Renewable Energy 174 (2021) 1–18

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Probabilistic wind power forecasting using selective ensemble of finite mixture Gaussian process regression models

Huaiping Jin [a, b, *], Lixian Shi [a, b], Xiangguang Chen [c], Bin Qian [a, b], Biao Yang [a, b], Huaikang Jin [d]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China
[b] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China
[c] School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing, 100081, China
[d] Huaneng Renewables Co., Ltd. Yunnan Branch, Kunming, 650000, China

ARTICLE INFO

Article history:
Received 5 August 2020
Received in revised form
24 March 2021
Accepted 6 April 2021
Available online 20 April 2021

Keywords:
Wind power forecasting
Ensemble learning
Gaussian process regression
Probabilistic modeling
Ensemble pruning
Model adaptation

ABSTRACT

Ensemble learning models have been widely used for wind power forecasting to facilitate efficient dispatching of power systems. However, traditional ensemble methods cannot always function well due to insufficient accuracy and diversity of base learners, ignorance of ensemble pruning, as well as the lack of adaptation capability. Therefore, a novel probabilistic wind power forecasting method is proposed based on selective ensemble of finite mixture Gaussian process regression models (SEFMGPR). First, a set of diverse local Gaussian process regression (GPR) models are constructed through multimodal perturbation mechanism, i.e., perturbing the training data and input attributes simultaneously. Then, a set of finite mixture GPR models (FMGPR) is built by integrating local GPR models through finite mixture mechanism (FMM). Next, the highly influential FMGPR models are selected using genetic algorithm (GA) based ensemble pruning. When a new test sample comes, the component predictions from the selected FMGPR models are adaptively combined by using FMM again and the probabilistic prediction results of the SEFMGPR model are obtained. Besides, an incremental adaptation mechanism is used to alleviate performance degradation of SEFMGPR. The application results from a real wind farm dataset show that SEFMGPR outperforms the traditional global and ensemble wind power prediction methods, and can maintain high prediction accuracy by effectively handling time-varying changes of wind power data.

1. Introduction

With the depletion of fossil energy, the exploitation of renewable energy has attracted considerable attention all over the world. Owing to the advantages of no pollution and wide distribution of wind energy, wind power production technology has been developed rapidly in recent years. However, the intermittence and fluctuation of wind energy will greatly affect the security and stability of the power system. Consequently, accurate and reliable wind power forecasting is of vital importance to facilitate reasonable power dispatching and arrange shutdown and maintenance to guarantee the stable operation of the power system [1,2].

Generally, according to the forecasting horizon, wind power forecasting can be categorized into four types: ultra-short-term, short-term, medium-term and long-term [3]. Ultra-short-term wind power forecasting aims to ensure the real-time and stable dispatching of the power grid and high quality of power supply. The purpose of short-term forecasting is to make power generation plans, arrange regional dispatching, and adjust maintenance plans. The medium and long-term predictions mainly serve the maintenance plan of wind turbines and transmission lines, which have low requirements on prediction accuracy. To meet the different needs of wind power forecasting, a variety of predictive models have been developed. In general, these forecasting methods can be split into two groups: physical methods and statistical methods. The former class of approaches mainly rely on numerical weather prediction (NWP) information, which is usually suitable for medium and long term wind power prediction [4]. In contrast, statistical methods try to develop empirical models based on historical time series data. The focus of this work is on the development of data-driven models for ultra-short term forecasting.

* Corresponding author. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China.
E-mail addresses: jinhuaiping@gmail.com, jinhuaiping@kust.edu.cn (H. Jin).

https://doi.org/10.1016/j.renene.2021.04.028
0960-1481/

In the early stage, statistical models are mainly based on autoregressive integrated moving average (ARIMA) [5], artificial neural network (ANN) [6], and Kalman Filter [7], etc. Subsequently, machine learning methods such as support vector machine (SVM) [8] and extreme learning machine (ELM) [9] have been introduced to wind power forecasting. In recent years, as a branch of machine learning, deep learning (DL) methods such as deep belief network (DBN) [10,11] and long short term memory network (LSTM) [12,13] have gained much attention for wind power prediction.

Despite the availability of statistical prediction methods, it is still difficult to achieve accurate forecasting performance because of the randomness and fluctuation of wind energy. Therefore, ensemble learning methods for wind power forecasting have received rapidly growing attention [14–16]. The base idea of such approaches is to pursue performance enhancement by training and combining multiple base learners for a prediction task [17,18]. One popular approach for this purpose is to build accurate and diverse base learners by perturbing the training data and then combine them. For example, two famous ensemble paradigms of this type, bagging and boosting methods, have been applied to wind power forecasting [19–22]. In addition, the ensemble forecasting models can also be constructed based on signal decomposition methods. To this end, the original power time series are decomposed into multiple sub-sequences, for which diverse base learners are built and integrated [23]. The commonly used decomposition methods include empirical mode decomposition (EMD) [24], wavelet transform (WT) [25], variational mode decomposition (VMD) [26,27] etc. While many ensemble prediction models have been proposed, there remain some problems in delivering accurate predictions of wind power.

A particular drawback of many of the current ensemble wind power forecasting models is the insufficiency of the diversity of base learners. Numerous studies have shown that the key to building a strong ensemble lies in generating accurate but diverse base learners [28,29]. In particular, guaranteeing the diversity among base learners is critical for the success of an ensemble model. Popular mechanisms for diversity generation include perturbing the data samples [19–22], input attributes [30], model parameters [31], and learning algorithms [32]. Nevertheless, many researchers only consider single perturbation for diversity creation, which may be inadequate for building high-performance ensemble models. Consequently, a multimodal perturbation mechanism [29] is strongly desirable.

Another factor limiting the ensemble prediction performance is the inappropriate combination of base learners. Two crucial problems should be addressed to achieve efficient base learner integration, i.e., deciding which learners are included in the ensemble and determining their weights. According to Zhou's finding [28], it may be better to ensemble many instead of all base learners, which brings the advantages of potentially improving the ensemble performance and reducing the model complexity. Therefore, selective ensemble has recently become a promising direction for enhancing ensemble performance [33,34]. After selecting the desirable base learners, suitable weights should be assigned to each learner. The frequently used methods for this purpose consist of simple averaging and weighting averaging by optimization [35–38] and learning methods [39,40], which often provide the fixed weights in advance. Since the base learners usually behave differently for different prediction points, the adaptive weighting strategies are preferable [41].

Traditional ensemble wind power prediction models are non-adaptive and thus cannot implement online updating according to the newly obtained information. Traditionally, once deployed into real-life operation, wind power forecasting models do not change anymore, which is not suitable for the time-varying characteristics of wind energy. To ensure high prediction performance for a long period of time, it is appealing to equip the ensemble models with adaptation mechanisms for enabling accurate wind power forecasting.

Moreover, it is difficult to evaluate the reliability of the prediction results. Generally, the deterministic ensemble forecasting methods focus on point prediction, whereas the estimations of prediction uncertainty are not considered. Alternatively, probabilistic forecasting models can provide confidence intervals of predictions, in addition to the point prediction outputs [42]. Hence, in this work, we attempt to employ Gaussian process regression (GPR) [43–45] as the base leaner for developing well-performing probabilistic ensemble forecasting models.

To address the above-mentioned issues, a novel wind power forecasting method, referred to as selective ensemble of finite mixture Gaussian process regression (SEFMGPR), is proposed. First, by exploiting multimodal perturbation mechanism, diverse subspaces are constructed by integrating bootstrapping and partial least squares regression (PLS), and local domains (LDs) are identified by performing Gaussian mixture model (GMM) clustering. Then, local GPR models are built for each of the LDs and further integrated as finite mixture Gaussian process regression (FMGPR) models through FMM. Next, a Genetic Algorithm (GA) based ensemble pruning strategy is used to select the highly influential FMGPR models. When an estimation is requested, the component predictions from the selected FMGPR models are adaptively weighted as the prediction output and variance of SEFMGPR. In addition, the SEFMGPR model is updated incrementally. Compared to traditional ensemble methods, the SEFMGPR method exhibits the following characteristics and advantages:

(1) A multimodal perturbation mechanism, which combines the perturbations on training data and input attributes together, is helpful to enhance the accuracy and diversity of base learners.
(2) A GA based ensemble pruning allows SEFMGPR to further enhance the ensemble prediction performance and significantly reduce the model complexity.
(3) With a two-level adaptive combination scheme, local GPR models and FMGPR models are adaptively integrated through FMM mechanism.
(4) The introduction of an incremental adaptation mechanism enables SEFMGPR to effectively alleviate performance degradation.
(5) As well as providing accurate and reliable estimations, the SEFMGPR method also gives the prediction confidence intervals, which is helpful for plant operators to evaluate the prediction reliability.

The rest of the paper proceeds as follows. Section 2 briefly introduces the basic principles of GPR, GMM, and PLS. Section 3 presents the details of the proposed SEFMGPR method. The case study is reported to verify the effectiveness and superiority of SEFMGPR in Section 4. Finally, the concluding remarks are drawn in Section 5.

2. Preliminaries

Gaussian process regression (GPR), Gaussian mixture model (GMM), partial least squares regression (PLS), and mutual information (MI) are introduced in Appendix A.

3. Proposed SEFMGPR for wind power forecasting

In this section, a novel wind power prediction framework,

referred to as selective ensemble of finite mixture Gaussian process regression (SEFMGPR), is presented. The basic concept of SEFMGPR modeling is illustrated in Fig. 1. The proposed SEFMGPR forecasting method can be divided into two stages: the offline modeling stage and the online prediction stage. In the offline stage, a set of diverse and accurate base FMGPR models are built through a multimodal perturbation mechanism and GA based ensemble pruning. In the online prediction stage, when a test sample arrives, the prediction outputs and variances with respect to different base FMGPR models are obtained. Then, an adaptive combination strategy is used to get the final prediction results of the SEFMGPR model. Moreover, to alleviate the performance deterioration of the SEFMGPR model, an incremental adaptation mechanism is implemented on the proposed model when the actual wind power data are available. The details of the SEFMGPR method are discussed in the following sections.

3.1. Multimodal perturbation mechanism

It is well known that to build a high-performance ensemble model, the base learners should be with high accuracy and diversity [46]. Especially, ensemble diversity, that is, the difference among the base learners, plays a vital role in ensemble modeling [47]. As the famous error decomposition theory reveals [48,49], the generalization error of an ensemble depends on the term of diversity. Though diversity is crucial to ensemble performance, there is still no well-accepted formal definition of diversity. The common basic ideas of generating diversity are manipulating input attributes, training samples, learning parameters, and output representations [46]. However, in the context of wind power forecasting, it is a common practice to use sole perturbation to create the diversity by applying heterogeneous models [50,51], resampled training subsets [19,20], etc. Consequently, to enhance the diversity of base learners, a multimodal perturbation mechanism is proposed by integrating perturbations on training data and input attributes, as depicted in Fig. 2. The basic idea of this approach is to build diverse subspaces by integrating bootstrapping and PLS regression analysis, and then construct diverse local domains (LDs) by GMM clustering for each of subspaces.

3.1.1. Building of diverse subspaces

A data set is usually described with a set of attributes, and thus different attribute subspaces, i.e. attribute subsets, might provide different views on the data [29]. A popular strategy for building subspaces is the random subspace method [52] where a series of subspaces are obtained by random resampling on input attributes. Though base learners trained from random resampling subsets of attributes might be quite diverse, such operation cannot guarantee

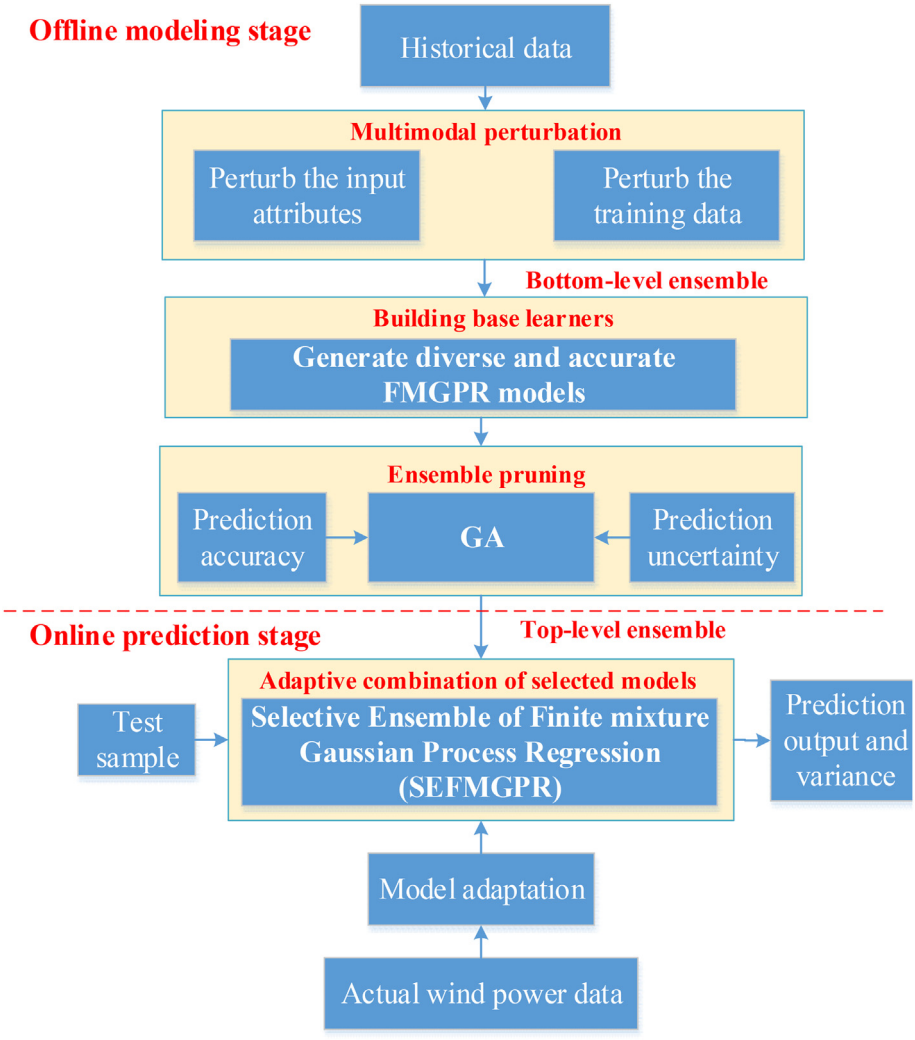

Fig. 1. Concept of SEFMGPR framework for wind power forecasting.

