Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

An ensemble learning model based on Bayesian model combination for solar energy prediction

Cite as: J. Renewable Sustainable Energy 11, 043702 (2019); doi: 10.1063/1.5094534
Submitted: 3 March 2019 · Accepted: 5 July 2019 ·
Published Online: 23 July 2019

Jian-Fang Chang,[1,a)] Na Dong,[1,b)] Wai Hung Ip,[2] and Kai Leung Yung[2]

AFFILIATIONS

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2]Department of Industrial and Systems Engineering of Hong Kong Polytechnic University, Hong Kong, China

[a)]**Present address:** School of Electrical and Information Engineering, Tianjin University.
[b)]**Author to whom correspondence should be addressed:** dongna@tju.edu.cn

ABSTRACT

To improve the reliability of solar irradiance prediction methods, an ensemble learning method based on the Bayesian model combination has been developed in this paper for solar utilization systems. First, a novel data sampling method has been proposed, including the advantages of clustering and cross validation, which can effectively ensure that the training subsets are different from each other and can cover a variety of different meteorological samples. Second, an ensemble learning model with multiple base learners has been designed. Each training subset is utilized to train the corresponding base learner. Then, a novel Bayesian model combination strategy expands hypothesis space E on Bayesian model averaging, which is applied to frame the combination strategy based on the accuracy of each base learner on the validation set. The prediction values of multiple learners are framed through the model combination strategy. Thus, a novel ensemble learning model based on Bayesian model combination has been established. Finally, experiments are carried out and the proposed method is compared with the Artificial Neural Network (ANN), K-means (Radial Basis Function) RBF, Support Vector Machine (SVM), and Multikernel SVM. The annual average mean absolute error of the ensemble learning method based on Bayesian model combination is reduced by 0.0374 MJ $\times$ m^{-2} compared with the ensemble learning method. The annual average mean absolute error of the proposed method is reduced by 42.6%, 38.2%, 52%, and 48.7%, respectively, compared with ANN, K-means RBF, SVM, and Multikernel SVM. The effectiveness as well as the reliability of the proposed method in solar energy prediction have been found to perform better and have verified our approach.

I. INTRODUCTION

As a kind of clean, substantial and renewable energy, solar energy does not produce pollutants,[1,2] which are vital features that other kinds of energies do not possess. Therefore, solar energy is one of the renewable energy sources and has been widely studied and applied.

Two issues that need to be addressed in the design and utilization of the solar energy system are the stability of individual energy producers and the creation of viable grid-connected systems which can reasonably manage and schedule individual energy producers. A particularly relevant aspect is the creation of a system that brings together many unstable individual producers to form a more stable energy network system.[3,4] Due to factors such as solar elevation angle, temperature, humidity, location, altitude, and other climatic factors, the energy generated by individual producers is mostly unstable,[5] and thus, applications of solar energy are often restricted. Despite the rapid development of smart grid systems, the stability of individual generators is critical to energy conservation and rational utilization, as well as grid security.[6] Therefore, solar irradiance prediction is of vital importance to the stable operation of the entire grid system and the formulation of energy dispatching plans.

Short-term predictions of solar energy are extremely critical.[1] Support Vector Machine (SVM)[7–10] and Artificial Neural Network (ANN)[11–13] have mainly been applied to the prediction of solar power, which are prone to train the prediction model through accuracy.[14] However, the reliability of prediction results is more significant in applications.[15] Ensemble learning (EL)[16] provides inspiration for improving the reliability of the prediction results. Ensemble learning combines multiple base learners together to achieve a better generalized performance and reliability than a single learner.

Nourani *et al.*[17] proposed an ensemble learning for multiregion daily global solar radiation estimation for Iraq. To ensure an appropriate selection of input variables, sensitivity analysis is conducted to determine the dominant parameters. Sun *et al.*[18] proposed a decomposition-clustering-ensemble (DCE) learning approach for solar

radiation forecasting. To achieve high-accuracy models, Li *et al.*[19] investigated various ensemble learning methods. Three types of global ensemble models, including homogeneous and heterogeneous ensembles, were constructed. Yao *et al.*[20] applied a machine learning algorithm developing a new GPP (Gross Primary Production) dataset for China with 0.1 spatial resolution and monthly temporal frequency based on eddy flux measurements from 40 sites in China and surrounding countries, most of which have not been explored in previous global GPP datasets. In order to improve the accuracy of the results of the size optimization algorithm, Zhang *et al.*[21] introduced weather forecasting together with artificial neural networks for solar radiation, ambient temperature, and wind speed prediction. Previous studies have mainly improved the performance of ensemble learning models from model and parameter selection. In order to increase the diversity of ensemble learning, we have improved the data sampling process and model combination strategy to improve the accuracy and reliability for solar irradiance prediction.

