



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

ARTICLE INFO

Article history:

Received 9 April 2019

Revised 11 October 2019

Accepted 11 October 2019

Available online 17 October 2019

Communicated by Prof. Zidong Wang

Keywords:

Nonlinear and time-varying system

Online modeling and prediction

Local model learning

Selective ensemble

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 on-line 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.

© 2019 Published by Elsevier B.V.

1. Introduction

Real-world systems often exhibit both nonlinear and nonstationary characteristics [1–13]. For these time-varying nonlinear systems, batch global nonlinear modeling approaches [14–26] 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 [27–34]. 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 [30–32] 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 processes 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

* Corresponding author at: School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.

E-mail addresses: liutong42@cqu.edu.cn (T. Liu), sqc@ecs.soton.ac.uk (S. Chen), lightsun@cqu.edu.cn (S. Liang), chrisharris57@msn.com (C.J. Harris).

[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 [39–43]. 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) [47–49]. 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 $\{\mathbf{x}(t), y(t)\}_{t=1}^N$ drawn from a process, where $\mathbf{x}(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}$ are the system’s input vector and output, respectively. Assume that the nonlinear characteristics

of the system over $\{\mathbf{x}(t), y(t)\}_{t=1}^N$ 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 $\{\mathbf{X}_l, \mathbf{y}_l\}_{l=1}^L$, where each \mathbf{X}_l contains W consecutive time samples of the input $\mathbf{x}(t)$ and \mathbf{y}_l consists of the corresponding output samples.

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

$$\hat{\mathbf{y}}_{\text{ini}} = f_{\text{ini}}(\mathbf{X}_{\text{ini}}) = \Phi \beta \quad (1)$$

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

$$\beta = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}_{\text{ini}}. \quad (2)$$

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

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

After an initial local model f_{ini} is built, a shifted window $\mathcal{W}_{\text{sft}} = \{\mathbf{X}_{\text{sft}}, \mathbf{y}_{\text{sft}}\}$ is sequentially obtained by moving the window one step ahead, that is, \mathcal{W}_{sft} contains the samples $\{\mathbf{x}(t), y(t)\}_{t=t_{\text{ini}}+1}^{t_{\text{ini}}+1+W}$. If the two local regions \mathcal{W}_{ini} and \mathcal{W}_{sft} are not significantly different, it can be considered that the process data within \mathcal{W}_{sft} follow the same distribution as in \mathcal{W}_{ini} and the window is continued to be shifted forward. Otherwise, \mathcal{W}_{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}_{sft} . Determining whether \mathcal{W}_{ini} and \mathcal{W}_{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}_{sft} be defined as

$$\mathbf{e}_{\text{sft}} = \mathbf{y}_{\text{sft}} - f_{\text{ini}}(\mathbf{X}_{\text{sft}}). \quad (4)$$

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

$$H_0^\mu : \mu_{\text{ini}} = \mu_{\text{sft}}, \quad (5)$$

$$H_0^{\sigma^2} : \sigma_{\text{sft}}^2 = \sigma_{\text{ini}}^2. \quad (6)$$

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

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

$$\chi^2 = (W-1) \sigma_{\text{sft}}^2 / \sigma_{\text{ini}}^2. \quad (8)$$

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

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

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 $f_{ini} = f_L$. When one or both conditions of (9) are violated, \mathcal{W}_{ini} and \mathcal{W}_{sft} are significantly different, and the new local linear model $f_{new} = f_{sft}$ 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 $\{\mathbf{X}_{sft}, \mathbf{y}_{sft}\}$ based on f_{new} and f_l be defined, respectively by

$$\mathbf{e}_{new} = \mathbf{y}_{sft} - f_{new}(\mathbf{X}_{sft}), \quad (10)$$

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

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

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

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

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

$$H_l^\mu: \mu_l = \mu_{new}, \quad (14)$$

$$H_l^{\sigma^2}: \sigma_l^2 = \sigma_{new}^2, \quad (15)$$

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 $l \in \{1, 2, \dots, L - 1\}$ such that

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

the hypotheses (14) and (15) are both valid, and \mathbf{e}_{new} and \mathbf{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 $f_L = f_{new}$.

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 $\{\mathbf{x}(t), y(t)\}$ is available, the data window \mathcal{W}_{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 effects. 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

Algorithm 1 Adaptation of local linear model set.

- 1: **Initialization**
- 2: Collect \mathcal{W}_{ini} with W consecutive samples from historical data, and construct the LS linear model f_{ini} on \mathcal{W}_{ini} .
- 3: Calculate \mathbf{e}_{ini} , and estimate μ_{ini} and σ_{ini}^2 .
- 4: Set $L = 1$, $\{\mathcal{W}_L, f_L\} = \{\mathcal{W}_{ini}, f_{ini}\}$ and $\mathcal{W}_{sft} = \mathcal{W}_L$.
- 5: **Step 1: New local model detection**
- 6: When a new data sample is available, shift \mathcal{W}_{sft} one sample ahead.
- 7: Calculate \mathbf{e}_{sft} , and estimate μ_{sft} and σ_{sft}^2 .
- 8: Construct T and χ^2 statistics using (7) and (8).
- 9: **If** both conditions of (9) are satisfied
- 10: Go to **Step 1**.
- 11: **End if**
- 12: Construct the LS linear model f_{sft} on \mathcal{W}_{sft} .
- 13: Set $\mathcal{W}_{new} = \mathcal{W}_{sft}$ and $f_{new} = f_{sft}$.
- 14: Calculate \mathbf{e}_{new} , and estimate μ_{new} and σ_{new}^2 .
- 15: **Step 2: Redundant local model deletion**
- 16: **For** $l = 1, 2, \dots, L - 1$
- 17: Compute \mathbf{e}_l , and estimate μ_l and σ_l^2 .
- 18: Construct T_l and χ_l^2 statistics using (12) and (13).
- 19: **If** both conditions of (16) are satisfied
- 20: Delete f_l , set $f_i = f_{i+1}$ for $i = l, l + 1, \dots, L - 1$, set $L = L - 1$, then go to **Step 3**.
- 21: **End if**
- 22: **End for**
- 23: **Step 3: Add new local model**
- 24: Set $L = L + 1$, $\mathcal{W}_L = \mathcal{W}_{new}$ and $f_L = f_{new}$.
- 25: Return to **Step 1**.

$t_{next} = t + 1$, the task of online modeling update is to produce the model prediction $\hat{y}(t_{next})$ for the processes true output $y(t_{next})$, given the process input $\mathbf{x}(t_{next})$ 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 [52–54]. 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, 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 $\{\mathbf{x}(t - i), y(t - i)\}_{i=0}^{p-1}$.

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

$$e_l(t - i) = y(t - i) - f_l(\mathbf{x}(t - i)), \quad 0 \leq i \leq p - 1. \quad (17)$$

The performance metric of the l th local model is defined as

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

By further defining

$$J_{lmax}(t) = \max_{1 \leq l \leq L} J_l(t), \quad (19)$$

we can normalize the performance metrics of (18) to

$$\bar{J}_l(t) = \frac{J_l(t)}{J_{lmax}(t)}, \quad 1 \leq l \leq L. \quad (20)$$

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

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

should be selected. Moreover, other local models whose performance metrics (20) are below a given threshold $0 < \varepsilon \leq 1$ 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 $\varepsilon \leq J_{l_{\min}}(t)$, only the best local model f_{l_1} is selected.