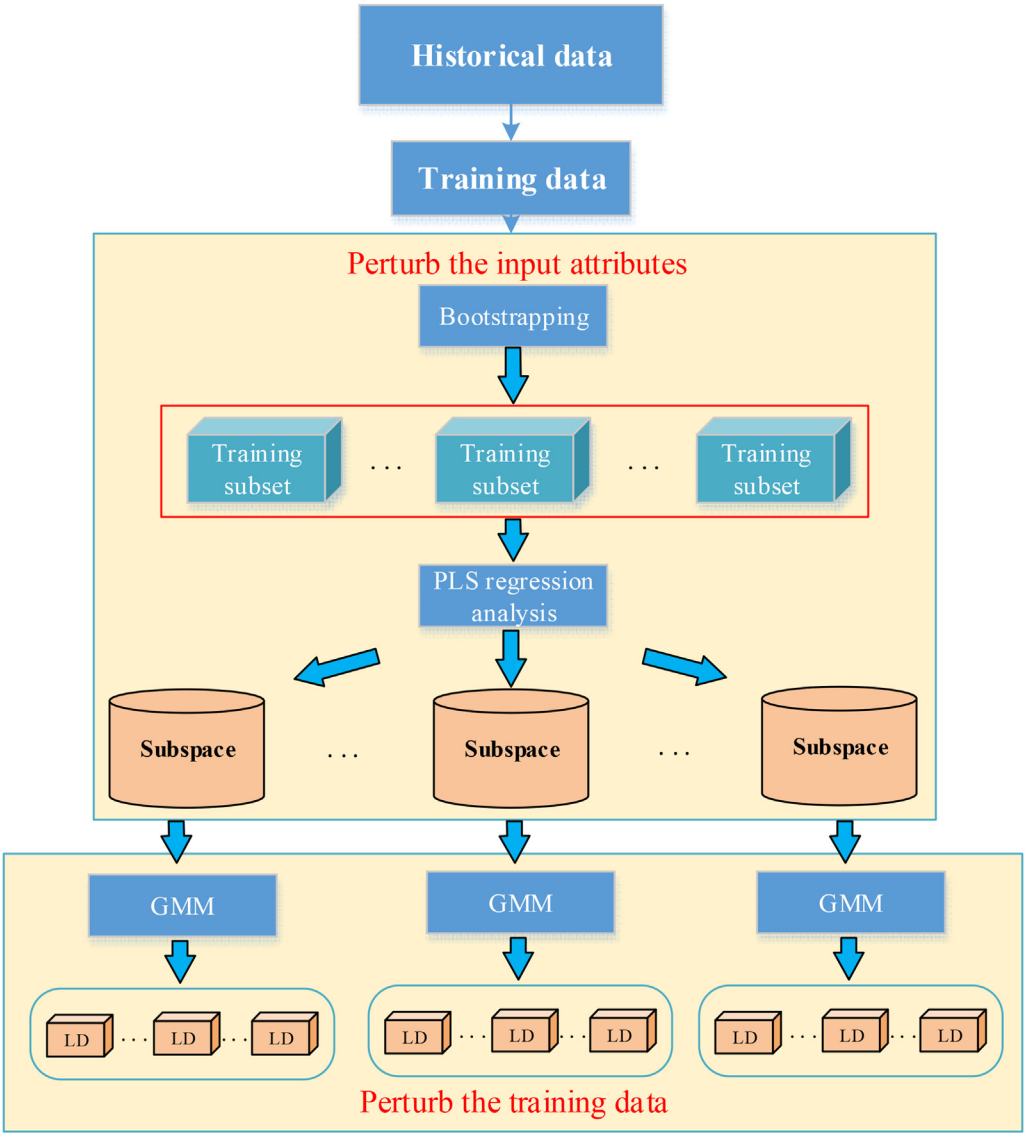

Fig. 2. Diagram of the proposed multimodal perturbation mechanism.

the performance of subspace models. Instead of selecting input attributes through random resampling via attribute direction, we attempt to obtain subspaces according to the relevance between input and output attributes by performing PLS regression on resampling training subsets.

Given the training sample set D_{trn}, a set of subsets is obtained through bootstrapping:

$$D_{trn} \xrightarrow{Bootstrapping} \{D_1, D_2, \ldots, D_r\} \tag{1}$$

where r is the number of resampling subsets.

Subsequently, PLS regression analysis is performed on each built training subsets $\{D_1, D_2, \ldots, D_r\}$. Since the regression coefficients of PLS models can reflect the importance of input attributes to output, a subspace can be obtained by selecting input attributes with large regression coefficients. Without loss of generality, assume the regression coefficients of a PLS model are represented as $\boldsymbol{\beta} = [b_1, b_2, b_3, \ldots, b_d]^{\mathrm{T}}$, where d is the number of input attributes. Then, the elements of $\boldsymbol{\beta}$ are rearranged as $\boldsymbol{\beta}' = [b'_1, b'_2, \ldots, b'_d]^{\mathrm{T}}$ in descending order according to their absolute values, i.e., $b'_1 > b'_2 > \ldots > b'_d$. Next, the ith candidate attribute x_i respect to b'_i is selected as the member of a subspace if satisfying the following condition:

$$\frac{|b'_1| + |b'_2| + \ldots + |b'_i|}{|b_1| + |b_2| + |b_3| + \ldots + |b_d|} > \eta \tag{2}$$

where η is a threshold controlling the size of selected input attributes. The parameter η should be neither too large nor too small. A large η will result in the addition of less irrelevant variables while a smallη may result in the removal of important variables.

Repeat the above attribute selection process for r times and subspaces $\{S_1, S_2, \ldots, S_m\}$ are obtained after removing the duplicated ones.

3.1.2. Building of local domains

Traditionally, wind power forecasting is achieved based on a single predictive model given the underlying assumption of a constant mode and conditions throughout the process of wind

power generation. In practice, however, wind power data often encounter multimode and non-Gaussian characteristics due to changes in climate conditions. Thus, the accuracy and reliability of wind power predictions may significantly degrade. To tackle this problem, Gaussian mixture models (GMM) are used for clustering on wind power data to capture local characteristics effectively [53]. Then, some ensemble predictive models can be built to deliver more accurate prediction results.

Provided a training set $D_{trn} = \{\mathbf{X}_{trn}, \mathbf{y}_{trn}\}$, assume that input data $\mathbf{X}_{trn}$ follow a Gaussian mixture distribution. Then, a GMM model with K components can be estimated from $\mathbf{X}_{trn}$ and D_{trn} can be divided into K local domains $\{LD_1, LD_2, \ldots, LD_K\}$.

Since the sole perturbation on attributes or training data cannot obtain adequate diversity and describe local process characteristics, the multimodal perturbation is desirable. Therefore, to get a high-performance ensemble predictive model, a set of subspaces is first built and then a group of LDs is obtained for each subspace by performing GMM clustering in this work.

3.2. Building of FMGPR models

After constructing a set of subspaces $\{S_1, S_2, \ldots, S_m\}$ and identifying the LDs $\{\{LD_i^{(1)}\}_{i=1}^{K_1}, \{LD_i^{(2)}\}_{i=1}^{K_2}, \cdots, \{LD_i^{(m)}\}_{i=1}^{K_m}\}$, where K_m is the number of mixture components for the mth subspace, one FMGPR model is built from a set of diverse local GPR models for each of the subspaces. Without loss of generality, for a given subspace S_m, diverse local GPR models are first built from each LD, and the local model set for the mth subspace can be denoted as

$$\mathscr{M}^{(m)} = \left\{GPR_1^{(m)}, GPR_2^{(m)}, \ldots, GPR_{K_m}^{(m)}\right\} \tag{3}$$

Subsequently, for a test sample $\mathbf{x}_*$, a set of local prediction outputs and variances can be obtained from the local model set $\mathscr{M}^{(m)}$ and local predictions of $GPR_k^{(m)}$ can be given as

$$GPR_k^{(m)}: \begin{cases} \hat{y}_{k,*}^{(m)} = \left(\mathbf{k}_{k,*}^{(m)}\right)^{\mathrm{T}} \left(\mathbf{C}_{r,k}^{(m)}\right)^{-1} \mathbf{y}_k \\ \left(\sigma_{k,*}^2\right)^{(m)} = C\left(\mathbf{x}_{k,*}^{(m)}, \mathbf{x}_{k,*}^{(m)}\right) - \left(\mathbf{k}_{k,*}^{(m)}\right)^{\mathrm{T}} \left(\mathbf{C}_{m,k}^{(m)}\right)^{-1} \mathbf{k}_{k,*}^{(\mathrm{m})}, k = 1, 2, \cdots, K_m \end{cases} \tag{4}$$

The next step is to combine the prediction outputs and variances by the FMM [54]:

$$\begin{cases} \hat{y}_*^{(m)} = \sum_{k=1}^{K_m} p\left(LD_k^{(m)} | \mathbf{x}_*^{(m)}\right) \hat{y}_{k,*}^{(m)} \\ \left(\hat{\sigma}_*^2\right)^{(m)} = \sum_{k=1}^{K_m} p\left(LD_k^{(m)} | \mathbf{x}_*^{(m)}\right) \left[\left(\sigma_{k,*}^2\right)^{(m)} + \left(\hat{y}_{k,*}^{(m)} - \hat{y}_*^{(m)}\right)^2\right] \end{cases} \tag{5}$$

where $\mathbf{x}_*^{(m)}$ is the test input data of the mth subspace, and $p(LD_k^{(m)} | \mathbf{x}_*^{(m)})$ is the posterior probability of $\mathbf{x}_*^{(m)}$ with respect to the kth local domain of the mth subspace, which is estimated by Bayes' rule as follows

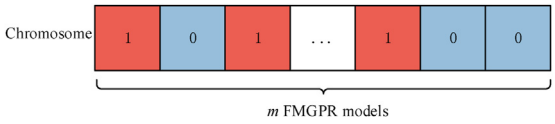

Fig. 3. A typical chromosome for selection of FMGPR models.

$$\begin{aligned} p\left(LD_k^{(m)} | \mathbf{x}_*^{(m)}\right) &= \frac{p\left(\mathbf{x}_*^{(m)} | LD_k^{(m)}\right) p\left(LD_k^{(m)}\right)}{p\left(\mathbf{x}_*^{(m)}\right)} \\ &= \frac{p\left(\mathbf{x}_*^{(m)} | LD_k^{(m)}\right) p\left(LD_k^{(m)}\right)}{\sum_{l=1}^{K_m} p\left(\mathbf{x}_*^{(m)} | LD_l^{(m)}\right) p\left(LD_l^{(m)}\right)}, k = 1, 2, \ldots, K_m \end{aligned} \tag{6}$$

where $p(\mathbf{x}_*^{(m)} | LD_k^{(m)})$ is the conditional probability, and $p(LD_k^{(m)})$ is the mixing coefficient π_k of GMM.

3.3. Selective ensemble of FMGPR models

After building diverse FMGPR models, this section focuses on how to integrate these base learners. A general approach for this purpose is to combine all base learners through a certain weighting strategy. However, such methods usually encounter two challenging issues, i.e., (i) combining all base learners without pruning may deteriorate the ensemble prediction rather than improvement while model complexity will increase [28], and (ii) traditional weighting methods such as simple averaging are non-adaptive and thus fail to accommodate the time-varying process characteristics [55]. To tackle these problems, we employ a GA based ensemble pruning method and a FMM based adaptive integration scheme.

