and $\varepsilon = 0.01$. As expected, small W results in better prediction accuracy but leads to large local model set which has adverse effort on online computational complexity. It can be seen from Fig. 9 (b) that $W = 15$ for FOS1 and FOS2, and $W = 14$ for FOS3 are appropriate.

Next, given $W = 15$ and $\varepsilon = 0.01$, Fig. 10 (a) and (b) show the impacts of the number of of latest labeled data samples p on online computational complexity and prediction accuracy, respectively. Not surprisingly, the online ACTpS increases with p, while the test MSE first decreases rapidly as p increases and it approaches some minimum value for large p. It can be seen from Fig. 10 (b) that the prediction MSEs reach the minimum values when $p \geq 25$ for FOS1 and FOS2 and when $p \geq 20$ for FOS3.



(a) (b)

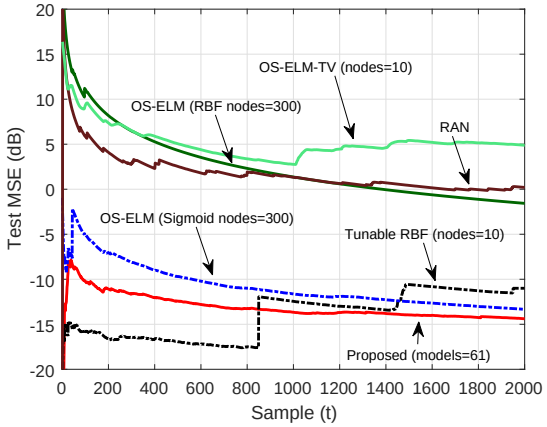
Figure 11: Influence of threshold ε on: (a) average selected ensemble size, and (b) online prediction accuracy, given $p = 25$, $W = 15$ for FOS1 and FOS2 and $W = 14$ for FOS3 of MHP.

Then, given $p = 25$, $W = 15$ for FOS1 and FOS2 as well as $W = 14$ for FOS3, Fig. 11 (a) and (b) illustrate how the threshold ε impacts on online computational complexity, in terms of average selected ensemble size, and prediction accuracy, respectively. Observe from Fig. 11 (a) that when ε is smaller than certain value, only the single best local linear model is selected. When ε is larger than this value, the average size of selected ensemble increases with ε. Also when $\varepsilon = 1$, all the local models are selected and the size of selected ensemble reaches the maximum value. Fig. 11 (b) indicates that the best online prediction MSEs are achieved with $\varepsilon = 0.01$ for all the three testing datasets of this MHP.

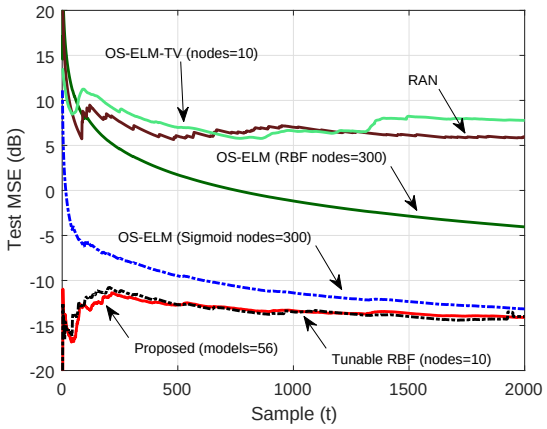
Finally, we compare the online prediction and adaptive modeling performance of the OS-ELM, the OS-ELM-TV, the RAN, the fast tunable RBF and our proposed selective ensemble of multiple local model learning in Table 4. The model construction and the choice of the algorithmic parameters for each method are similar to the previous cases. Furthermore, the online MSE learning curves of these modeling methods are depicted in in Fig. 12 (a) to (c), for the

Table 4: Real-world industrial MHP: comparison of online prediction and adaptive modeling performance for the OS-ELM, OS-ELM-TV, RAN, fast tunable RBF and the proposed selective ensemble of multiple local model learning