Assume that $M(\geq 1)$ 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

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

This selection procedure yields the M local model outputs

$$\hat{y}_{l_m}(t-i) = f_{l_m}(\mathbf{x}(t-i)), \quad 1 \leq m \leq M, \quad (23)$$

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

$$\hat{y}(t-i) = \sum_{m=1}^M \theta_m(t) \hat{y}_{l_m}(t-i), \quad 0 \leq i \leq p-1, \quad (24)$$

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

$$\sum_{m=1}^M \theta_m(t) = 1. \quad (25)$$

The estimation errors

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

are utilized to determine the combining coefficients.

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

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

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

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

where $\boldsymbol{\theta}(t) = [\theta_1(t) \dots \theta_M(t)]^T$ and $\bar{\mathbf{E}}(t)$ is the estimated error covariance matrix which is given as

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

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

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \frac{1}{2} \boldsymbol{\theta}^T(t) \bar{\mathbf{E}}(t) \boldsymbol{\theta}(t), \\ \text{s.t.} \quad & \sum_{m=1}^M \theta_m(t) = 1. \end{aligned} \quad (30)$$

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

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

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

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

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

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

which is followed by the normalization

$$\hat{\theta}_m(t) = \frac{1}{\sum_{j=1}^M \tilde{\theta}_j(t)} \tilde{\theta}_m(t), \quad 1 \leq m \leq M. \quad (34)$$

The prediction $\hat{y}(t_{\text{next}})$ for the processes true output $y(t_{\text{next}})$ is produced as the selected ensemble

$$\hat{y}(t_{\text{next}}) = \sum_{m=1}^M \hat{\theta}_m(t) f_{l_m}(\mathbf{x}(t_{\text{next}})) \quad (35)$$

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 $\mathbf{x}(t_{\text{next}})$ at new sample time $t_{\text{next}} = t + 1$.
 - 6: Calculate the performance metrics $\bar{J}_l(t)$ using (20) for $1 \leq l \leq L$ on past p data points.
 - 7: Select the subset models with the index set Γ of (22).
 - 8: Calculate the error covariance matrix $\bar{\mathbf{E}}(t)$ using (29).
 - 9: Calculate the optimal combining coefficients $\hat{\boldsymbol{\theta}}(t)$ 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 $\{\mathbf{x}(t_{\text{next}}), y(t_{\text{next}})\}$ to the dataset with $t = t + 1$.
 - 14: Shift \mathcal{W}_{sft} one sample ahead, and perform relevant 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)) [30–32], 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)

$$\text{MSE}(t) = \frac{1}{t} \sum_{i=1}^t (y(i) - \hat{y}(i))^2, \quad (36)$$

and the mean absolute error (MAE)

$$\text{MAE}(t) = \frac{1}{t} \sum_{i=1}^t |y(i) - \hat{y}(i)|, \quad (37)$$

are utilized to evaluate the online prediction performance, where $\hat{y}(i)$ 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 $\text{ACTpS} = \frac{\text{Total time}}{\text{Total number of samples}}$. The experiments are carried out on Matlab 2017a, running on a PC with i7-3770 3.40 GHz processor of 4 cores and 16 GB of RAM.

3.1. Online time series prediction

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

$$\begin{cases} \frac{dx(t)}{dt} = a(y(t) - x(t)), \\ \frac{dy(t)}{dt} = cx(t) - x(t)z(t) - y(t), \\ \frac{dz(t)}{dt} = x(t)y(t) - bz(t), \end{cases} \quad (38)$$

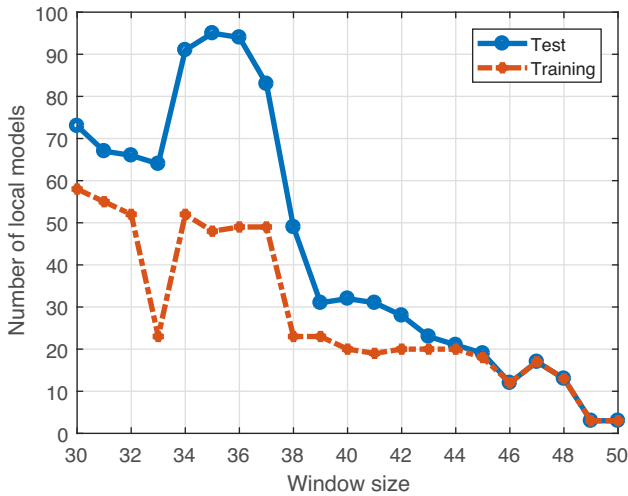
where a , b and c are the parameters that control the behaviour of Lorenz system. The fourth-order Runge-Kutta method with a step size of 0.01 is used to generate the samples, and only Y -dimension samples $\{y(t)\}$ are used for the time-series prediction. The 60-steps

ahead prediction is considered, which predicts $y(t)$ with the past samples

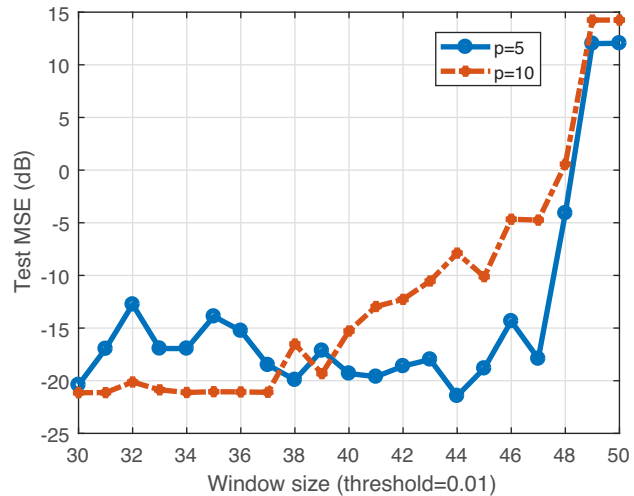
$$\mathbf{x}(t) = [y(t-60) \ y(t-66) \ y(t-72) \ y(t-78)]^T. \quad (39)$$

In all the simulations, after removing a large number of initial samples, 4000 data samples are generated. The first 1000 samples are employed for initial model training and the last 3000 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.