The basic idea of the proposed GA based ensemble pruning is to choose influential base learners by solving a binary optimization problem through a GA method. Let $\mathbf{s} \in \mathbb{R}^m$ be a solution with each element of $\mathbf{s}$ indicating whether a FMGPR model is selected or not, then the pruning of FMGPR models can be achieved by solving the optimization problem as follows:

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} f(\mathbf{s}) \tag{7}$$

where $f(\mathbf{s})$ is the objective function, which is defined based on the

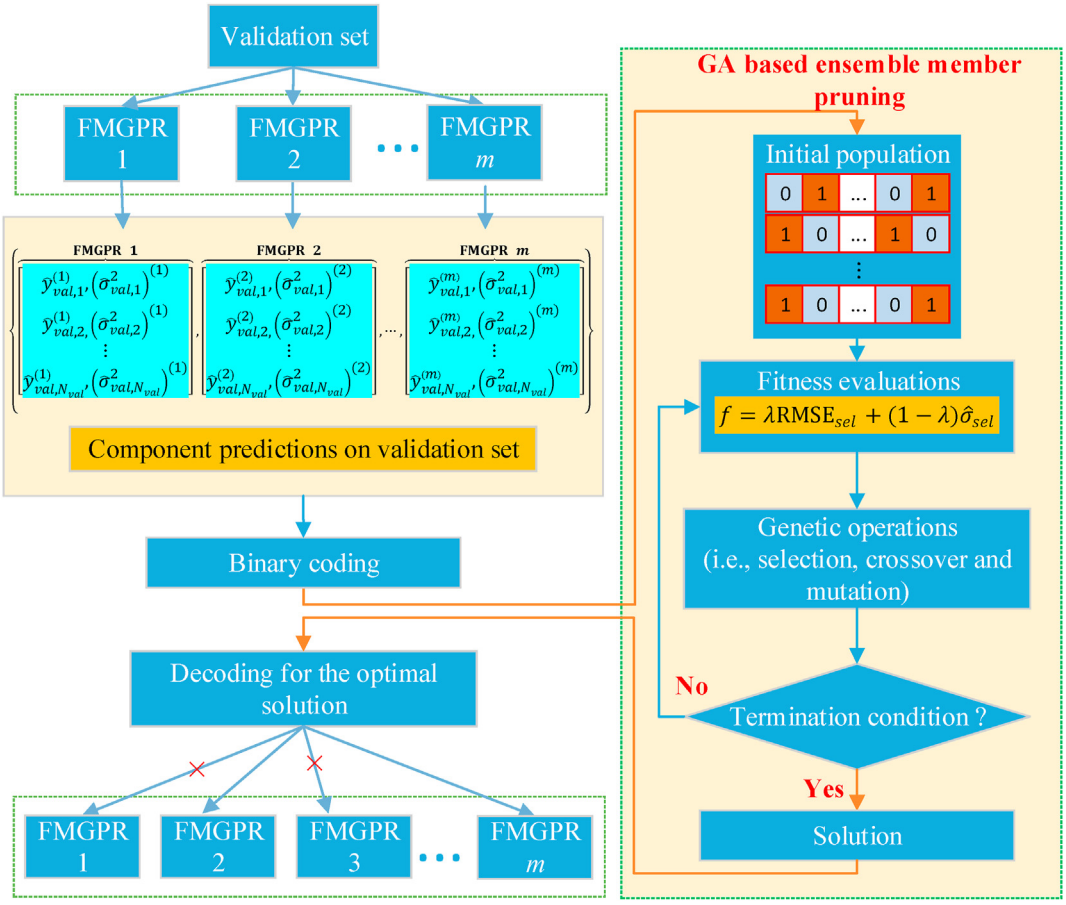

Fig. 4. Flow diagram of the GA based ensemble pruning.

prediction accuracy and uncertainty, i.e.,

$$f(\mathbf{s}) = \lambda \mathrm{RMSE}_{sel} + (1-\lambda)\hat{\sigma}_{sel} \tag{8}$$

where RMSE_{sel} is the root mean square error and $\hat{\sigma}_{sel}$ is the mean standard deviation of the selected FMGPR model, and λ is a tradeoff parameter for balancing the importance of prediction accuracy and uncertainty.

Since it is impossible to obtain the actual prediction performance of FMGPR models, RMSE_{sel} and $\hat{\sigma}_{sel}$ are estimated through an independent validation set D_{val}. To this end, the component prediction outputs $\{\hat{\mathbf{y}}_{val}^{(1)}, \hat{\mathbf{y}}_{val}^{(2)}, \ldots, \hat{\mathbf{y}}_{val}^{(m)}\}$ and variances $\{(\hat{\boldsymbol{\sigma}}_{val}^2)^{(1)}, (\hat{\boldsymbol{\sigma}}_{val}^2)^{(2)}, \ldots, (\hat{\boldsymbol{\sigma}}_{val}^2)^{(m)}\}$ for the validation set using diverse base FMGPR models are obtained, where

$$\begin{cases} \hat{\mathbf{y}}_{val}^{(j)} = \left[\hat{y}_{val,1}^{(j)}, \hat{y}_{val,2}^{(j)}, \ldots, \hat{y}_{val,N_{val}}^{(j)}\right]^{\mathrm{T}} \\ \left(\hat{\boldsymbol{\sigma}}_{val}^2\right)^{(j)} = \left[\left(\hat{\sigma}_{val,1}^2\right)^{(j)}, \left(\hat{\sigma}_{val,2}^2\right)^{(j)}, \ldots, \left(\hat{\sigma}_{val,N_{val}}^2\right)^{(j)}\right]^{\mathrm{T}} \end{cases} \tag{9}$$

where N_{val} is the number of validation samples, and $\hat{y}_{val,i}^{(j)}, (\hat{\sigma}_{val,i}^2)^{(j)}$ denote the prediction output and variance of the ith validation sample using the jth FMGPR model, respectively.

Then, RMSE_{sel} and $\hat{\sigma}_{sel}$ are calculated as follows:

$$\mathrm{RMSE}_{sel} = \sqrt{\frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \left(y_i - \hat{y}_i^{ens}\right)^2} \tag{10}$$

$$\hat{\sigma}_{sel} = \sqrt{\frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \left(\hat{\sigma}_i^{ens}\right)^2} \tag{11}$$

where y_i is the actual output of the ith validation sample, and $\hat{y}_i^{ens}, (\hat{\sigma}_i^{ens})^2$ are the ith selective ensemble prediction output and variance, which are calculated from $\{\hat{y}_{val,i}^{(1)}, \hat{y}_{val,i}^{(2)}, \ldots, \hat{y}_{val,i}^{(m)}\}$ and $\{\hat{\sigma}_{val,i}^{(1)}, \hat{\sigma}_{val,i}^{(2)}, \ldots, \hat{\sigma}_{val,i}^{(m)}\}$, respectively. The combination approach for determining the ensemble prediction results will be presented later in

this subsection.

Considering that evolutionary algorithms have gained great success as powerful optimization tools, GA, a well-known member of this family, is employed for solving the problem in Eq. (8). Therefore, as shown in Fig. 3, **s** is expressed as the style of a chromosome, where 0 and 1 indicate exclusion and inclusion of a FMGPR model, respectively, and the fitness of each chromosome is evaluated by $^{1}/_{f(\mathbf{s})}$.

Then, a population with a number of chromosomes can be generated randomly and further evolved through selection, crossover, and mutation operations. Next, the chromosome with the best fitness is selected to decide which FMGPR models are removed when stopping condition is achieved. The details of the GA based ensemble pruning process are described in **Algorithm 1** and the flow diagram is illustrated in Fig. 4.

Algorithm 1. (*GA based ensemble pruning*)

In the following, the combination strategy for achieving ensemble prediction is discussed in detail. Similar to the combination method used in FMGPR modeling, once again, the FMM mechanism is used to enable an adaptive combination of diverse FMGPR models. Suppose m_{sel} FMGPR models have been selected, the ensemble prediction output and variance for a test sample $\mathbf{x}_*$ can be calculated as:

$$\widehat{y}_*^{ens} = \sum_{i=1}^{m_{sel}} w_i \widehat{y}_*^{(i)}, m_{sel} \le m \tag{12}$$

$$\left(\widehat{\sigma}_*^{ens}\right)^2 = \sum_{i=1}^{m_{sel}} w_i \left(\left(\widehat{\sigma}_*^2\right)^{(i)} + \left(\widehat{y}_*^{(i)} - \widehat{y}_*^{ens}\right)^2 \right) \tag{13}$$

where $\widehat{y}_*^{(i)}$ and $(\widehat{\sigma}_*^2)^{(i)}$ are the prediction output and variance of the ith FMGPR model, respectively; and w_i denotes the mixture combination of diverse FMGPR models, which is usually determined

Algorithm 1: GA based ensemble pruning

INPUT:

$\left\{\widehat{\mathbf{y}}_{val}^{(1)}, \widehat{\mathbf{y}}_{val}^{(2)}, \ldots, \widehat{\mathbf{y}}_{val}^{(m)}\right\}$: the prediction outputs of m FMGPR models on validation set

$\left\{(\widehat{\boldsymbol{\sigma}}_{val}^2)^{(1)}, (\widehat{\boldsymbol{\sigma}}_{val}^2)^{(2)}, \ldots, (\widehat{\boldsymbol{\sigma}}_{val}^2)^{(m)}\right\}$: the prediction variances of m FMGPR models on validation set

N^{pop}: the population size of GA optimization

T: maximum generations

λ: tradeoff parameter for balancing the prediction accuracy and uncertainty

Objective function:

$$f(\mathbf{s}) = \lambda \mathrm{RMSE}_{sel} + (1-\lambda)\hat{\sigma}_{sel}$$

BEGIN:

1 Generate an initial population $P_{t=0}$ of size N^{pop} by randomly initializing each individual using a binary-coded chromosome;

2 **FOR** $t = 1$ to T **DO**

3 Evaluate each individual in the population P_t based on $f(\mathbf{s})$ by Eqs. (7)~(11);

4 Generate a new population through performing selection, crossover and mutation;

5 $t \leftarrow t + 1$;

6 **END FOR**

7 Choose the best individual from the finally obtained P_T in the last generation:

$$s^* \leftarrow \arg\min_{s \in P_T} f(\mathbf{s});$$

8 Decode the individual s^* to determine a small-sized set of FMGPR models $\left\{\mathrm{FMGPR}_{sel}^{(1)}, \mathrm{FMGPR}_{sel}^{(2)}, \ldots, \mathrm{FMGPR}_{sel}^{(m_{sel})}\right\}$ for ensemble prediction.

END

OUTPUT:

The selected FMGPR models: $\left\{\mathrm{FMGPR}_{sel}^{(1)}, \mathrm{FMGPR}_{sel}^{(2)}, \ldots, \mathrm{FMGPR}_{sel}^{(m_{sel})}\right\}$

according to practical application demands. Considering using the weights both in the selection of the base FMGPR models and in the combination of the component predictions is prone to suffering from overfitting [48], during the GA optimization process, the component predictions from FMGPR models are combined via simple averaging instead of weighted averaging, i.e.,

$$w_i = \frac{1}{m_{sel}}, i = 1, 2, \ldots, m_{sel} \tag{14}$$

During the online prediction stage, w_i is estimated as the posterior probability of $\mathbf{x}_*^{(i)}$ with respect to component FMGPR models according to Bayes' rule:

$$w_i = \frac{p\left(\mathbf{x}_*^{(i)}\mathrm{FMGPR}_{sel}^{(i)}\right)p\left(\mathrm{FMGPR}_{sel}^{(i)}\right)}{\sum_{k=1}^{m_{sel}} p\left(\mathbf{x}_*^{(k)}\middle|\mathrm{FMGPR}_{sel}^{(k)}\right)p\left(\mathrm{FMGPR}_{sel}^{(k)}\right)} \tag{15}$$

where $\mathbf{x}_*^{(i)}$ is the test input for the ith FMGPR model and $p(\mathbf{x}_*^{(i)}\middle|\mathrm{FMGPR}_{sel}^{(i)})$ is the conditional probability, and $p(\mathrm{FMGPR}_{sel}^{(i)})$ is the prior probability. Since the priorities of different FMGPR models are unknown in advance due to the lack of sufficient prior knowledge, we assume that $p(\mathrm{FMGPR}_{sel}^{(i)})$ of each model is equal for the sake of simplicity:

$$p\left(\mathrm{FMGPR}_{sel}^{(i)}\right) = \frac{1}{m_{sel}}, i = 1, 2, \ldots, m_{sel} \tag{16}$$

Subsequently, according to the component prediction variances, $p(\mathbf{x}_*^{(i)}\middle|\mathrm{FMGPR}_{sel}^{(i)})$ can be estimated as

$$p\left(\mathbf{x}_*^{(i)}\middle|\mathrm{FMGPR}_{sel}^{(i)}\right) = \exp\left(-\gamma \times \frac{\left(\widehat{\sigma}_*^2\right)^{(i)}}{\widehat{y}_*^{(i)}}\right) \tag{17}$$

where γ is a scaling parameter.

After the above step of GA based ensemble pruning, a set of most influential FMGPR models are selected, which are then integrated as a SEFMGPR model using a FMM based adaptive combination strategy.