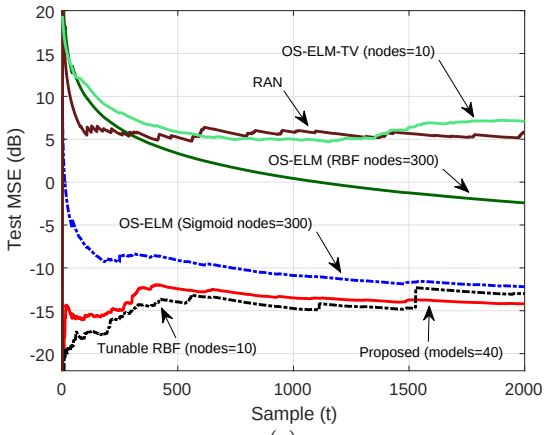
Sensor	Model	MSE (dB)	MAE	Online ACTpS (ms)	Models/Nodes	
					Initial	Final
FOS1	OS-ELM (Sigmoid)	18.7159	0.3275	0.17	100	100
		-13.3432	0.1230	1.26	300	300
	OS-ELM (RBF)	2.1618	0.1993	0.46	100	100
		-1.5694	0.1836	1.85	300	300
	OS-ELM-TV	4.8825	1.2540	0.16	10	10
	RAN	0.2035	0.3238	0.39	0	39
	Tunable RBF	-11.6108	0.1075	0.34	10	10
	Proposed ($W = 15$, $\varepsilon = 0.01$, $p = 25$)	**-14.0782**	**0.1335**	**0.65**	14	61
FOS2	OS-ELM (Sigmoid)	13.1016	0.2450	0.18	100	100
		-13.1594	0.1414	1.33	300	300
	OS-ELM (RBF)	16.2024	0.4114	0.43	100	100
		-4.0463	0.1747	1.89	300	300
	OS-ELM-TV	7.7808	1.8379	0.16	10	10
	RAN	5.9574	0.6522	0.45	0	50
	Tunable RBF	-13.5971	0.1375	0.37	10	10
	Proposed ($W = 15$, $\varepsilon = 0.01$, $p = 25$)	**-14.1107**	**0.1323**	**0.71**	22	56
FOS3	OS-ELM (Sigmoid)	-1.7993	0.1889	0.17	100	100
		-12.1990	0.1690	1.34	300	300
	OS-ELM (RBF)	9.1531	0.2936	0.45	100	100
		-2.4284	0.2110	1.89	300	300
	OS-ELM-TV	7.1171	1.6951	0.16	10	10
	RAN	5.8149	0.6819	0.31	0	37
	Tunable RBF	-13.1200	0.1207	0.34	10	10
	Proposed ($W = 14$, $\varepsilon = 0.01$, $p = 20$)	**-14.2038**	**0.1168**	**0.76**	16	40



(a)



(b)



(c)

Figure 12: Test MSE learning curves for various modeling methods: (a) FOS1, (b) FOS2, and (c) FOS3 of MHP.

three test datasets of the real-world industrial MHP. For this MHP, again the OS-ELM-TV achieves a very poor online prediction accuracy although it imposes the lowest ACTpS. The OS-ELM with 100 hidden nodes attains the worst online prediction accuracy while imposing a very low ACTpS. By contrast, the OS-ELM with 300 hidden nodes significantly improves online prediction accuracy but imposing the highest ACTpS. The RAN considerably outperforms the OS-ELM-TV, in terms of online prediction accuracy, and it also enjoys relatively low ACTpS. In general, the fast tunable RBF achieves the second best test MSE performance at the cost of very low online computational complexity. Our proposed method outperforms all the other models in online prediction accuracy, and it also has a very low online computational complexity. Specifically, for FOS1, our method attains the test MSE of -14.078 dB at the cost of 0.65 ms of ACTpS, while the fast tunable RBF achieves the test MSE of -11.611 dB at the cost of 0.34 ms of ACTpS. For FOS2, our method attains the test MSE of -14.111 dB and its ACTpS is 0.71 ms, while by contrast the fast tunable RBF achieves the test MSE of -13.597 dB and its ACTpS is 0.37 ms. For FOS3, our method achieves the online prediction MSE of -14.204 dB at the expense of 0.76 ms of ACTpS, in comparison to the test MSE of 13.120 dB and the ACTpS of 0.34 ms achieved by the fast tunable RBF. The results of Fig. 12 further demonstrate that our selective ensemble of multiple local model learning approach can much better track the nonlinear and time-varying characteristics of the underlying system.

4. Conclusions and Future Research

In this paper, a novel selective ensemble of multiple local model learning approach has been proposed for adaptive online modeling of nonlinear and nonstationary systems. Our learning approach automatically identifies the newly emerging process state during online operation and fits a local linear model to the newly identified process state. Adaptive modeling is achieved by a selective ensemble strategy which selects a number of best local linear models from the

local model set and optimally combines them to produce the online prediction. Extensive experimental results have demonstrated that our proposed selective ensemble of multiple local model learning is capable of fast tracking the nonlinear and time-varying characteristics of the underlying system. In particular, it has been shown that our proposed method not only achieves the best online prediction accuracy, in comparison with some state-of-the-art online modeling methods, but also offers acceptably low online computational complexity.