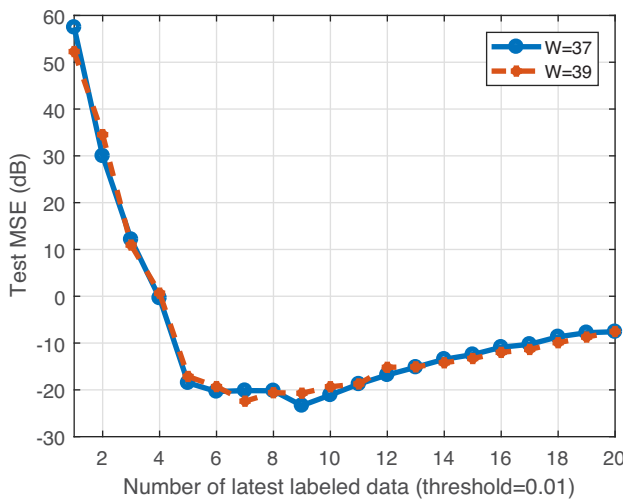
The OS-ELM-TV employs the polynomial function $p_f(x) = x^2 + x$ as the hidden layer activation function and the 3-order Legendre function as the output weight basis function [35], while the training data are used for weight initialization. For the fast tunable RBF, the training is done by the orthogonal 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,



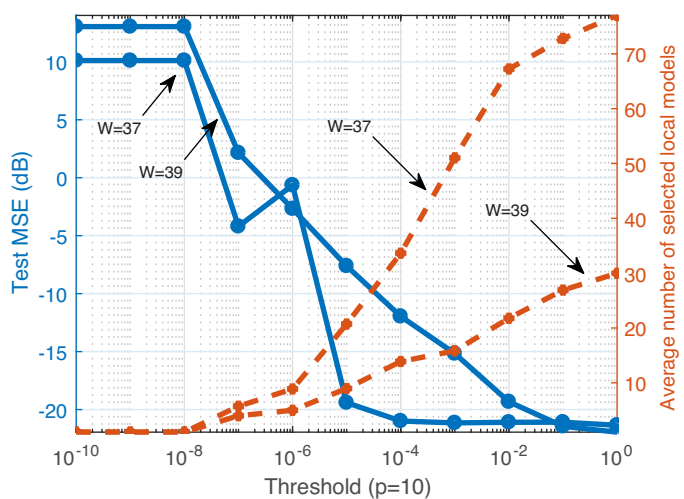
(a)



(b)



(c)



(d)

Fig. 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 |

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 Lorenz 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 hid-

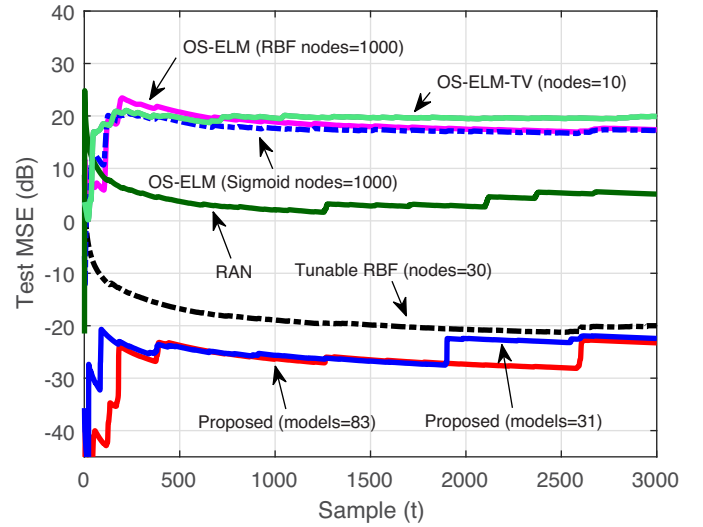


Fig. 2. Lorenz time series with fixed parameters: test MSE learning curves for various modeling methods.

den 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 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

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

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

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)–(d). Table 2 com-

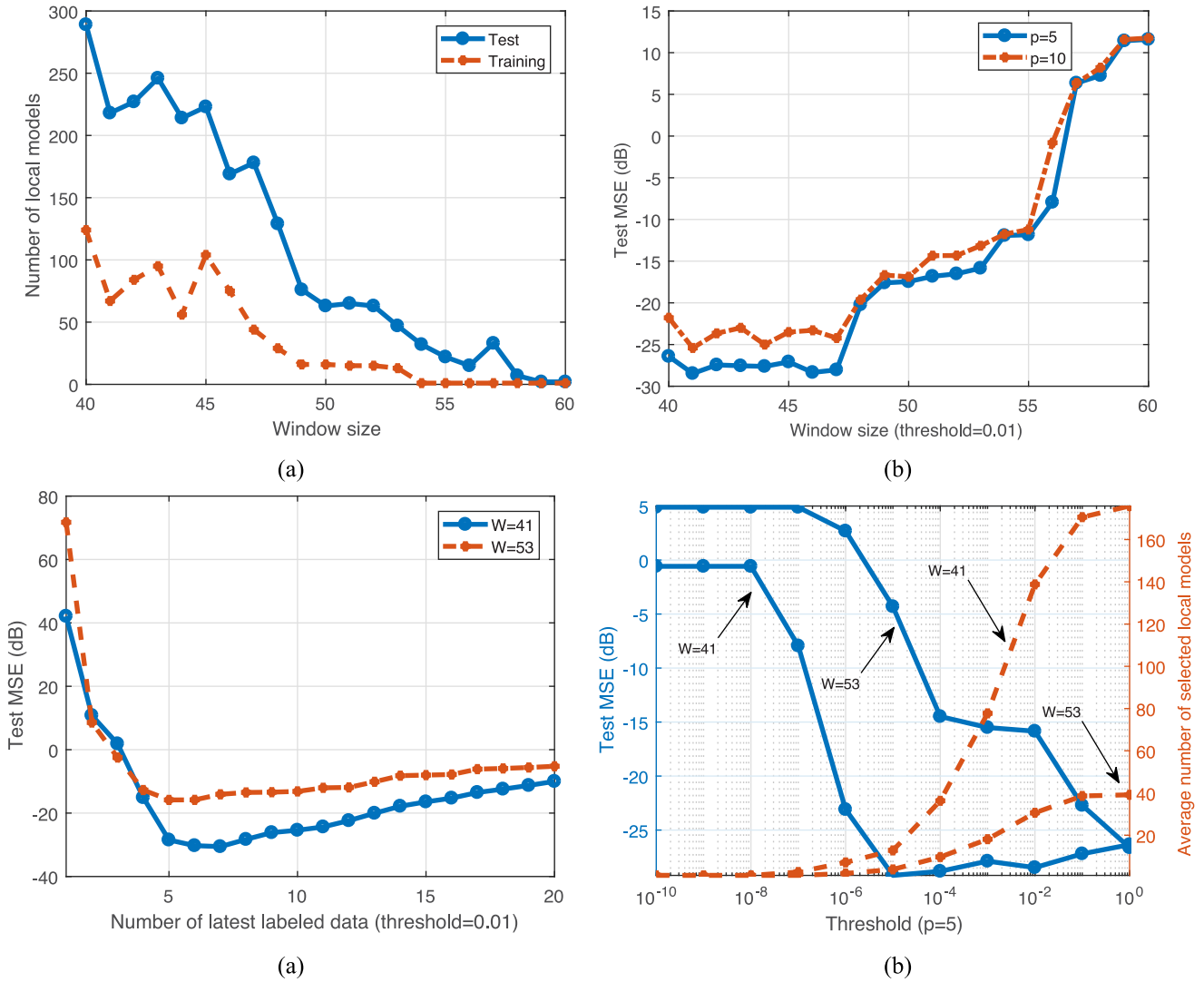


Fig. 3. Lorenz time series with time-varying parameters: (a) influence of window size W on number of local models, (b) influence of W on prediction accuracy given $\varepsilon = 0.01$ and two values of p , (c) influence of number of latest labeled data p on prediction accuracy given $\varepsilon = 0.01$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