3.4. Adaptation of SEFMGPR model

Although the wind power forecasting models can be built based on a large number of historical data, the already obtained predictive models may encounter performance degradation because of the time-varying changes of wind energy, e.g., climate change and seasonal factors. Therefore, it is necessary to endow the prediction model with adaptation capability, which is seldom paid attention to previous studies. In recent years, some research work attempts to retrain the predictive models during the online implementation stage. However, frequent model reconstruction will lead to heavy computational burden and in-depth knowledge is required for determining appropriate model structure, hyper-parameters, etc. To address this issue, we propose an incremental adaptation mechanism to allow online updating of the SEFMGPR model, which is achieved by updating local domains and GPR models of the selected FMGPR models.

In the proposed SEFMGPR wind power forecasting method, the final prediction results are obtained by a two-level adaptive ensemble approach. At the top-level ensemble, the final prediction results are obtained through adaptively combining the component predictions from diverse FMGPR models, while at the bottom-level ensemble, the predictions of local GPR models are integrated as the outputs of FMGPR models. The combination weights for the two levels of ensemble are determined adaptively through the FMM mechanism. In addition, the updating of local domains and GPR models are achieved by employing an incremental adaptation mechanism, which are described in detail in the following subsections.

3.4.1. Adaptation of local domains

During the offline modeling, the training input data at instant t are divided as local domains $\{\mathrm{LD}_{1,t}, \ldots, \mathrm{LD}_{k,t}, \ldots, \mathrm{LD}_{K,t}\}$ by performing GMM clustering. However, due to the time-varying behavior of wind power data, the probability distribution may change and thus online updating needs to be implemented. Suppose the number of components does not change during online operation, then the mean vectors, covariance matrices and mixture coefficients $\{\boldsymbol{\mu}_{k,t}, \sum_{k,t}, \pi_{k,t}\}_{k=1}^{K}$ at time instant t for the chosen LD are recursively updated as follows

$$\boldsymbol{\mu}_{k,t+1} = \boldsymbol{\mu}_{k,t} + \alpha\frac{p(\mathrm{LD}_{k,t}|\mathbf{x}_{t+1})}{\pi_{k,t}}(\mathbf{x}_{t+1} - \boldsymbol{\mu}_{k,t}) \tag{18}$$

$$\boldsymbol{\Sigma}_{k,t+1} = \boldsymbol{\Sigma}_{k,t} + \alpha\frac{p(\mathrm{LD}_k|\mathbf{x}_{t+1})}{\pi_{k,t}}\left((\mathbf{x}_{t+1} - \boldsymbol{\mu}_{k,t})(\mathbf{x}_{t+1} - \boldsymbol{\mu}_{k,t})^{\mathrm{T}} - \boldsymbol{\Sigma}_{k,t}\right) \tag{19}$$

$$\pi_{k,t+1} = \pi_{k,t} + \alpha\left(p(\mathrm{LD}_{k,t}|\mathbf{x}_{t+1}) - \pi_{k,t}\right) \tag{20}$$

where $\boldsymbol{\mu}_{k,t+1}, \sum_{k,t+1}, \pi_{k,t+1}$ are the mean vector, covariance matrix and mixture coefficient, respectively, which are updated based on the new wind power data $\mathbf{x}_{t+1}$; and α is a parameter for controlling the influence of the samples, which is usually set to $1/T$, where T is the number of updated samples [56].

3.4.2. Adaptation of local GPR models

With the changes in wind power data characteristics, the regression relationship between input and output data also shows time-varying behavior. Thus, the local GPR models built from historical LDs are updated using a moving window (MW) approach [57]. When a new sample arrives, the window moves forward to remove the oldest sample and add the new one. Since updating the whole local GPR models will significantly increase the computational cost, it is more reasonable to update the one with the highest relevance to the incoming wind power data. Then the main problem is shifted to finding out the most suitable local GPR model. With the help of posterior probabilities, the local GPR model corresponding to the LD with the maximum value of $p(\mathrm{LD}_{k,t+1}|\mathbf{x}_{t+1})$ is used as the target for the update.

During online operation at time instant t, the samples in the window with the size of L the kth GPR model can be denoted as

$$\mathrm{LD}_{k,t}=\left\{\mathbf{X}_{k,t},\mathbf{y}_{k,t}\right\}=\left\{\left(\mathbf{x}_{k,t,1},y_{k,t,1}\right),\cdots,\left(\mathbf{x}_{k,t,L},y_{k,t,L}\right)\right\} \tag{21}$$

with the covariance matrix calculated as

$$\mathbf{C}_{k,t}=\begin{bmatrix} C(\mathbf{x}_{k,t,1},\mathbf{x}_{k,t,1}) & \cdots & C(\mathbf{x}_{k,t,1},\mathbf{x}_{k,t,L}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{x}_{k,t,L},\mathbf{x}_{k,t,1}) & \cdots & C(\mathbf{x}_{k,t,L},\mathbf{x}_{k,t,L}) \end{bmatrix} \tag{22}$$

At time instant$t+1$, the window data are updated as

$$\begin{aligned}\mathrm{LD}_{k,t+1}&=\left\{\mathbf{X}_{k,t+1},\mathbf{y}_{k,t+1}\right\}\\&=\left\{\left(\mathbf{x}_{k,t,2},y_{k,t,2}\right),\cdots,\left(\mathbf{x}_{k,t,L},y_{k,t,L}\right),\left(\mathbf{x}_{t+1},y_{t+1}\right)\right\}\end{aligned} \tag{23}$$

Then the adaptation of the kth local GPR model can be achieved by simply updating the corresponding covariance matrix $\mathbf{C}_{k,t}$. As can be seen from Eq. (39), the inverse of $\mathbf{C}_{k,t}$, i.e., $[\mathbf{C}_{k,t}]^{-1}$ has been obtained at time instant$t$, thus it is appealing to obtain the updated inverse of covariance matrix $[\mathbf{C}_{k,t+1}]^{-1}$ from $[\mathbf{C}_{k,t}]^{-1}$ rather than calculating directly from $\mathbf{C}_{k,t}$ in order to reduce the computational cost [58]. To this end, two stages are required to obtain $[\mathbf{C}_{k,t+1}]^{-1}$.

At the first stage, a temporary covariance matrix $\tilde{\mathbf{C}}_{k,t}$ is obtained by removing the first row and column from $\mathbf{C}_{k,t}$, corresponding to removal of the oldest sample. If we rewrite $[\mathbf{C}_{k,t}]^{-1}$as

$$[\mathbf{C}_{k,t}]^{-1}=\begin{bmatrix} e & \mathbf{f}^{\mathrm{T}} \\ \mathbf{f} & \mathbf{G} \end{bmatrix} \tag{24}$$

where e is the first element of $[\mathbf{C}_{k,t}]^{-1}$ and $\mathbf{f}$ is a column vector with $(L-1)$ elements, then $[\tilde{\mathbf{C}}_{k,t}]^{-1}$ can be derived as

$$\tilde{\mathbf{C}}_{k,t}^{-1}=\mathbf{G}-\frac{\mathbf{f}\mathbf{f}^{\mathrm{T}}}{e} \tag{25}$$

Subsequently, at the second stage, a new sample is added to the model by adding a new row and column to the matrix $\tilde{\mathbf{C}}_{k,t}$:

$$\mathbf{C}_{k,t+1}=\begin{bmatrix} \tilde{\mathbf{C}}_{k} & \mathbf{b} \\ \mathbf{b}^{\mathrm{T}} & a \end{bmatrix} \tag{26}$$

where $\mathbf{b}=[C(\mathbf{x}_{t+1},\mathbf{x}_{k,t,2}),\cdots,C(\mathbf{x}_{t+1},\mathbf{x}_{k,t,L})]^{\mathrm{T}}$, and $a=C(\mathbf{x}_{t+1},\mathbf{x}_{t+1})$. To calculate the inverse of $\mathbf{C}_{k,t+1}$, we first express the target inverse matrix $[\mathbf{C}_{k,t+1}]^{-1}$ as

$$\left[\mathbf{C}_{k,t+1}\right]^{-1}=\begin{bmatrix} \mathbf{E} & \mathbf{h} \\ \mathbf{h}^{\mathrm{T}} & r \end{bmatrix} \tag{27}$$

where r is the last element of $[\mathbf{C}_{k,t+1}]^{-1}$ and $\mathbf{h}$ is a column vector with $(L-1)$ elements.

By taking Eqs. (26) and (27) into $\mathbf{C}_{k,t+1}[\mathbf{C}_{k,t+1}]^{-1}=\mathbf{I}_L$, we can obtain

$$\begin{cases} \tilde{\mathbf{C}}_{k,t}E+\mathbf{b}\mathbf{h}^{\mathrm{T}}=\mathbf{I}_{L-1} \\ \tilde{\mathbf{C}}_{k,t}h+br=0 \\ \mathbf{b}^{\mathrm{T}}+ar=1 \end{cases}$$

$$\Rightarrow\left[\mathbf{C}_{k,t+1}\right]^{-1}=\begin{bmatrix} \tilde{\mathbf{C}}_{k,t}^{-1}\left(\mathbf{I}+\mathbf{b}\mathbf{b}^{\mathrm{T}}\left(\tilde{\mathbf{C}}_{k,t}^{-1}\right)^{\mathrm{T}}g\right) & -\tilde{\mathbf{C}}_{k,t}^{-1}\mathbf{b}g \\ -\left(\tilde{\mathbf{C}}_{k,t}^{-1}\mathbf{b}\right)^{\mathrm{T}}g & g \end{bmatrix} \tag{28}$$

where $\tilde{\mathbf{C}}_{k,t}^{-1}$ is the inverse of $\tilde{\mathbf{C}}_{k,t}$, and g is calculated as follows:

$$g=\left(a-\mathbf{b}^{\mathrm{T}}\tilde{\mathbf{C}}_{k,t}^{-1}\mathbf{b}\right)^{-1} \tag{29}$$

To this end, the target GPR model from FMGPR is updated as

$$\begin{cases} \widehat{y}_{k,\mathrm{new}}=\mathbf{k}_{k,\mathrm{new}}^{\mathrm{T}}\left[\mathbf{C}_{k,t+1}\right]^{-1}\mathbf{y}_{k,t+1} \\ \sigma_{k,\mathrm{new}}^{2}=C(\mathbf{x}_{\mathrm{new}},\mathbf{x}_{\mathrm{new}})-\mathbf{k}_{k,\mathrm{new}}^{\mathrm{T}}\left[\mathbf{C}_{k,t+1}\right]^{-1}\mathbf{k}_{k,\mathrm{new}} \end{cases} \tag{30}$$

3.5. Implementation process of proposed model

Overall, the proposed SEFMGPR method consists of two stages: the offline modeling stage and the online prediction stage. In the offline modeling stage, a set of diverse FMGPR models is first built from various subspaces and then ensemble pruning is carried out to enhance the generalization performance and reduce the number of base learners. In the online prediction stage, the final ensemble prediction output and variance are obtained by performing two-level adaptive combination of component predictions. Moreover, the incremental updating of the SEFMGPR model is implemented when a new sample is available. The step-by-step procedure of the SEFMGPR approach for wind power forecasting is summarized below.

(i) Collect the historical wind speed, wind direction, wind power, air temperature, surface air pressure and density data as the modeling data, which are further divided into the training set D_{trn} and validation set D_{val}.
(ii) Generate a set of training subsets $\{\mathrm{D}_{tra}^{(1)},\ldots,\mathrm{D}_{tra}^{(m)}\}$ through bootstrapping resampling and construct subspaces $\{\mathrm{S}_1,\ldots,\mathrm{S}_m\}$by PLS regression from these data sets.
(iii) FMGPR models $\{\mathrm{FMGPR}^{(1)},\ldots,\mathrm{FMGPR}^{(m)}\}$ are constructed by GPR modeling and GMM clustering.
(iv) Select the most influential FMGPR models$\{\mathrm{FMGPR}_{sel}^{(1)},\mathrm{FMGPR}_{sel}^{(2)},\ldots,\mathrm{FMGPR}_{sel}^{(m_{sel})}\}$ to the ensemble by **Algorithm 1.**
(v) For a test sample $\mathbf{x}_*$, the final prediction output $\widehat{y}_*^{ens}$ and variance $(\widehat{\sigma}_*^{ens})^2$ are given through two-level adaptive ensemble.
(vi) When a new sample $(\mathbf{x}_{t+1},y_{t+1})$ comes, the SEFMGPR model is updated through an incremental manner.