Although our approach has been shown to outperform the fast tunable RBF method, the latter offers lower online computational complexity. A key property of our method is its ability to identifying newly emerging characteristics of the underlying system and grows the local linear model set online. Online modeling is carried by an ensemble of small subset local linear models selected from this local model set. For a highly nonlinear and time-varying system, during online operation, the local linear model set is inevitably growing large. In order to reduce online computational complexity, it is desired to remove some 'oldest' local models from the local model set. However, this is not as simple as it appears. A local model exists in the local model set because it has appeared in the system's past history. The fact that it is not used in the most recent selective ensemble does not imply that it will not be needed in the future. Further research is warranted to develop reliable mechanism of removing 'unwanted' past local linear model online. This is expected to be challenging.

Acknowledgement

T. Liu would like to thank the sponsorship of Chinese Scholarship Council for funding his research at School of Electronics and Computer Science, University of Southampton, UK. This work was partly supported by the National Natural Science Foundation of China under grant 61771077 and the Key Research Program of Chongqing Science & Technology Commission under Grant No.CSTC2017jcyjBX0025.

References

[1] L. Rutkowski, "Generalized regression neural networks in time-varying environment," *IEEE Trans. Neural Networks*, vol. 15, no. 3, pp. 576–596, May 2004.

[2] J. Liu and D.-S. Chen, "Nonstationary fault detection and diagnosis for multimode processes," *AIChE J.*, vol. 56, no. 1, pp. 207–219, Jan. 2010.

[3] W. Bartelmus, F. Chaari, R. Zimroz, and M. Haddar, "Modelling of gearbox dynamics under time-varying nonstationary load for distributed fault detection and diagnosis," *European J. Mechanics -A/Solids*, vol. 29, no. 4, pp. 637–646, Jul.-Aug. 2010.

[4] M. Kano and M. Ogawa, "The state of the art in chemical process control in Japan: Good practice and questionnaire survey," *J. Process Control*, vol. 20, no. 9, pp. 969–982, Oct. 2010.

[5] I. B. Khediri, M. Limam, and C. Weihs, "Variable window adaptive kernel principal component analysis for nonlinear nonstationary process monitoring," *Computers & Industrial Eng.*, vol. 61, no. 3, pp. 437-446, Oct. 2011.

[6] Y. Zhang, T. Chai, Z. Li, and C. Yang, "Modeling and monitoring of dynamic processes," *IEEE Trans. Neural Networks and Learning Syst.*, vol. 23, no. 2, pp. 277–284, Feb. 2012.

[7] Y. Zhang, J. An, and H. Zhang, "Monitoring of time-varying processes using kernel independent component analysis," *Chemical Eng. Sci.*, vol. 88, pp. 23–32, Jan. 2013.

[8] X. Yuan, Z. Ge, and Z. Song, "Spatio-temporal adaptive soft sensor for nonlinear time-varying and variable drifting processes based on moving window LWPLS and time difference model," *Asia-Pacific J. Chem. Eng.*, vol. 11, no. 2, pp. 209–219, 2015.

[9] H. Jin, *et al.*, "Dual learning-based online ensemble regression approach for adaptive soft sensor modeling of nonlinear time-varying processes," *Chemometrics Intell. Lab. Syst.*, vol. 151, pp. 228–244, Feb. 2016.

[10] C. A. Vriezinga, S. Sánchez-Pedreño, and J. Grasman, "Thermal runaway in microwave heating: A mathematical analysis," *Applied Mathematical Modelling*, vol. 26, no. 11, pp. 1029–1038, Nov. 2002.

[11] S. Chandrasekaran, S. Ramanathan, and T. Basak, "Microwave material processing - a review," *AIChE J.*, vol. 58, no. 2, pp. 330–363, Feb. 2012.

[12] G. Ditzler, M. Roveri, G. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Computational Intelligence Mag.*, vol. 10, no. 4, pp. 12–25, Nov. 2015.

[13] J. L. Lobo, *et al.*, "Evolving spiking neural networks for online learning over drifting data streams," *Neural Networks*, vol. 108, pp. 1–19, Dec. 2018.