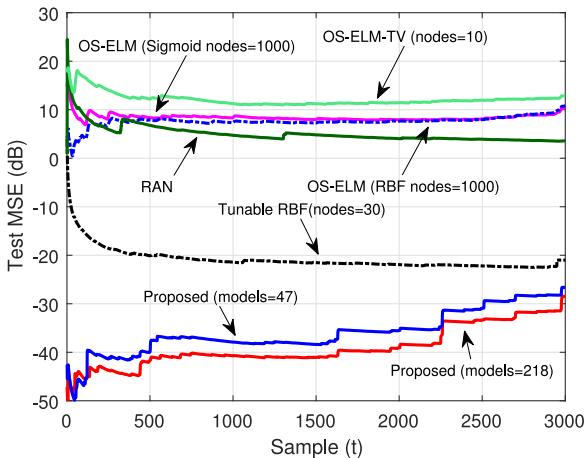


Fig. 4. Lorenz time series with time-varying parameters: test MSE learning curves for various modeling methods.

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 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. Lorenz 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 $\{\tilde{y}(t)\}$ according to

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

The new time series $\{\tilde{y}(t)\}$ is used for the time series prediction. In this case, $\{\tilde{y}(t)\}$ is even more nonstationary than the dataset in the previous simulation with time-varying parameters. In particular, the dynamic range of $\tilde{y}(t)$ 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 com-

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 |

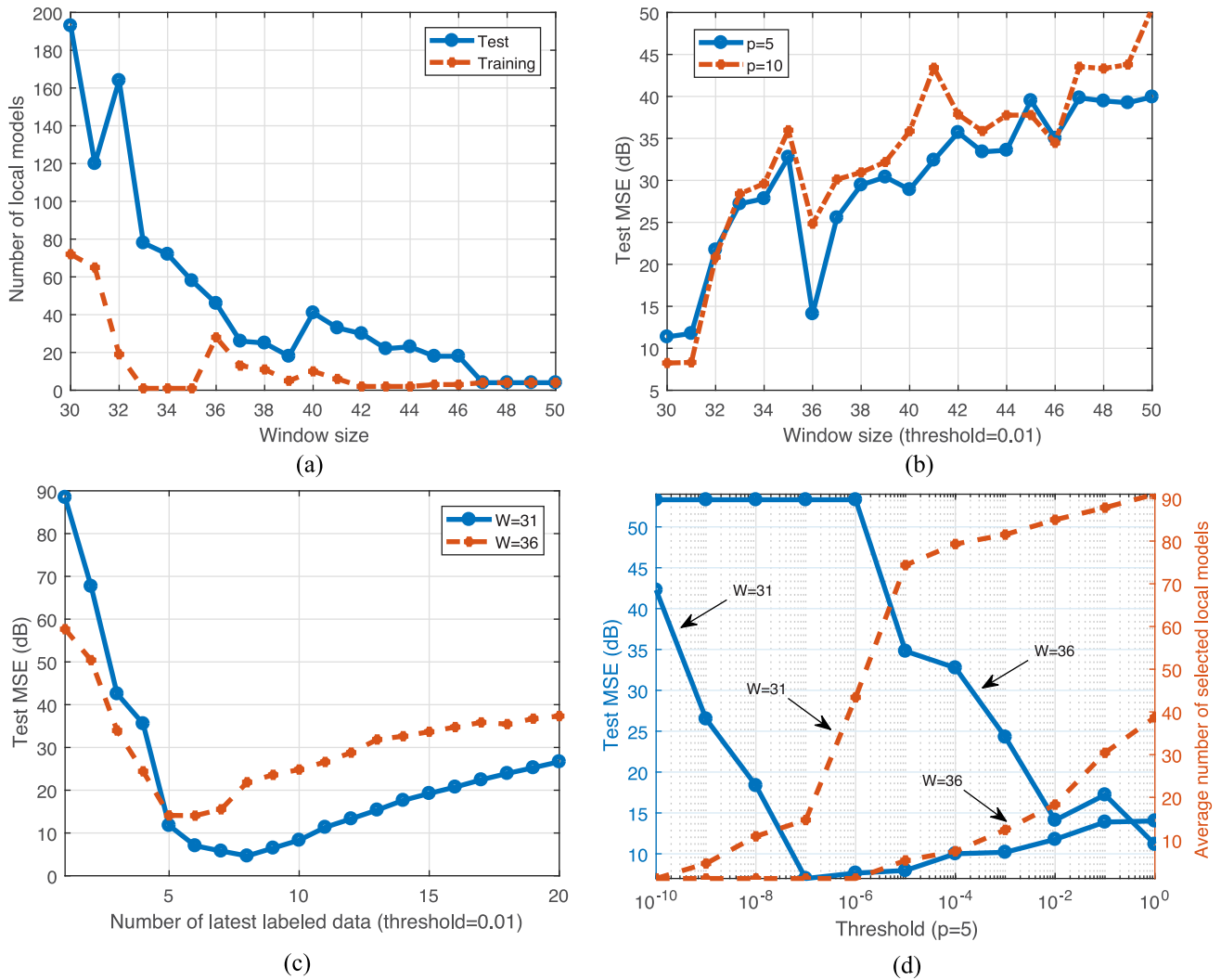


Fig. 5. Lorenz time series with time-based drift: (a) influence of window size W on number of local models, (b) influence of W on prediction accuracy given $\varepsilon = 0.01$ and two values of p , (c) influence of number of latest labeled data p on prediction accuracy given $\varepsilon = 0.01$ and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

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 previous 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

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 |
| | 52.7802 | 250.2566 | 5.68 | 500 | 500 |
| OS-ELM (RBF) | 52.7402 | 248.4557 | 32.55 | 1000 | 1000 |
| | 54.1309 | 268.0022 | 0.15 | 10 | 10 |
| OS-ELM-TV | 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 ($W = 36, \epsilon = 0.01, p = 5$) | 14.1114 | 1.3605 | 0.34 | 28 | 46 |
| Proposed ($W = 31, \epsilon = 10^{-7}, p = 5$) | 6.9794 | 0.7076 | 0.51 | 65 | 120 |

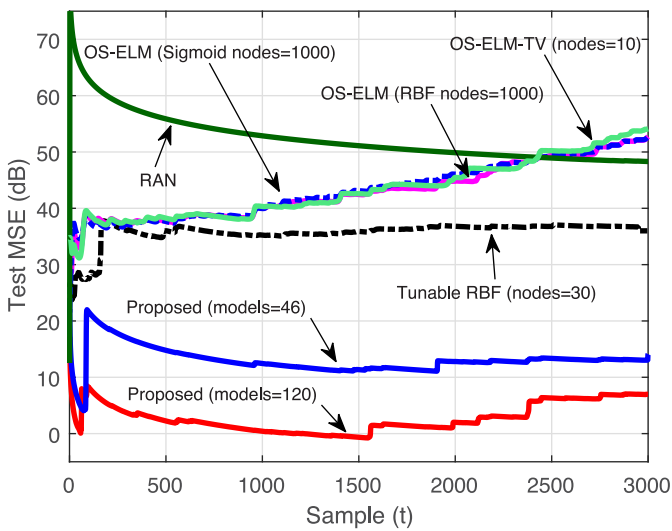


Fig. 6. Lorenz series with time-based drift: test MSE learning curves for various modeling methods.