4. Case study

In this section, the effectiveness and superiority of the proposed SEFMGPR method for wind power forecasting are verified through a real wind power dataset. First, the experiment regarding the selection of covariance functions of GPR models is presented. Then, the prediction performance of SEFMGPR is compared to that of different wind power forecasting models. Finally, the characteristics and advantages of SEFMGPR are analyzed. The methods for comparison are as follows:

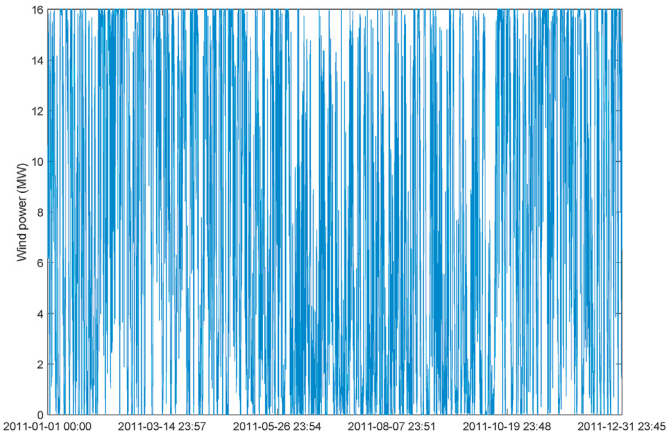

Fig. 5. Trend plot of real wind power time series.

(i) Persistence method: a classic method.
(ii) GPR: global GPR model.
(iii) EGPR: ensemble of GPR models built from subspaces.
(iv) EFMGPR: ensemble of finite mixture GPR models without ensemble pruning.
(v) $\text{SEFMGPR}_{\text{nonadapt}}$: selective ensemble of finite mixture GPR models without update, which is a nonadaptive variant of the proposed SEFMGPR.
(vi) SEFMGPR (the proposed method): selective ensemble of finite mixture GPR models with update.

To evaluate the prediction performance of wind power forecasting, root-mean-square error (RMSE) and coefficient of determination (R^2) are adopted:

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left(\widehat{y}_i - y_i\right)^2} \tag{31}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=1}^{n_{\text{test}}}\left(y_i - \overline{y}\right)^2} \tag{32}$$

where y_i and $\widehat{y}_i$ denote the actual and predicted outputs, respectively; $\overline{y}$ represents the mean value of outputs; and n_{test} is the number of testing samples.

The computer configurations for the experiments are as follows. CPU: Intel core i7-6700 (3.40 GHz), RAM: 8 GB, OS: Windows 10, Software: MATLAB R2018a.

4.1. Wind power data and pre-processing

This case study uses a real wind farm dataset from National Renewable Energy Laboratory (NREL) [59]. With a sampling interval of 5 min, 105,124 data points were collected in 2011 from a wind turbine with the installed capacity of 16 MW. The dataset consists of wind power, wind speed, wind direction, air temperature, surface air pressure, and density at hub height. In this paper, all of these features are considered for modeling. In addition, as a typical time series, wind power data are greatly affected by their historical states. Thus, in this work, a dynamic model structure [60] is used for wind power forecasting:

$$\widehat{\text{WP}}(t+h) = f\left(\begin{bmatrix} WP(t), WP(t-1), \ldots, WP(t-l), WS(t), WS(t-1), \ldots, WS(t-l), \\ WD(t), WD(t-1), \ldots, WD(t-l), Tem(t), Tem(t-1), \ldots, Tem(t-l), \\ Den(t), Den(t-1), \ldots, Den(t-l), Pre(t), Pre(t-1), \ldots, Pre(t-l) \end{bmatrix}\right) \tag{33}$$

where h is the ahead step, $\widehat{\text{WP}}$ is the predicted wind power, $f(\cdot)$ is the unknown function; $\text{WS}, \text{WD}, \text{WP}, \text{Tem}, \text{Pre}$ and Den denote wind speed, wind direction, wind power, air temperature, surface air pressure and density at hub height, respectively; and l is the number of time lags. In this paper, l is set to 8, and h is set to $\{1, 4, 8, 16\}$, which correspond to the forecast horizons of 15 min, 1 h, 2 h and 4 h, respectively. However, when considering all time-lagged variables, the model complexity of the resulting prediction model will grow significantly. Thus, some weakly relevant variables to the wind power are removed by using mutual information criterion (MI) (Appendix A).

Table 1
Prediction performance of GPR models with different covariance functions.

Forecasting horizon	Covariance function	RMSE	R^2
1-step ahead	Matérn with noise term	0.9118	0.9751
	Diagonal squared exponential with noise term	0.8796	0.9769
	Rational quadratic with noise term	0.9118	0.9751
	Linear with bias and noise terms	**0.8421**	**0.9788**
4-step ahead	Matérn with noise term	2.2028	0.8551
	Diagonal squared exponential with noise term	2.4613	0.8191
	Rational quadratic with noise term	2.1064	0.8675
	Linear with bias and noise terms	**1.9152**	**0.8905**
8-step ahead	Matérn with noise term	2.9806	0.7355
	Diagonal squared exponential with noise term	3.1346	0.7075
	Rational quadratic with noise term	2.8920	0.7510
	Linear with bias and noise terms	**2.6194**	**0.7957**
16-step ahead	Matérn with noise term	5.1649	0.2115
	Diagonal squared exponential with noise term	5.6915	0.0425
	Rational quadratic with noise term	4.9922	0.2633
	Linear with bias and noise terms	**3.4864**	**0.6407**

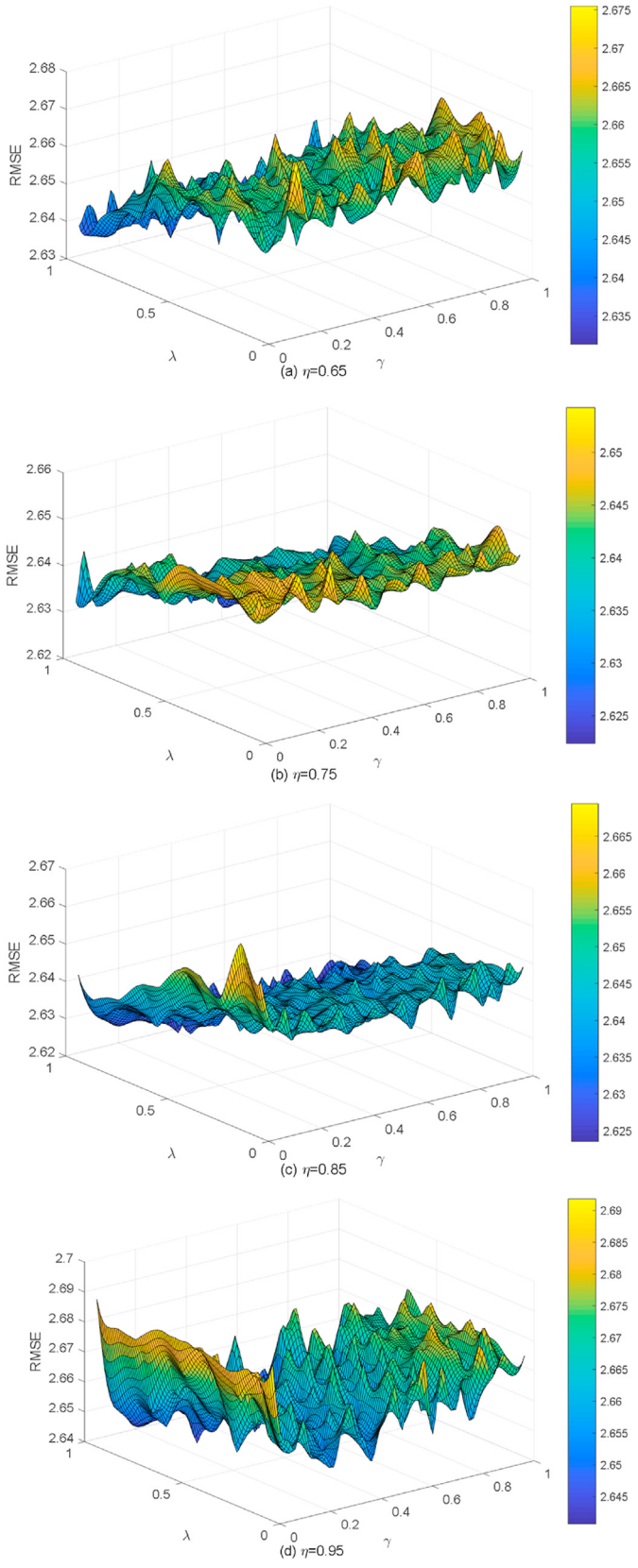

Fig. 6. Influence of model parameters $\{\eta, \lambda, \gamma\}$ on forecasting RMSE of SEFMGPR (validation set 2).

Table 2
Performance of different prediction models on various forecasting horizons (testing set 1).

Forecasting horizon	Method	RMSE	R^2
1-step ahead	Persistence	0.8462	0.9786
	GPR	0.8421	0.9788
	EGPR	0.8317	0.9793
	EFMGPR	0.7933	0.9812
	$\text{SEFMGPR}_{\text{nonadapt}}$	0.7901	0.9813
	SEFMGPR	**0.7771**	**0.9819**
4-step ahead	Persistence	1.9392	0.8880
	GPR	1.9152	0.8905
	EGPR	1.9097	0.8911
	EFMGPR	1.7844	0.9049
	$\text{SEFMGPR}_{\text{nonadapt}}$	1.7795	0.9055
	SEFMGPR	**1.7771**	**0.9057**
8-step ahead	Persistence	2.7061	0.7835
	GPR	2.6194	0.7957
	EGPR	2.6184	0.7959
	EFMGPR	2.4887	0.8156
	$\text{SEFMGPR}_{\text{nonadapt}}$	2.4821	0.8166
	SEFMGPR	**2.4443**	**0.8221**
16-step ahead	Persistence	3.7179	0.5945
	GPR	3.4864	0.6407
	EGPR	3.4802	0.6420
	EFMGPR	3.3566	0.6669
	$\text{SEFMGPR}_{\text{nonadapt}}$	3.3449	0.6693
	SEFMGPR	**3.3333**	**0.6716**

4.2. Prediction results and analysis

After performing the downsampling of the collected samples, the samples with the interval of 15min from year 2011 are used for the experiments, where the wind power series are shown in Fig. 5. The wind power data from January to February are randomly divided into three subsets: the training set with 3000 samples, the validation set 1 with 1332 samples, and validation set 2 with 1332 samples. Then, the data from March to December are used for model testing, where the data from March is defined as testing set 1 while the remaining testing data are defined as testing set 2. During the offline development and online implementation of SEFMGPR, the purposes of using different data sets are:

(i) Training set: used for constructing diverse subspaces and component GPR models;
(ii) Validation set 1: used for ensemble pruning of diverse FMGPR models;
(iii) Validation set 2: used for selecting suitable modeling parameters $\{\eta\lambda,\gamma\}$;
(iv) Testing set 1: used for near-term prediction performance evaluation, where the prediction time series are close to the modeling time series;
(v) Testing set 2: used for far-term prediction performance evaluation, where the prediction time series are far away from the modeling time series.

4.2.1. Experiment I: selection of covariance functions of GPR models

Selecting suitable covariance functions, which reveal the prior hypothesis of an unknown function, is of vital importance for developing high-performance GPR models. Therefore, four different covariance functions are evaluated for wind power prediction, including the Matérn covariance function with noise term, the diagonal squared exponential covariance function with noise

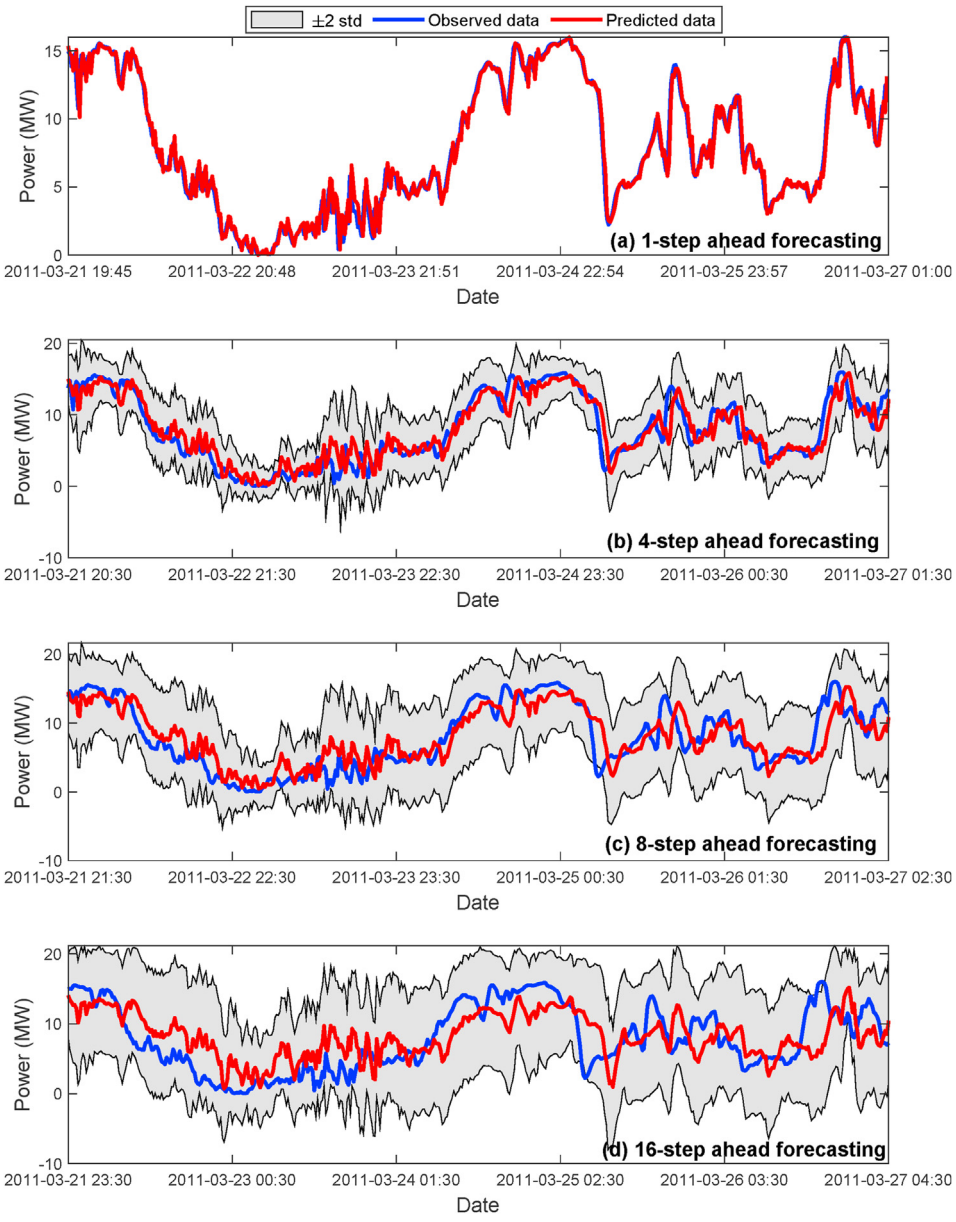

Fig. 7. Trend plots of wind power forecasting with different forecasting steps (testing set 1).

term, the rational quadratic covariance function with noise term, and the linear covariance function with bias and noise terms.