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

[15] S. Chen, S. A. Billings, and P. M. Grant, "Non-linear systems identification using neural networks," *Int. J. Control*, vol. 51, no. 6, pp. 1191–1214, 1990.

[16] S. Chen, C. F N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 302–309, Mar. 1991.

[17] S. Chen, X. Hong, C. J. Harris, and P. M. Sharkey, "Sparse modelling using orthogonal forward regression with PRESS statistic and regularization," *IEEE Trans. Syst., Man and Cybernetics, Part B*, vol. 34, no. 2, pp. 898–911, Apr. 2004.

[18] S. Chen, X. X. Wang, and C. J. Harris, "NARX-based nonlinear system identification using orthogonal least squares basis hunting," *IEEE Trans. Control Systems Technology*, vol. 16, no. 1, pp. 78–84, Jan. 2008.

[19] X. Hong, *et al.*, "Model selection approaches for nonlinear system identification: A review," *Int. J. Syst. Sci.*, vol. 39, no. 10, pp. 925–946, 2008.

[20] S. Chen, X. Hong, B. L. Luk, and C. J. Harris, "Non-linear system identification using particle swarm optimisation tuned radial basis function models," *Int. J. Bio-Inspired Computation*, vol. 1, no. 4, pp. 246–258, 2009.

[21] S. Chen, X. Hong, and C. J. Harris, "Particle swarm optimization aided orthogonal forward regression for unified data modelling," *IEEE Trans. Evolutionary Computation*, vol. 14, no. 4, pp. 477–499, Aug. 2010

[22] S. Chen, X. Hong, J. B. Gao, and C. J. Harris, "Complex-valued B-spline neural networks for modeling and inverting Hammerstein systems," *IEEE Trans. Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1673–1685, Nov. 2014.

[23] X. Hong and S. Chen, "Elastic net orthogonal forward regression," *Neurocomputing*, vol. 148, pp. 551–560, Jan. 2015.

[24] X. Hong, S. Chen, J. Gao, and C. J. Harris, "Nonlinear identification using orthogonal forward regression with nested optimal regularization," *IEEE Trans. Cybernetics*, vol. 45, no. 12, pp. 2925–2936, Dec. 2015.

[25] X. Hong, S. Chen, Y. Guo, and J. Gao, "l^1-norm penalized orthogonal forward regression," *Int. J. Systems Science*, vol. 48, no. 10, pp. 2195–2201, Apr. 2017.

[26] S. Chen, *et al.*, "Comparative performance of complex-valued B-spline and polynomial models applied to iterative frequency-domain decision feedback equalization of Hammerstein channels," *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 2872–2884, Dec. 2017.

[27] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural computation*, vol. 1, no. 2, pp. 281–294, 1989.

[28] S. Chen and S. A. Billings, "Recursive prediction error estimator for non-linear models," *Int. J. Control*, vol. 49, no. 2, pp. 569–594, 1989.

[29] S. Chen, "Nonlinear time series modelling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning," *Electronics Letters*, vol. 31, no. 2, pp. 117–118, Jan. 1995.

[30] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1-3, pp. 489–501, Dec. 2006.

[31] N. Liang, G. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Trans. Neural Networks*, vol. 17, no. 6, pp. 1411–1423, Nov 2006.

[32] G.-B. Huang and L. Chen, "Enhanced random search based incremental extreme learning machine," *Neurocomputing*, vol. 71, no. 16, pp. 3460–3468, Oct 2008.

[33] F. Ding, P. X. Liu, and G. Liu, "Multiinnovation least-squares identification for system modeling," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 40, no. 3, pp. 767–778, Jun. 2010.

[34] X. Hong, G. D. Fatta, H. Chen, and S. Wang, "Sparse least squares support vector regression for nonstationary systems," in *Proc. IJCNN 2018* (Rio de Janeiro, Brazil), Jul. 8-13, 2018, pp. 1–8.

[35] Y. Ye, S. Squartini, and F. Piazza, "Online sequential extreme learning machine in nonstationary environments," *Neurocomputing*, vol. 116, pp. 94–101, Sep. 2013.

[36] J. Platt, "A resource-allocating network for function interpolation," *Neural computation*, vol. 3, no. 2, pp. 213–225, Jun. 1991.

[37] V. Kadirkamanathan and M. Niranjan. "A function estimation approach to sequential learning with neural networks," *Neural computation*, vol. 5, no. 6, pp. 954–975, Nov. 1993.