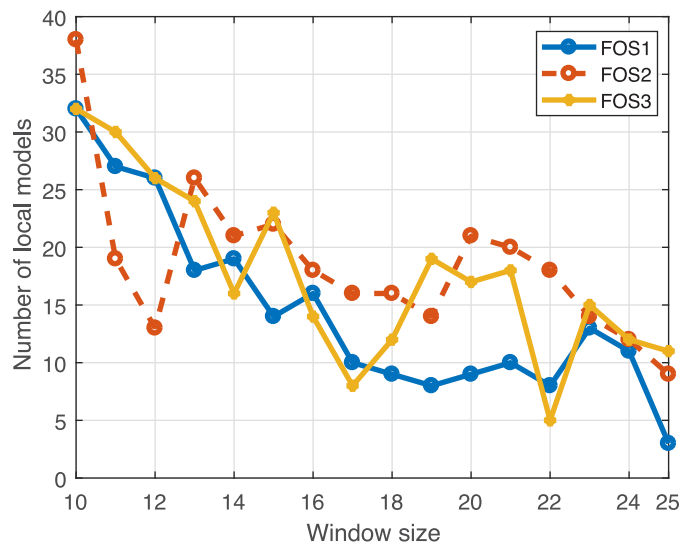


Fig. 8. Influence of window size W on number of local models obtained for three training datasets of MHP.

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 [47–50,57]. This is a challenging task, because MHP is a complex thermal process with nonlinear dynamics and nonstationary characteristics. Unlike conventional

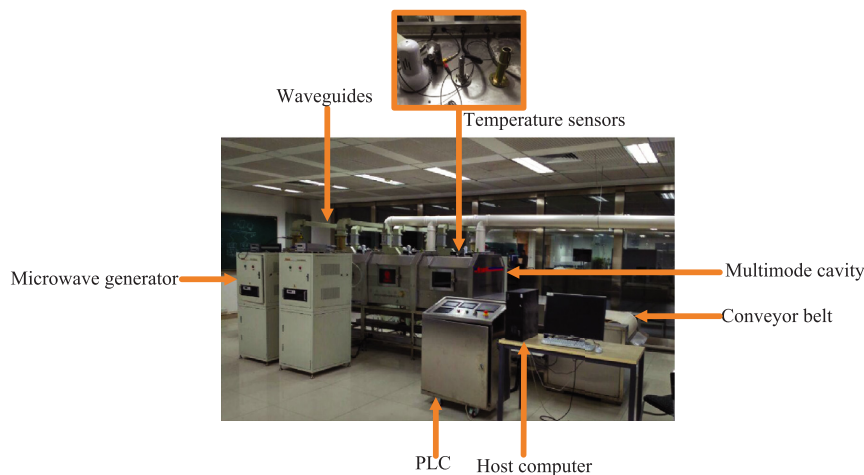


Fig. 7. An industrial microwave heating system.

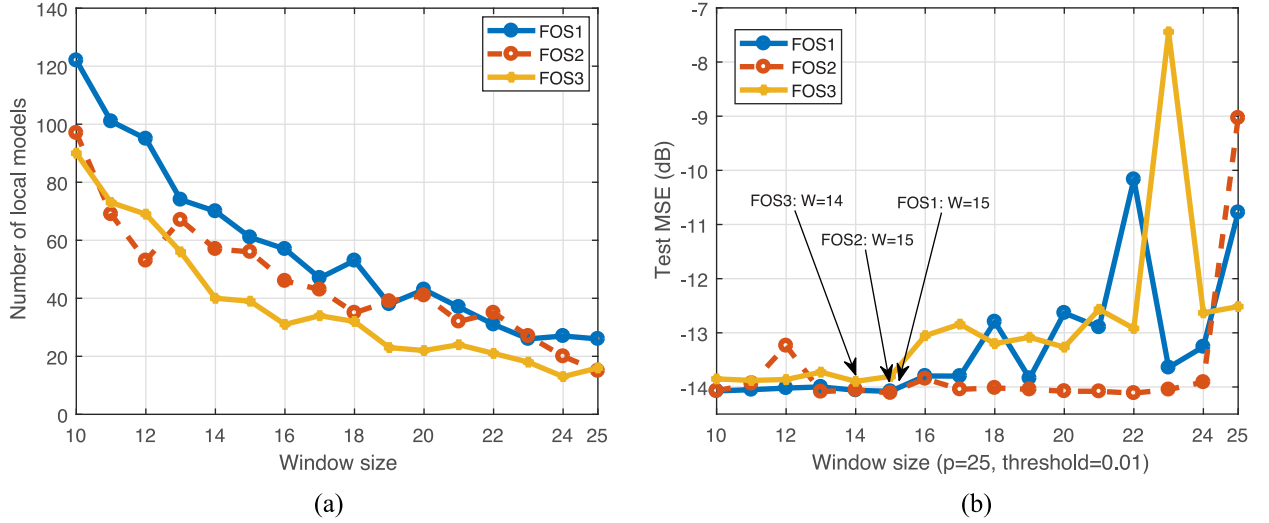


Fig. 9. Influence of window size W on: (a) number of total local models obtained, and (b) online prediction accuracy given $p = 25$ and $\varepsilon = 0.01$, for three testing datasets of MHP.