Table 1 shows the prediction performance of GPR models using different covariance functions on the testing set 1. As can be seen from the table, the linear covariance function performs best among the compared covariance functions for 1-step, 4-step, 8-step, and 16-step ahead forecasting. Obviously, the forecasting performance of the linear covariance function is significantly superior to the nonlinear ones for this case study. This may be mainly because the GPR models with nonlinear covariance functions are prone to encountering overfitting issues though they can obtain high training accuracy. Therefore, the linear covariance function is adopted in this work.

4.2.2. Experiment II: comparison to different wind power forecasting models

To build high-performance wind power forecasting models, the optimal model parameters are determined as ones that minimize the prediction errors on the validation set 2. The search ranges of η, λ and γ are $\{0.65, 0.75, 0.85, 0.95\}$, $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, \ldots, 1\}$, and $\{0.01, 0.04, 0.07, \ldots, 1\}$, respectively. Meanwhile, the maximum number of generations and the population size are set as $T = 100$ and $N^{pop} = 100$, respectively. As a result, the prediction RMSE values of SEFMGPR on validation set 2 using different combinations of model parameters are illustrated in Fig. 6. It is readily observed that the cutoff threshold η should be neither too small nor too large, a relatively large tradeoff parameter λ for balancing accuracy and uncertainty is preferable, and the scaling parameter γ exhibits a complicated relation with the SEFMGPR prediction performance.

Based on the above search settings, the optimal model parameters for different forecasting intervals are obtained as follows:

(i) 1-step forecasting: $\eta = 0.85$, $\lambda = 0.95$, $\gamma = 0.01$.
(ii) 4-step forecasting: $\eta = 0.85$, $\lambda = 0.95$, $\gamma = 1$.
(iii) 8-step forecasting: $\eta = 0.85$, $\lambda = 0.95$, $\gamma = 0.01$.

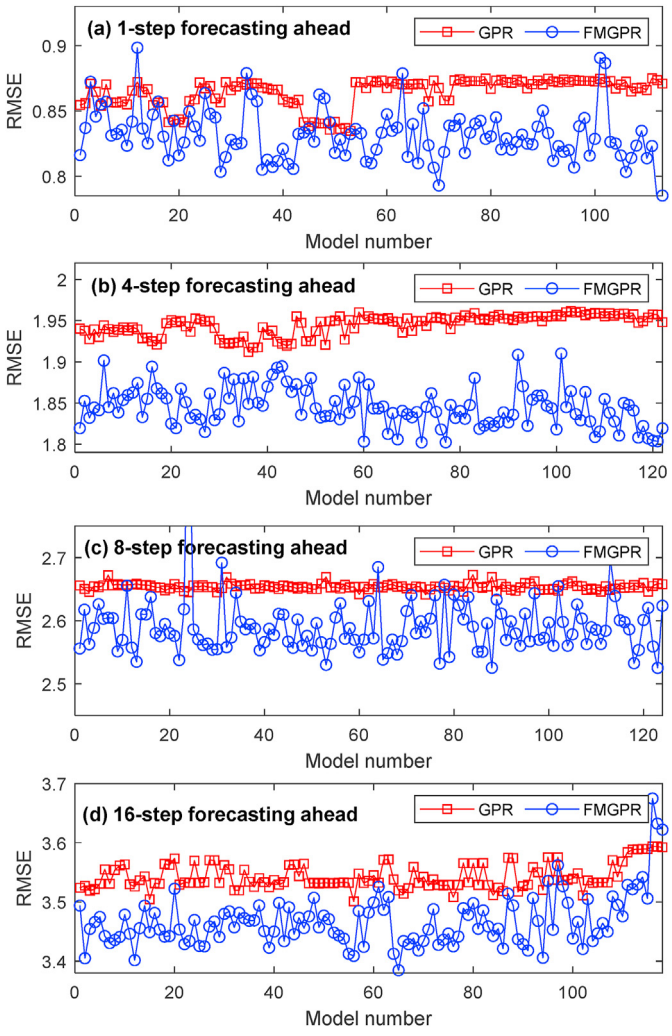

Fig. 8. Performance comparison between GPR and FMGPR models on different forecasting steps.

(iv) 16-step forecasting: $\eta = 0.85$, $\lambda = 0.95$, $\gamma = 0.1$.

The prediction performance of the proposed SEFMGPR approach on the testing set 1 is compared to that of the other 5 methods, as shown in Table 2. Notice that, to ensure the validity of the predicted wind power, the prediction outputs of different models should be kept within the rated power range of wind turbine, i.e. 0–16 MW in this case study.

The detailed comparisons listed in Table 2 can be summarized as follows:

(i) The proposed SEFMGPR method achieves the best performance on different forecasting horizons while the Persistence method performs worst. Compared to the Persistence method, in terms of prediction RMSE, SEFMGPR improves the prediction accuracy by 8.16%, 8.67%, 9.67%, and 10.34%, respectively. With the increase of prediction horizons, undoubtedly, the prediction accuracy gets worse.

(ii) Compared with the single GPR model, it is easy to observe that the ensemble models obtain significantly better prediction results. However, the performance improvement gained from EGPR is negligible compared to that from EFMGPR. This is because the sole perturbation used for EGPR is not enough for generating the diversity of base learners.

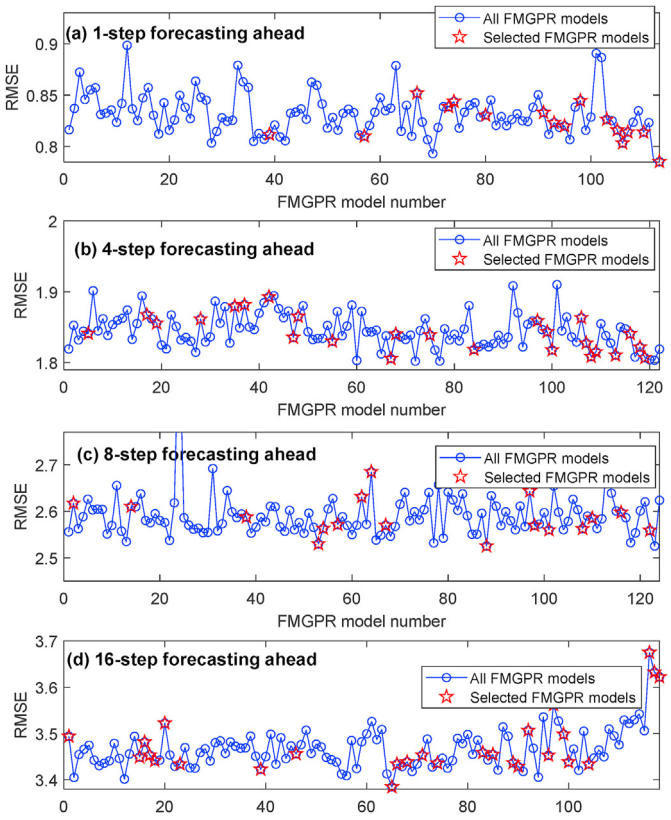

Fig. 9. Illustration of GA based ensemble member selection.

(iii) Comparing EFMGPR with $\text{SEFMGPR}_{\text{nonadapt}}$ indicates that the introduction of ensemble pruning enables the enhancement of model prediction performance as well as the sharp reduction of base FMGPR models.

(iv) Online updating of wind power forecasting models is indispensable for maintaining high model performance. As shown in Table 2, in comparison with the nonadaptive prediction model $\text{SEFMGPR}_{\text{nonadapt}}$, SEFMGPR provides improved prediction accuracy due to the introduction of model adaptation strategy.

Besides, the trend plots of wind power prediction on different prediction steps based on the SEFMGPR model are illustrated in Fig. 7, where the prediction confidence intervals are also provided. It can be observed that the predictions and actual values of wind power achieve good agreement, demonstrating the effectiveness of SEFMGPR in providing accurate estimations of wind power. In addition, we can also see that the confidence interval of predictions becomes larger with the decrease of the prediction accuracy, which implies that the ensemble prediction variance of SEFMGPR enables the operators to evaluate the prediction reliability.

4.2.3. Experiment III: characteristics and advantages of the proposed SEFMGPR method

Traditionally, ensemble models for wind power forecasting are built through perturbing on input attributes or training samples, which is insufficient in generating accurate and diverse base learners. Alternatively, we employ a multimodal perturbation mechanism for the SEFMGPR approach, which combines the perturbations of input attributes and training data through subspace construction and GMM clustering. As a result, a set of FMGPR models are required to build SEFMGPR. Meanwhile, a set of GPR

Table 3
Performance comparison between EFMGPR and SEFMGPR models.

Forecasting horizon	EFMGPR			SEFMGPR			P_1^*(%)	P_2^*(%)
	m	RMSE	R^2	m_{sel}	RMSE	R^2		
1-step	113	0.7933	0.9812	17	**0.7901**	**0.9813**	84.96	0.40
4-step	122	1.7844	0.9049	25	**1.7771**	**0.9057**	79.51	0.41
8-step	124	2.4887	0.8156	17	**2.4821**	**0.8166**	86.29	0.26
16-step	118	3.3566	0.6669	27	**3.3449**	**0.6716**	77.11	0.34

* P1.. x 100%, P2=...x 100%

Table 4
The 4-step ahead prediction performance of Persistence and SEFMGPR under different ramp states.

Duration		Ramp state	Persistence		SEFMGPR	
Start time	End time		RMSE	R^2	RMSE	R^2
2011-03-03 14:15	2011-03-04 02:45	Up-Ramp	2.1967	0.8400	1.9971	0.8678
2011-03-08 14:15	2011-03-09 05:15	Up-Ramp	1.5105	0.9247	1.0977	0.9602
2011-03-02 18:45	2011-03-03 06:45	Down-Ramp	2.5031	0.7898	2.2118	0.8359
2011-03-29 17:30	2011-03-30 02:45	Down-Ramp	2.1171	0.8224	1.7229	0.8824

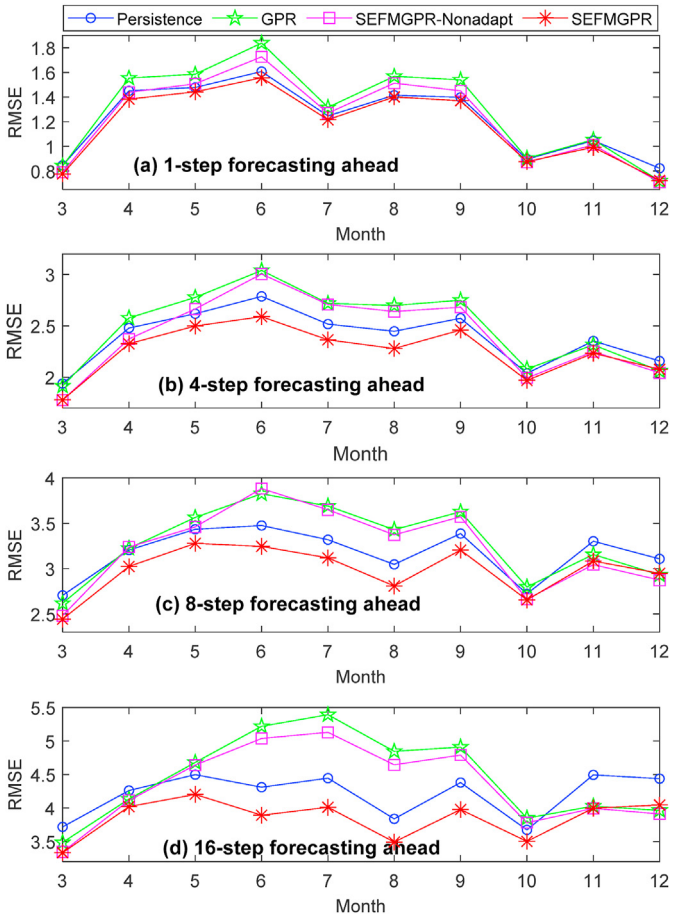

Fig. 10. Prediction RMSE values of four forecasting methods for different forecasting horizons in 2011 (testing sets 1 and 2).