[38] H. Chen, Y. Gong, X. Hong, and S. Chen, "A fast adaptive tunable RBF network for nonstationary systems," *IEEE Trans. Cybernetics*, vol. 46, no. 12, pp. 2683–2692, Dec. 2016.

[39] H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *J. Royal Statistical Society*, vol. 42, no. 3, pp. 245–292, 1980.

[40] H. Tong, *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag: New York, 1983.

[41] S. A. Billings and S. Chen, "Extended model set, global data and threshold model identification of severely non-linear systems," *Int. J. Control*, vol. 50, no. 5, pp. 1897–1923, 1989.

[42] X. Hong and Y. Gong, "A constrained recursive least squares algorithm for adaptive combination of multiple models," in *Proc. IJCNN 2015* (Killarney, Ireland), Jul. 11-16, 2015, pp. 1–6.

[43] H. Chen, Y. Gong, and X. Hong, "A new adaptive multiple modelling approach for non-linear and non-stationary systems," *Int. J. Systems Science*, vol. 47, no. 9, pp. 2100–2110, 2016.

[44] W. Shao, *et al.*, "Online soft sensor design using local partial least squares models with adaptive process state partition," *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 108–121, May 2015.

[45] W. Shao, S. Chen, and C. J. Harris, "Adaptive soft sensor development for multi-output industrial processes based on selective ensemble learning," *IEEE Access*, vol. 6, pp. 55628–55642, Oct. 2018.

[46] E. N. Lorenz. "Deterministic nonperiodic flow," *J. Atmospheric Sci.*, vol. 20, pp. 130–141, Mar. 1963.

[47] K. Wang, *et al.*, "Learning to detect local overheating of the high-power microwave heating process with deep learning," *IEEE Access*, vol. 6, pp. 10288–10296, Feb. 2018.

[48] J. Zhong, S. Liang, and Q. Xiong, "Improved receding horizon H_∞ temperature spectrum tracking control for Debye media in microwave heating process," *J. Process Control*, vol. 71, pp. 14–24, Nov. 2018.

[49] T. Liu, S. Liang, Q. Xiong, and K. Wang, "Adaptive critic based optimal neurocontrol of a distributed microwave heating system using diagonal recurrent network," *IEEE Access*, vol. 6, pp. 68839–68840, Dec. 2018.

[50] T. Liu, S. Liang, Q. Xiong, and K. Wang, "Two-stage method for diagonal recurrent neural network identification of a high-power continuous microwave heating system," *Neural Processing Letter*, pp. 1–22, Feb. 2019.

[51] D. H. Kaye and D. A. Freedman, *Reference Guide on Statistics* (3rd ed.). Washington DC: National Academy Press, 2011.

[52] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, Feb. 1991.

[53] X. Hong and C. J. Harris, "A mixture of experts network structure construction algorithm for modelling and control," *Applied Intelligence*, vol. 16, no. 1. pp. 59–69, Jan. 2002.

[54] V. Cherkassky and Y. Ma, "Multiple model regression estimation," *IEEE Trans. Neural Networks*, vol. 16, no.4, pp. 785–798, Jul. 2005.

[55] Z.-H. Zhou, J. Wu, and W. Tang, "Ensemble neural networks: Many could be better than all," *Artificial Intelligence*, vol. 137, nos. 1-2, pp. 239–263, May 2002.

[56] L.-J. Zhao, T.-Y. Chai, and D.-C. Yuan, "Selective ensemble extreme learning machine modeling of effluent quality in wastewater treatment plants," *Int. J. Automation and Computing*, vol. 9, no. 6, pp. 627–633, Dec. 2012.

[57] J. Zhong, L. Shan, and Q. Xiong, "Receding horizon H^∞ guaranteed cost tracking control for microwave heating medium with temperature-

dependent permittivity," *ISA Transactions*, vol. 73, no. 2018, pp. 249–256, Feb. 2018.

[58] J. Zhong, S. Liang, Y. Yuan, and Q. Xiong, "Coupled electromagnetic and heat transfer ODE model for microwave heating with temperature-dependent permittivity," *IEEE Trans. Microwave Theory & Techniques*, vol. 64, no. 8, pp. 2467–2477, Aug. 2016.

[59] X. Shi, *et al.*, "Research of uniformity evaluation model based on entropy clustering in the microwave heating processes," *Neurocomputing*, vol. 173, no. 3, pp. 562–572, Jan. 2016.