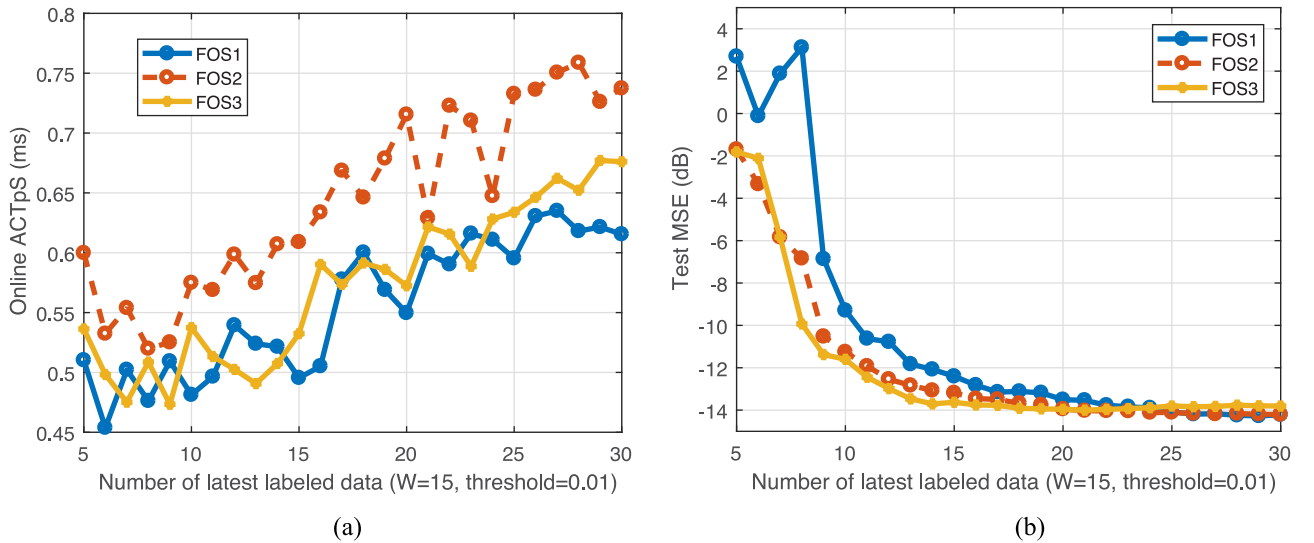


Fig. 10. Influence of number of latest labeled data samples p on: (a) online average computational time per sample, and (b) online prediction accuracy, both obtained given $W = 15$ and $\varepsilon = 0.01$ for three testing datasets of MHP.

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 physicochemical 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 commands, 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

$$\mathbf{u}(t) = [u_{p_1}(t) \ u_{p_2}(t) \ u_{p_3}(t) \ u_{p_4}(t) \ u_{p_5}(t) \ v(t)]^T. \quad (43)$$

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 j th FOS's location can be adequately represented by Wang [47,49,59]

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

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

$$\mathbf{x}_j(t) = [y_{s_j}(t-1) \ \mathbf{u}^T(t-1)]^T \in \mathbb{R}^7. \quad (45)$$

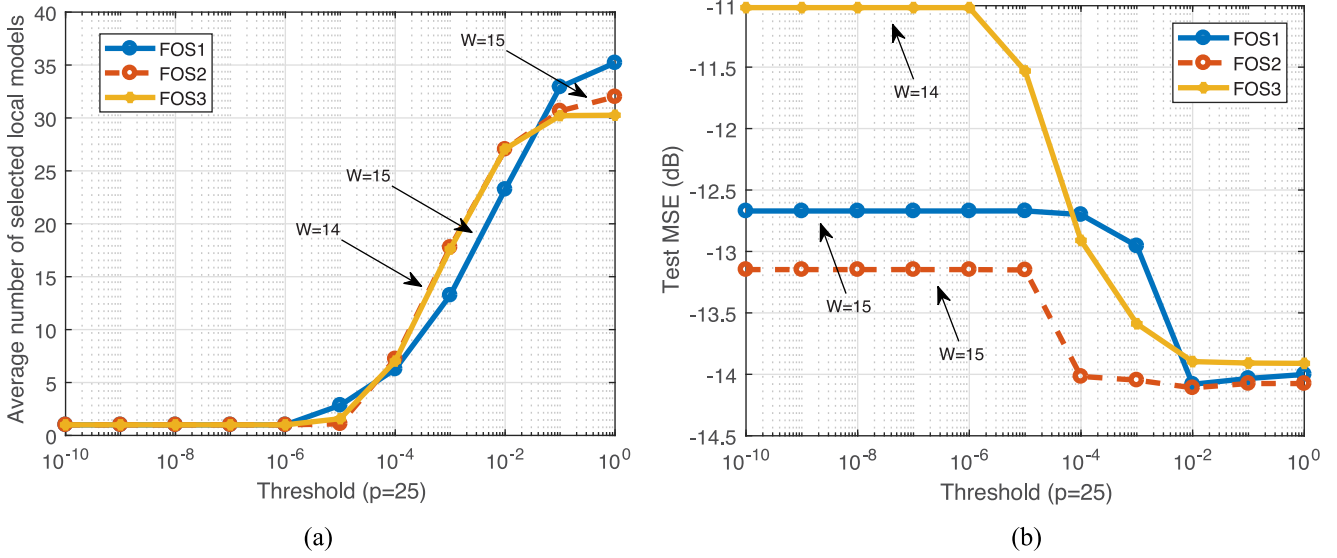


Fig. 11. Influence of threshold ε on: (a) average selected ensemble size, and (b) online prediction accuracy, given $p = 25$, $W = 15$ for FOS1 and FOS2 and $W = 14$ for FOS3 of MHP.

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 |

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 contains 3000 data samples. We first normalize the five microwave power inputs and the temperature measurements according to

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

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

where y_{\min,s_j} and y_{\max,s_j} are the minimum and maximum temperature measurements of the j th FOS, respectively. For each FOS's dataset, we use the first 1000 samples for model training, and the last 2000 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 $p = 25$ 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 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.

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 Fig. 12(a)–(c), for the 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

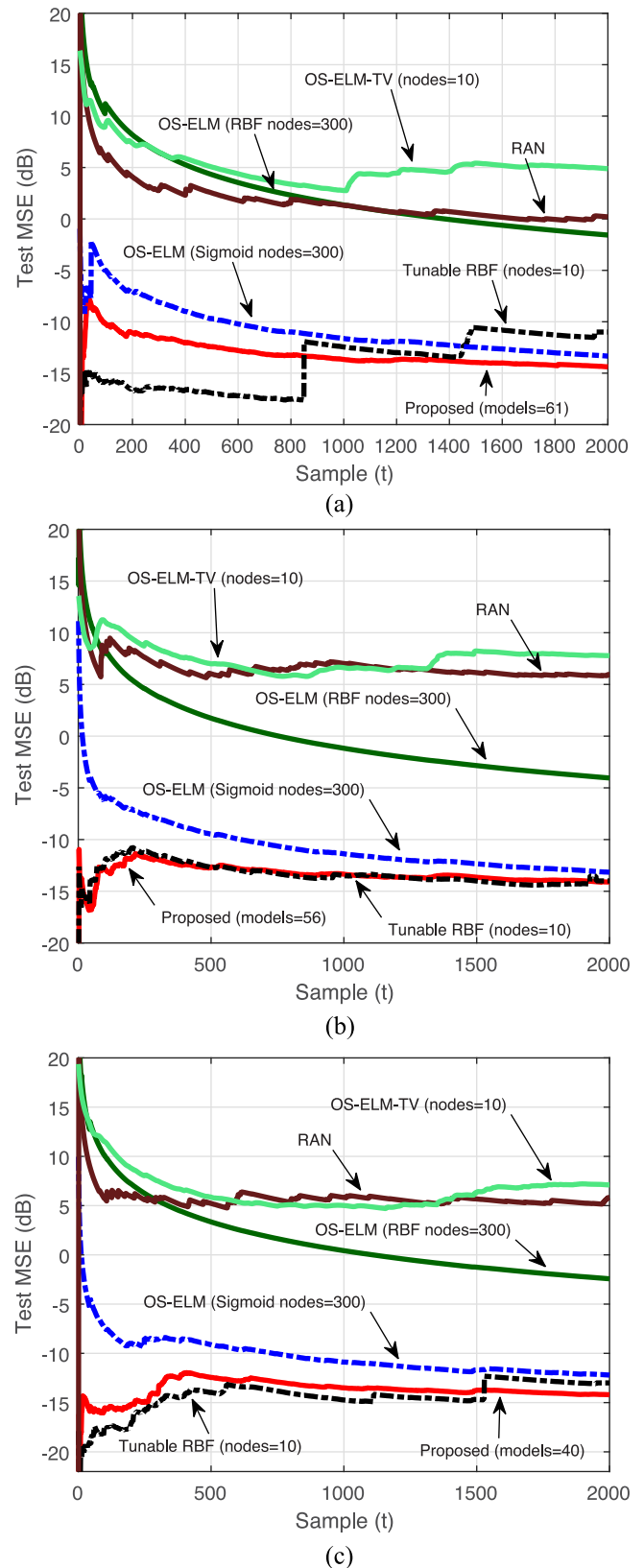


Fig. 12. Test MSE learning curves for various modeling methods: (a) FOS1, (b) FOS2, and (c) FOS3 of MHP.