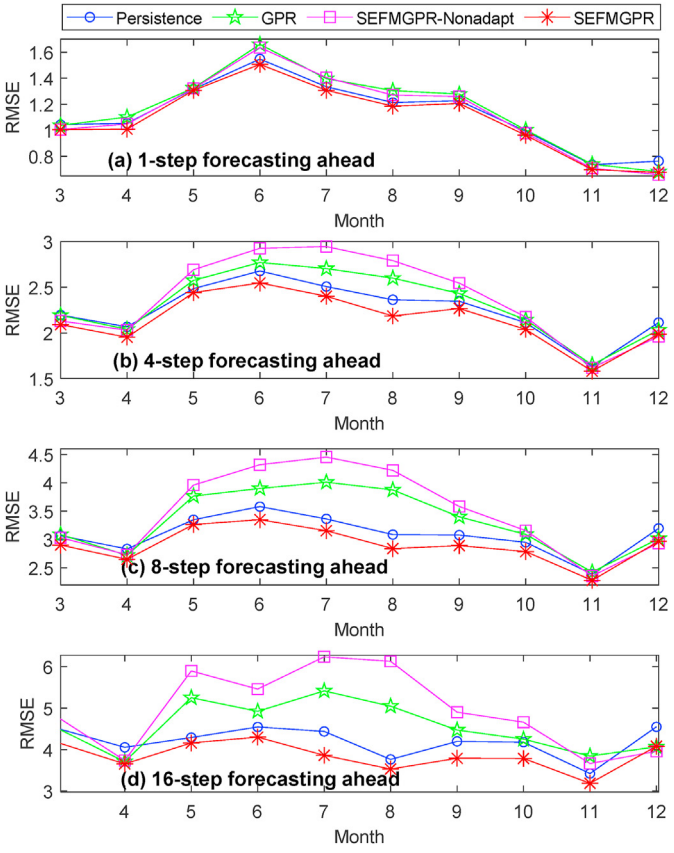

Fig. 11. Prediction RMSE values of four forecasting methods for different forecasting horizons in 2012.

models are built from diverse subspaces for comparison, corresponding to the case of using single-modal perturbation.

Fig. 8 shows the prediction RMSE values of base GPR models and FMGPR models for constructing EGPR and EFMGPR models. Noticeably, no matter what the prediction horizon is, FMGPR models have much better performance than GPR models. In detail, the average prediction RMSE values from FMGPR models are improved by 3.80%, 5.13%, 2.35% and 2.13%, respectively, in comparison with GPR models. In terms of model diversity, it is easy to find that FMGPR models are much more diverse than GPR models. These experiments reveal that the resulting base FMGPR models are more accurate and diverse than GPR models, which is the key to building high-performance EFMGPR and SEFMGPR models.

Subsequently, the advantage of the proposed GA based ensemble pruning is verified. The selection results of FMGPR

models for different forecasting horizons are depicted in Fig. 9. As can be seen, only a small part of base FMGPR models are selected for the ensemble, though a large number of base FMGPR models have been built. Meanwhile, we can see that some of the selected models perform worse than some of the unselected ones. A benefit of such results is to guarantee the diversity of base FMGPR models.

To further investigate the advantages of ensemble pruning, Table 3 compares the prediction accuracy and ensemble member sizes of EFMGPR and SEFMGPR models on different forecasting steps. It is readily seen that slightly superior results are achieved with SEFMGPR after performing ensemble pruning while the ensemble member sizes are reduced significantly. For example, in the case of the 8-step ahead forecasting, only 17 FMGPR models are selected from 124 FMGPR models using the GA based ensemble pruning. Compared with EFMGPR, the number of base models of SEFMGPR has been reduced by 86.29%, while producing comparable forecasting performance. These results demonstrate the effectiveness and superiority of the proposed ensemble pruning method.

In addition, we are also interested about the forecasting performance of the proposed method when encountering the extreme ramp events. Thus, taking the 4-step ahead forecasting as an example, the prediction accuracy of the Persistence and SEFMGPR methods under different ramp states are tabulated in Table 4. It can be seen that SEFMGPR can provide much better prediction accuracy than Persistence method, which implies the superior capability of SEFMGPR in handling the ramp events and delivering high-performance predictions.

4.2.4. Experiment IV: prediction performance across different months and years for different wind power forecasting methods

Though Table 2 has confirmed the effectiveness of the proposed SEFMGPR approach by the online prediction results in March, it is also desirable to investigate the prediction performance of SEFMGPR across different months during one year. This is mainly because the data characteristics may suffer from great changes along with the online implementation of wind power prediction, which may cause severe model performance degradation. Fortunately, the proposed SEFMGPR approach, which is equipped with adaptation mechanism to enable automatic model updating, has the potential of dealing with time-varying behavior of wind power generation.

Fig. 10 depicts the prediction RMSE values from March to December in 2011 for different forecasting methods. Not surprisingly, the nonadaptive prediction models, i.e., GPR, and SEFMGPR-$_{\text{nonadapt}}$, start encountering significant performance deterioration in April. In comparison, due to the continuous inclusion of newly measured wind power data, the two adaptive prediction methods, i.e., Persistence and SEFMGPR, can effectively alleviate the prediction accuracy reduction, especially from May to September. Overall, SEFMGPR achieves much better prediction accuracy than other forecasting methods, which indicates that the combination of adaptive ensemble and incremental adaptation mechanism is very helpful to maintain high prediction performance.

To further verify the usefulness and superiority of SEFMGPR, the wind power data in 2012 are used for model training, validation and testing via the same partition manner as the modeling data in 2011. As we expect, similar conclusions can be drawn according to the prediction results in Fig. 11, where SEFMGPR performs best among the compared methods.

The above prediction results show that the proposed SEFMGPR approach is capable of delivering accurate estimations of wind power, providing the possibility of evaluating the prediction reliability using the ensemble prediction variance, and accommodating the time-varying characteristics of wind power data. Also, the superiority of SEFMGPR over traditional wind power forecasting methods is confirmed.

5. Conclusions

In this paper, a novel ensemble method SEFMGPR is proposed for probabilistic wind power forecasting, which is designed based on the selective ensemble of GPR models. First, a set of diverse and accurate local GPR models are built through a multimodal perturbation mechanism, i.e., perturbing the input attributes and training data. Then, a GA based ensemble pruning is performed to select the ensemble members and a two-level adaptive combination is achieved through the FMM mechanism. When a new sample is available, the SEFMGPR model is updated incrementally. The experimental results indicate that the SEFMGPR method is superior to traditional wind power forecasting methods in providing accurate estimations, prediction confidence intervals, as well as accommodating the time-varying characteristics of wind energy.

Though the experimental results have shown that the proposed SEFMGPR wind power forecasting approach outperforms the traditional Persistence and nonadaptive prediction methods, more efforts on the following issues are encouraged in order to further enhance the prediction performance. First, SEFMGPR only consider the perturbations on input attributes and training data for building diverse and accurate base learners, however, the perturbations on learning algorithms and parameters may also be helpful. Second, the influences of ramp events on the model prediction accuracy are ignored in this study, which may be vital for obtaining reliable wind power predictions. Third, the effective handling of ramp events may be crucial to improve the model adaptation capability. Finally, exploring the ensemble pruning through multi-objective optimization is also an interesting issue.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 61763020, 61863020), and Yunnan Fundamental Research Projects (grant number 2018FD040).

Appendix A

A. 1. Gaussian process regression

A Gaussian process is defined as a collection of random variables, any finite number of which has a joint Gaussian distribution. Given a data set $D = \{\mathbf{X}, \mathbf{y}\} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where $\mathbf{x} \in \mathbb{R}^{1\times d}$ and $y \in \mathbb{R}^{1\times 1}$ denote d-dimensional input and single output, respectively, then the relationship between input and output can be described as

$$y = f(\mathbf{x}) + \varepsilon \tag{34}$$

where ε is the Gaussian noise with zero mean and variance σ_n^2, and $f(\mathrm{n})$ represents the unknown function. From the function-space view, a Gaussian process can be written as

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), C(\mathbf{x}, \mathbf{x}')) \tag{35}$$

where the mean function $m(\mathbf{x})$ and covariance function $C(\mathbf{x}, \mathbf{x}')$ are defined as follows:

$$\begin{cases} m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \\ C(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \end{cases} \tag{36}$$

Consider a Gaussian process with zero mean, the output variable follows a Gaussian distribution

$$\mathbf{y} \sim \mathcal{N}(0, \mathbf{C}) \tag{37}$$

where $\mathbf{C}$ is an $n \times n$ covariance matrix with $C_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$.

For a new input vector $\mathbf{x}_*$, the joint distribution of the training outputs $\mathbf{y}$ and the new output y_* is also Gaussian:

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{C} & \mathbf{k}_* \\ \mathbf{k}_*^{\mathrm{T}} & C(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{38}$$

where $\mathbf{k}_* = [C(\mathbf{x}_*, \mathbf{x}_1), \cdots, C(\mathbf{x}_*, \mathbf{x}_n)]^{\mathrm{T}}$. Then, the prediction mean $\hat{y}_*$ and variance σ_*^2 are calculated as

$$\begin{cases} \hat{y}_* = \mathbf{k}_*^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{y} \\ \sigma_*^2 = C(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{k}_* \end{cases} \tag{39}$$

As can be seen from Eq. (39), the definition of covariance functions has a significant influence on GPR models. Under the view of Gaussian process, it is the covariance function that defines nearness or similarity between samples. In this paper, the following covariance functions are considered for wind power forecasting, which are defined as follows:

(1) Matérn covariance function with noise term:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \left(1 + \frac{\sqrt{3}\mathbf{x}_i - \mathbf{x}_j}{l}\right) \exp\left(-\frac{\sqrt{3}\mathbf{x}_i - \mathbf{x}_j}{l}\right) + \sigma_n^2 \delta_{ij} \tag{40}$$

(2) Diagonal squared exponential covariance function with noise term:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}\mathbf{x}_i - \mathbf{x}_j\right) + \sigma_n^2 \delta_{ij} \tag{41}$$

(3) Rational quadratic covariance function with noise term:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \left(1 + \frac{1}{2\alpha l^2}\mathbf{x}_i - \mathbf{x}_j\right)^{-\alpha} + \sigma_n^2 \delta_{ij} \tag{42}$$

(4) Linear covariance function with bias and noise terms:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \frac{(\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j + 1)}{l^2} + \sigma_n^2 \delta_{ij} \tag{43}$$

The hyper-parameters $\Theta = \{\sigma_f^2, l, \alpha, \sigma_n^2\}$ can be estimated by maximizing the following log-likelihood function:

$$L = -\frac{1}{2}\log\det\mathbf{C} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} - \frac{n}{2}\log 2\pi \tag{44}$$

A. 2. Gaussian mixture models

Suppose a data set $\mathbf{x} \in \mathbb{R}^{n \times d}$ is drawn from a mixture of K component Gaussian distribution:

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mathbf{\Theta}_k) \tag{45}$$

Let $\mathbf{\Theta} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sum_1, \ldots, \sum_K, \ldots, \pi_1, \ldots, \pi_K\}$ be the parameters of the Gaussian mixture models, where $\boldsymbol{\mu}_k$ is the mean vector, $\sum_k$ is the covariance matrix of the kth Gaussian component, and π_k is the mixing coefficient satisfying

$$0 \le \pi_k \le 1, \sum_{k=1}^{K} \pi_k = 1 \tag{46}$$

A multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \sum_k)$ is completely specified by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\sum_k$, with the probability density function expressed as

$$p(\mathbf{x}|\mathbf{\Theta}_k) = \frac{1}{\sqrt{(2\pi)^d \left|\sum_k\right|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right] \tag{47}$$

The hyper-parameters of GMM model can be estimated from the modified expectation-maximization algorithm (EM) [61]. According to Bayes' rule, the posterior probability of $\mathbf{x}$ with respect to the kth component is calculated as

$$p(\mathbf{\Theta}_k|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)} \tag{48}$$

A. 3. Partial least squares regression

For a data set $\{\mathbf{X}, \mathbf{Y}\}$ with the input $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the output $\mathbf{Y} \in \mathbb{R}^{n \times q}$, the goal of PLS modeling is to project the scaled and mean-centered input and output data to latent variables

$$\mathbf{X} = \sum_{i=1}^{r} \mathbf{t}_i \mathbf{p}_i^{\mathbf{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{49}$$

$$\mathbf{Y} = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{q}_i^{\mathbf{T}} + \mathbf{F} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{50}$$

where $\mathbf{T} \in \mathbb{R}^{n \times r}$ and $\mathbf{U} \in \mathbb{R}^{n \times r}$ are score matrices and $r \le d$ is the number of the latent variables, $\mathbf{P} \in \mathbb{R}^{d \times r}$ and $\mathbf{Q} \in \mathbb{R}^{q \times r}$ are load matrices, and $\mathbf{E}$ and $\mathbf{F}$ are the residual matrices of input and output data, respectively. As a result, the PLS regression model is given as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{R} \tag{51}$$

where $\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{U}$ is the input weight matrix, $\boldsymbol{\beta} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})\mathbf{Q}^{\mathrm{T}}$ and $\mathbf{R}$ is the residual matrix.