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.

Declaration of Competing Interest

The authors have no conflict of interest to declare.

Acknowledgment

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 Netw.* 15 (3) (2004) 576–596.
- [2] J. Liu, D.S. Chen, Nonstationary fault detection and diagnosis for multimode processes, *AIChE J.* 56 (1) (2010) 207–219.
- [3] W. Bartelmus, F. Chaari, R. Zimroz, M. Haddar, Modelling of gearbox dynamics under time-varying nonstationary load for distributed fault detection and diagnosis, *Eur. J. Mech. A/Solids* 29 (4) (2010) 637–646.
- [4] M. Kano, M. Ogawa, The state of the art in chemical process control in Japan: good practice and questionnaire survey, *J. Process Control* 20 (9) (2010) 969–982.
- [5] I.B. Khediri, M. Limam, C. Weihs, Variable window adaptive kernel principal component analysis for nonlinear nonstationary process monitoring, *Comput. Ind. Eng.* 61 (3) (2011) 437–446.
- [6] Y. Zhang, T. Chai, Z. Li, C. Yang, Modeling and monitoring of dynamic processes, *IEEE Trans. Neural Netw. Learn. Syst.* 23 (2) (2012) 277–284.
- [7] Y. Zhang, J. An, H. Zhang, Monitoring of time-varying processes using kernel independent component analysis, *Chem. Eng. Sci.* 88 (2013) 23–32.
- [8] X. Yuan, Z. Ge, 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.* 11 (2) (2015) 209–219.
- [9] H. Jin, et al., Dual learning-based online ensemble regression approach for adaptive soft sensor modeling of nonlinear time-varying processes, *Chemomet. Intell. Lab. Syst.* 151 (2016) 228–244.
- [10] C.A. Vriezina, S. Sánchez-Pedreño, J. Grasman, Thermal runaway in microwave heating: a mathematical analysis, *Appl. Math. Model* 26 (11) (2002) 1029–1038.
- [11] S. Chandrasekaran, S. Ramanathan, T. Basak, Microwave material processing - a review, *AIChE J.* 58 (2) (2012) 330–363.
- [12] G. Ditzler, M. Roveri, G. Alippi, R. Polikar, Learning in nonstationary environments: a survey, *IEEE Comput. Intell. Mag.* 10 (4) (2015) 12–25.
- [13] J.L. Lobo, et al., Evolving spiking neural networks for online learning over drifting data streams, *Neural Netw.* 108 (2018) 1–19.
- [14] S. Chen, S.A. Billings, W. Luo, Orthogonal least squares methods and their application to non-linear system identification, *Int. J. Control* 50 (5) (1989) 1873–1896.
- [15] S. Chen, S.A. Billings, P.M. Grant, Non-linear systems identification using neural networks, *Int. J. Control* 51 (6) (1990) 1191–1214.
- [16] S. Chen, C.F.N. Cowan, P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, *IEEE Trans. Neural Netw.* 2 (2) (1991) 302–309.
- [17] S. Chen, X. Hong, C.J. Harris, P.M. Sharkey, Sparse modelling using orthogonal forward regression with PRESS statistic and regularization, *IEEE Trans. Syst., Man Cybern. Part B* 34 (2) (2004) 898–911.
- [18] S. Chen, X.X. Wang, C.J. Harris, NARX-Based nonlinear system identification using orthogonal least squares basis hunting, *IEEE Trans. Control Syst. Technol.* 16 (1) (2008) 78–84.
- [19] X. Hong, et al., Model selection approaches for nonlinear system identification: a review, *Int. J. Syst. Sci.* 39 (10) (2008) 925–946.
- [20] S. Chen, X. Hong, B.L. Luk, C.J. Harris, Non-linear system identification using particle swarm optimisation tuned radial basis function models, *Int. J. Bio-Inspired Comput.* 1 (4) (2009) 246–258.
- [21] S. Chen, X. Hong, C.J. Harris, Particle swarm optimization aided orthogonal forward regression for unified data modelling, *IEEE Trans. Evol. Comput.* 14 (4) (2010) 477–499.
- [22] S. Chen, X. Hong, J.B. Gao, C.J. Harris, Complex-valued b-spline neural networks for modeling and inverting Hammerstein systems, *IEEE Trans. Neural Netw. Learn. Syst.* 25 (9) (2014) 1673–1685.
- [23] X. Hong, S. Chen, Elastic net orthogonal forward regression, *Neurocomputing* 148 (2015) 551–560.
- [24] X. Hong, S. Chen, J. Gao, C.J. Harris, Nonlinear identification using orthogonal forward regression with nested optimal regularization, *IEEE Trans. Cybern.* 45 (12) (2015) 2925–2936.
- [25] X. Hong, S. Chen, Y. Guo, J. Gao, l^1 -Norm penalized orthogonal forward regression, *Int. J. Syst. Sci.* 48 (10) (2017) 2195–2201.
- [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 Netw. Learn. Syst.* 28 (12) (2017) 2872–2884.
- [27] J. Moody, C.J. Darken, Fast learning in networks of locally-tuned processing units, *Neural Comput.* 1 (2) (1989) 281–294.
- [28] S. Chen, S.A. Billings, Recursive prediction error estimator for non-linear models, *Int. J. Control* 49 (2) (1989) 569–594.
- [29] S. Chen, Nonlinear time series modelling and prediction using gaussian RBF networks with enhanced clustering and RLS learning, *Electron Lett.* 31 (2) (1995) 117–118.
- [30] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, *Neurocomputing* 70 (1–3) (2006) 489–501.
- [31] N. Liang, G. Huang, P. Saratchandran, N. Sundararajan, A fast and accurate online sequential learning algorithm for feedforward networks, *IEEE Trans. Neural Netw.* 17 (6) (2006) 1411–1423.
- [32] G.-B. Huang, L. Chen, Enhanced random search based incremental extreme learning machine, *Neurocomputing* 71 (16) (2008) 3460–3468.
- [33] F. Ding, P.X. Liu, G. Liu, Multinnovation least-squares identification for system modeling, *IEEE Trans. Syst. Man Cybern. Part B* 40 (3) (2010) 767–778.
- [34] X. Hong, G.D. Fatta, H. Chen, S. Wang, Sparse Least Squares Support Vector Regression for Nonstationary Systems, in: *Proceedings of the IJCNN, 2018*, pp. 1–8. (Rio de Janeiro, Brazil)
- [35] Y. Ye, S. Squartini, F. Piazza, Online sequential extreme learning machine in nonstationary environments, *Neurocomputing* 116 (2013) 94–101.
- [36] J. Platt, A resource-allocating network for function interpolation, *Neural Comput.* 3 (2) (1991) 213–225.
- [37] V. Kadirkamanathan, M. Niranjan, A function estimation approach to sequential learning with neural networks, *Neural Comput.* 5 (6) (1993) 954–975.
- [38] H. Chen, Y. Gong, X. Hong, S. Chen, A fast adaptive tunable RBF network for nonstationary systems, *IEEE Trans. Cybern.* 46 (12) (2016) 2683–2692.
- [39] H. Tong, K.S. Lim, Threshold autoregression, limit cycles and cyclical data, *J. R. Stat. Soc.* 42 (3) (1980) 245–292.
- [40] H. Tong, *Threshold Models in Non-Linear Time Series Analysis*, Springer-Verlag, New York, 1983.
- [41] S.A. Billings, S. Chen, Extended model set, global data and threshold model identification of severely non-linear systems, *Int. J. Control* 50 (5) (1989) 1897–1923.
- [42] X. Hong, Y. Gong, A Constrained Recursive Least Squares Algorithm for Adaptive Combination of Multiple Models, in: *Proceedings of the IJCNN 2015, 2015*, pp. 1–6. (Killarney, Ireland)
- [43] H. Chen, Y. Gong, X. Hong, A new adaptive multiple modelling approach for non-linear and non-stationary systems, *Int. J. Syst. Sci.* 47 (9) (2016) 2100–2110.
- [44] W. Shao, et al., Online soft sensor design using local partial least squares models with adaptive process state partition, *Chemomet. Intell. Labor. Syst.* 144 (2015) 108–121.
- [45] W. Shao, S. Chen, C.J. Harris, Adaptive soft sensor development for multi-output industrial processes based on selective ensemble learning, *IEEE Access* 6 (2018) 55628–55642.
- [46] E.N. Lorenz, Deterministic nonperiodic flow, *J. Atmosph. Sci.* 20 (1963) 130–141.
- [47] K. Wang, et al., Learning to detect local overheating of the high-power microwave heating process with deep learning, *IEEE Access* 6 (2018) 10288–10296.