A. 4. Mutual information

Mutual information (MI) [62] is a nonparametric and nonlinear method to evaluate relevance from information theory. For two random variables X and Y, the MI between X and Y can be calculated by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X;Y) \quad (52)$$

where $H(\mathrm{n})$ represents entropy, $H(X|Y), H(Y|X)$ are conditional entropies, respectively, and $H(X;Y)$ is the joint entropy of X and Y, which are calculated as

$$H(X) = -\int p_X(x)\log p_X(X)dx \quad (53)$$

$$H(Y) = -\int p_Y(y)\log p_Y(y)dy \quad (54)$$

$$H(X;Y) = -\iint p_{X,Y}(x,y)\log p_{X,Y}(x,y)dxdy \quad (55)$$

where $p_{X,Y}(x,y)$ is the joint probability density function and $p_X(x)$, $p_Y(y)$ can be denoted as

$$p_X(x) = \int p_{X,Y}(x,y)dy \quad (56)$$

$$p_Y(y) = \int p_{X,Y}(x,y)dx \quad (57)$$

By substituting Eqs. 53–55 into Eq. (52), the MI values can be calculated as

$$I(X;Y) = \iint p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}dxdy \quad (58)$$

In practical applications, the integration is substituted by summation overall all available discrete samples. Then, the MI values can be computed by using k-nearest neighbor statistics to estimate the entropies [63]. The basic idea of this approach is to estimate the entropy based on an average distance to the k-nearest neighbors.

Given a training set $\mathrm{D}_{trn} = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X} \in \mathbb{R}^{n\times d}$, the input variable selection using MI is as follows:

(i) Set a suitable cutoff threshold ζ for selecting the relevant variables.
(ii) Calculate the mutual information value between $\mathbf{x}_i$ and $\mathbf{y}$, where $\mathbf{x}_i$ denotes the vector of the ith input variables.
(iii) Repeat the step (ii) for d times and get a set of MI values.
(iiii) Select the most relevant input variables whose MI values exceed the cutoff threshold.

References

[1] P.S. Georgilakis, Technical challenges associated with the integration of wind power into power systems, Renew. Sustain. Energy Rev. 12 (3) (2008) 852–863.
[2] I. Colak, S. Sagiroglu, M. Yesilbudak, Data mining and wind power prediction: a literature review, Renew. Energy 46 (2012) 241–247.
[3] J. Yan, Y. Liu, S. Han, Y. Wang, S. Feng, Reviews on uncertainty analysis of wind power forecasting, Renew. Sustain. Energy Rev. 52 (2015) 1322–1330.
[4] Q. Xu, D. He, N. Zhang, C. Kang, Q. Xia, J. Bai, J. Huang, A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining, IEEE Transactions on sustainable energy 6 (4) (2015) 1283–1291.
[5] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.
[6] T. Hill, L. Marquez, M. O'Connor, W. Remus, Artificial neural network models for forecasting and decision making, Int. J. Forecast. 10 (1) (1994) 5–15.
[7] G. Welch, G. Bishop, An Introduction to the Kalman Filter, Citeseer, 1995.
[8] D. Liu, D. Niu, H. Wang, L. Fan, Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm, Renew. Energy 62 (2014) 592–597.
[9] T. Mahmoud, Z. Dong, J. Ma, An advanced approach for optimal wind power generation prediction intervals by using self-adaptive evolutionary extreme learning machine, Renew. Energy 126 (2018) 254–269.
[10] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[11] H. Wang, G. Wang, G. Li, J. Peng, Y. Liu, Deep belief network based deterministic and probabilistic wind speed forecasting approach, Appl. Energy 182 (2016) 80–93.
[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[13] Z. Ma, H. Chen, J. Wang, X. Yang, R. Yan, J. Jia, W. Xu, Application of hybrid model based on double decomposition, error correction and deep learning in short-term wind speed prediction, Energy Convers. Manag. 205 (2020), 112345.
[14] H. Liu, C. Chen, X. Lv, X. Wu, M. Liu, Deterministic wind energy forecasting: a review of intelligent predictors and auxiliary methods, Energy Convers. Manag. 195 (2019) 328–345.
[15] A. Tascikaraoglu, M. Uzunoglu, A review of combined approaches for prediction of short-term wind speed and power, Renew. Sustain. Energy Rev. 34 (2014) 243–254.
[16] Y. Ren, P. Suganthan, N. Srikanth, Ensemble methods for wind and solar power forecasting—a state-of-the-art review, Renew. Sustain. Energy Rev. 50 (2015) 82–91.
[17] Y. Liu, C. Yang, Z. Gao, Y. Yao, Ensemble deep kernel learning with application to quality prediction in industrial polymerization processes, Chemometr. Intell. Lab. Syst. 174 (2018) 15–21.
[18] Z. Ge, Z. Song, Performance-driven ensemble learning ICA model for improved non-Gaussian process monitoring, Chemometr. Intell. Lab. Syst. 123 (2013) 1–8.
[19] L. Xiao, Y. Dong, Y. Dong, An improved combination approach based on Adaboost algorithm for wind speed time series forecasting, Energy Convers. Manag. 160 (2018) 273–288.
[20] T. Peng, J. Zhou, C. Zhang, Y. Zheng, Multi-step ahead wind speed forecasting using a hybrid model based on two-stage decomposition technique and AdaBoost-extreme learning machine, Energy Convers. Manag. 153 (2017) 589–602.
[21] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, M.U. Rehman, A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting, IEEE Access 7 (2019) 28309–28318.
[22] Y. Li, H. Shi, F. Han, Z. Duan, H. Liu, Smart wind speed forecasting approach using various boosting algorithms, big multi-step forecasting strategy, Renew. Energy 135 (2019) 540–553.
[23] Z. Qian, Y. Pei, H. Zareipour, N. Chen, A review and discussion of decomposition-based hybrid models for wind energy forecasting applications, Appl. Energy 235 (2019) 939–953.
[24] W. Sun, M. Liu, Wind speed forecasting using FEEMD echo state networks with RELM in Hebei, China, Energy Convers. Manag. 114 (2016) 197–208.
[25] H. Liu, X.-w. Mi, Y.-f. Li, Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network, Energy Convers. Manag. 156 (2018) 498–514.
[26] D. Wang, H. Luo, O. Grunder, Y. Lin, Multi-step ahead wind speed forecasting using an improved wavelet neural network combining variational mode decomposition and phase space reconstruction, Renew. Energy 113 (2017) 1345–1358.
[27] Y. Hao, C. Tian, A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting, Appl. Energy 238 (2019) 368–383.
[28] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artif. Intell. 137 (1–2) (2002) 239–263.
[29] Z.-H. Zhou, Y. Yu, Ensembling local learners through Multimodal perturbation, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 35 (4) (2005) 725–735.
[30] A. Lahouar, J.B.H. Slama, Hour-ahead wind power forecast based on random forests, Renew. Energy 109 (2017) 529–541.
[31] J. Chen, G.-Q. Zeng, W. Zhou, W. Du, K.-D. Lu, Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization, Energy Convers. Manag. 165 (2018) 681–695.
[32] Z. Qu, K. Zhang, W. Mao, J. Wang, C. Liu, W. Zhang, Research and application of ensemble forecasting based on a novel multi-objective optimization algorithm for wind-speed forecasting, Energy Convers. Manag. 154 (2017) 440–454.
[33] T. Liu, S. Chen, S. Liang, C.J. Harris, Selective ensemble of multiple local model learning for nonlinear and nonstationary systems, Neurocomputing 378 (2020) 98–111.
[34] H. Jin, S. Huang, L. Wang, X. Chen, B. Pan, J. Li, Selective ensemble learning based on evolutionary multi-objective optimization for soft sensor development, J. Chem. Eng. Chin. Univ. 33 (3) (2019) 680–691.
[35] Z. Yang, J. Wang, A combination forecasting approach applied in multistep wind speed forecasting based on a data processing strategy and an optimized artificial intelligence algorithm, Appl. Energy 230 (2018) 1108–1125.
[36] J. Song, J. Wang, H. Lu, A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting, Appl. Energy 215 (2018) 643–658.
[37] W. Zhang, Z. Qu, K. Zhang, W. Mao, Y. Ma, X. Fan, A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting, Energy Convers. Manag. 136 (2017) 439–451.
[38] H. Liu, Z. Duan, Y. Li, H. Lu, A novel ensemble model of different mother

wavelets for wind speed multi-step forecasting, Appl. Energy 228 (2018) 1783–1800.
[39] P. Jiang, C. Li, Research and application of an innovative combined model based on a modified optimization algorithm for wind speed forecasting, Measurement 124 (2018) 395–412.
[40] L. Cheng, H. Zang, T. Ding, R. Sun, M. Wang, Z. Wei, G. Sun, Ensemble recurrent neural network based probabilistic wind speed forecasting approach, Energies 11 (8) (2018) 1958.
[41] T. Ouyang, H. Huang, Y. He, Z. Tang, Chaotic wind power time series prediction via switching data-driven modes, Renew. Energy 145 (2020) 270–281.
[42] Y. Zhang, J. Wang, X. Wang, Review on probabilistic forecasting of wind power generation, Renew. Sustain. Energy Rev. 32 (2014) 255–270.
[43] H. Jin, J. Li, M. Wang, B. Qian, B. Yang, Z. Li, L. Shi, Ensemble just-in-time learning-based soft sensor for mooney viscosity prediction in an industrial rubber mixing process, Adv. Polym. Technol. (2020) 2020.
[44] S. Zhu, X. Yuan, Z. Xu, X. Luo, H. Zhang, Gaussian mixture model coupled recurrent neural networks for wind speed interval forecast, Energy Convers. Manag. 198 (2019), 111772.
[45] L. Wang, H. Jin, X. Chen, J. Dai, K. Yang, D. Zhang, Soft sensor development based on the hierarchical ensemble of Gaussian process regression models for nonlinear and non-Gaussian chemical processes, Ind. Eng. Chem. Res. 55 (28) (2016) 7704–7719.
[46] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC press, Boca Raton, 2012.
[47] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Inf. Fusion 6 (1) (2005) 5–20.
[48] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, Adv. Neural Inf. Process. Syst. (1995) 231–238.
[49] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1) (1992) 1–58.
[50] M. Sun, C. Feng, J. Zhang, Multi-distribution ensemble probabilistic wind power forecasting, Renew. Energy 148 (2020) 135–149.
[51] Z. Wu, X. Xia, L. Xiao, Y. Liu, Combined model with secondary decomposition-model selection and sample selection for multi-step wind power forecasting, Appl. Energy 261 (2020), 114345.
[52] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[53] X. Xie, H. Shi, Dynamic multimode process modeling and monitoring using adaptive Gaussian mixture models, Ind. Eng. Chem. Res. 51 (15) (2012) 5497–5505.
[54] B. Muthén, K. Shedden, Finite mixture modeling with mixture outcomes using the EM algorithm, Biometrics 55 (2) (1999) 463–469.
[55] T. Chen, J. Ren, Bagging for Gaussian process regression, Neurocomputing 72 (7–9) (2009) 1605–1610.
[56] Z. Zivkovic, F. van der Heijden, Recursive unsupervised learning of finite mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 26 (5) (2004) 651–656.
[57] R. Grbić, D. Slišković, P. Kadlec, Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models, Comput. Chem. Eng. 58 (2013) 84–97.
[58] S. Van Vaerenbergh, J. Via, I. Santamaria, Nonlinear system identification using a new sliding-window kernel RLS algorithm, JCM 2 (3) (2007) 1–8.
[59] C. Draxl, A. Clifton, B.-M. Hodge, J. McCaa, The wind integration national dataset (wind) toolkit, Appl. Energy 151 (2015) 355–366.
[60] J. Wang, Y. Song, F. Liu, R. Hou, Analysis and application of forecasting models in wind power integration: a review of multi-step-ahead wind speed forecasting models, Renew. Sustain. Energy Rev. 60 (2016) 960–981.
[61] M.A.T. Figueiredo, A.K. Jain, Unsupervised learning of finite mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 381–396.
[62] F. Amiri, M.R. Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, J. Netw. Comput. Appl. 34 (4) (2011) 1184–1199.
[63] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. 69 (6) (2004), 066138.