- [48] J. Zhong, S. Liang, Q. Xiong, Improved receding horizon h_∞ temperature spectrum tracking control for Debye media in microwave heating process, *J. Process Control* 71 (2018) 14–24.
- [49] T. Liu, S. Liang, Q. Xiong, K. Wang, Adaptive critic based optimal neurocontrol of a distributed microwave heating system using diagonal recurrent network, *IEEE Access* 6 (2018) 68839–68840.
- [50] T. Liu, S. Liang, Q. Xiong, K. Wang, Two-stage method for diagonal recurrent neural network identification of a high-power continuous microwave heating system, *Neural Process. Lett.* 1–22 (2019).
- [51] D.H. Kaye, D.A. Freedman, *Reference Guide on Statistics*, 3rd ed., National Academy Press, Washington DC, 2011.
- [52] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, *Neural Comput* 3 (1) (1991) 79–87.
- [53] X. Hong, C.J. Harris, A mixture of experts network structure construction algorithm for modelling and control, *Appl. Intell.* 16 (1) (2002) 59–69.
- [54] V. Cherkassky, Y. Ma, Multiple model regression estimation, *IEEE Trans. Neural Netw.* 16 (4) (2005) 785–798.
- [55] Z.-H. Zhou, J. Wu, W. Tang, Ensemble neural networks: many could be better than all, *Artif. Intell.* 137 (1–2) (2002) 239–263.
- [56] L.-J. Zhao, T.-Y. Chai, D.C. Yuan, Selective ensemble extreme learning machine modeling of effluent quality in wastewater treatment plants, *Int. J. Autom. Comput.* 9 (6) (2012) 627–633.
- [57] J. Zhong, L. Shan, Q. Xiong, Receding horizon h^∞ guaranteed cost tracking control for microwave heating medium with temperature-dependent permittivity, *ISA Trans.* 73 (2018) (2018) 249–256.
- [58] J. Zhong, S. Liang, Y. Yuan, Q. Xiong, Coupled electromagnetic and heat transfer ODE model for microwave heating with temperature-dependent permittivity, *IEEE Trans. Microwave Theory Techn.* 64 (8) (2016) 2467–2477.
- [59] X. Shi, et al., Research of uniformity evaluation model based on entropy clustering in the microwave heating processes, *Neurocomputing* 173 (3) (2016) 562–572.



Tong Liu received the B.Sc. degree in automation from the College of Automation, Chongqing University, Chongqing, China, in 2016, where he is currently pursuing the Ph.D. degree in control theory and control engineering. Since 2018, he has been a Visiting Ph.D. Student with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. His research interests include non-stationary time series modeling, online learning and neural networks, intelligent control system design, and with their industrial applications.



Sheng Chen received the B.Eng degree from East China Petroleum Institute (now China University of Petroleum), China, in January 1982, and the Ph.D degree from City University, London, in September 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Science (DSc), by the University of Southampton. From October 1986 to August 1999, he held various research and academic posts with University of Sheffield, University of Edinburgh and University of Portsmouth, all in UK. Since September, 1999, he has been with School of Electronics and Computer Science, University of Southampton, where he holds the post of Professor in Intelligent Systems and Signal Processing.

Professor Chen's research interests are in computation intelligence, wireless communications and signal processing. He is a Fellow of the United Kingdom Royal Academy of Engineering, a fellow of IEEE and fellow of IET. He is one of the original ISI most highly cited researchers in engineering (March 2004).



Shan Liang received his M.Sc. degree in control science and engineering from the College of Automation, Chongqing University, Chongqing, China, in 1995, and the Ph.D. degree from the Department of Mechanical Systems Engineering, Kumamoto University, Kumamoto, Japan, in 2004. His current research interests include numerical modeling, electromagnetic theory, nonlinear systems, adaptive control, and sensor networks. He is a Member of IEEE.



Chris J. Harris received his BSc and MA degrees from the University of Leicester and the University of Oxford in UK, respectively, and his PhD degree from the University of Southampton, UK, in 1972. He was awarded the higher doctoral degree, the Doctor of Sciences (DSc), by the University of Southampton in 2001. He is Emeritus Research Professor at the University of Southampton, having previously held senior academic appointments at Imperial College, Oxford and Manchester Universities, as well as Deputy Chief Scientist for the UK Government.

Professor Harris was awarded the IEE senior Achievement Medal for Data Fusion research and the IEE Faraday Medal for distinguished international research in Machine Learning. He was elected to the UK Royal Academy of Engineering in 1996. He is the co-author of over 500 scientific research papers during a 50 year research career.