In this paper, a Bayesian model combination based ensemble learning (BMC-EL) prediction method has been proposed for solar energy prediction to improve the reliability of prediction methods. First, clustering[22] and cross validation[23] have been introduced in the data sampling process to generate multiple training subsets so as to ensure that the training subsets are different from each other and can cover a variety of different meteorological samples. Second, an ensemble learning model with a multiple base learner is established; each training subset is utilized to train the corresponding base learner. Here, a random forest is applied as a base learner for ensemble learning. Then, a Bayesian model combination[24] is applied to frame the combination strategy based on the accuracy of each base learner on the validation set. The prediction values of multiple learners are framed through the model combination strategy.

To verify the accuracy and reliability of the proposed method in solar energy prediction, the American Meteorological Society 2013–2014 Solar Energy Prediction Contest dataset[25] is used for experiments. Multikernel-SVM, K-means (Radial Basis Function) RBF, as well as classical SVM and ANN are introduced to establish comparison tests. The experimental results verify the accuracy and reliability of the proposed method in solar energy prediction.

II. DATA SAMPLING

The ensemble learning model will have a better performance when the diversity between base learners is more significant.[24] Therefore, the differences of training subsets should be increased to improve the diversity of the base learners.

The pseudocode of K-means clustering is shown as Fig. 1.

Input: Meteorological data $D=\{x_1,x_2,...,x_m\}$
Number of clusters: K
Randomly select K samples as the initial cluster center $\{\mu_1,\mu_2,...,\mu_m\}$
repeat
Set $C_i=\varnothing \quad (1\le i\le k)$
for $j = 1, 2, \ldots, m$
$d_{ji}=\|x_j-\mu_i\|_2$
$\lambda_j=\arg\min_{i\in\{1,2,...k\}} d_{ji}$
$C_{\lambda_j}=C_{\lambda_j}\cup\{x_j\}$
end for
for $i = 1, 2, \ldots, k$
$\mu_i'=\frac{1}{|C_i|}\sum_{x\in C_i} x$
If $\mu_i'\neq\mu_i$
$\mu_i=\mu_i'$
else
$\mu_i=\mu_i$
end for
until no update
Output $C=\{C_1,C_2,...,C_k\}$

FIG. 1. The pseudocode of K-means clustering.

In order to alleviate the above problems, K means clustering and K fold cross validation have been applied at the same time to increase the diversity of training subsets, as shown in the Fig. 2. In order to distinguish the subscripts of K means clustering, K fold cross validation is replaced by M fold cross validation.

Assuming that the training set is designed to be sampled as training subsets $\{D_1, D_2, \ldots, D_M\}$. The training set is first divided into clusters $\{C_1, C_2, \ldots, C_k\}$ through K means clustering, where each cluster contains samples with similar (adjacent) weather conditions.

After M fold cross validation, clusters $C_1 - C_k$ can be randomly separated into packages $\{b_1^1, b_2^1, \ldots, b_M^1\}, \{b_1^2, b_2^2, \ldots, b_M^2\}, \ldots, \{b_1^k, b_2^k, \ldots, b_M^k\}$. When these packages are imported into the training subsets, M fold cross validation will remove one package in turn. For example, import $\{b_2^1, b_3^1, \ldots, b_M^1\}$ into the training subset D_1 and import $\{b_2^1, b_3^1, \ldots, b_M^1\}$ into the training subset D_2, similarly, different $M - 1$ packets are imported into the corresponding training subsets until $\{b_1^k, b_2^k, \ldots, b_{M-1}^k\}$ is imported into the training subset D_M.

In the process of sampling, we utilize K-means clustering to divide the meteorological data into K clusters. However, subsets are not constructed by combining randomly divided data from each cluster. When each cluster is divided into M packages by M-fold cross validation, only $M - 1$ packages are added to the training subset. In other words, each training subset contains all clusters, but only contains $\frac{M-1}{M}$ samples of each cluster. Since M packages are missing 1 package in turn when added to each training subset, all training subsets can cover complete dataset. Therefore, there are $\frac{1}{M-1}$ samples that are different between any two training subsets.

Therefore, the proposed data sampling method not only increases the diversity of training subsets, but also ensures that the training subset covers all types of meteorological data.

III. BASE LEARNER

Ensemble learning can reduce the prediction risk through the combination of multiple base learners. To meet the stringent requirements of applications, ensemble learning is utilized to improve the reliability of solar energy prediction. The pruning operation of Classification and Regression Tree (CART) in a random forest can effectively alleviate the risk of over-fitting. Random forest is simple, efficient and easy to implement, equipped with the characteristic of small computational cost and excellent generalization ability. So the random forest has been defined as the basic learner of ensemble learning. An individual random forest is composed of multiple CART, and

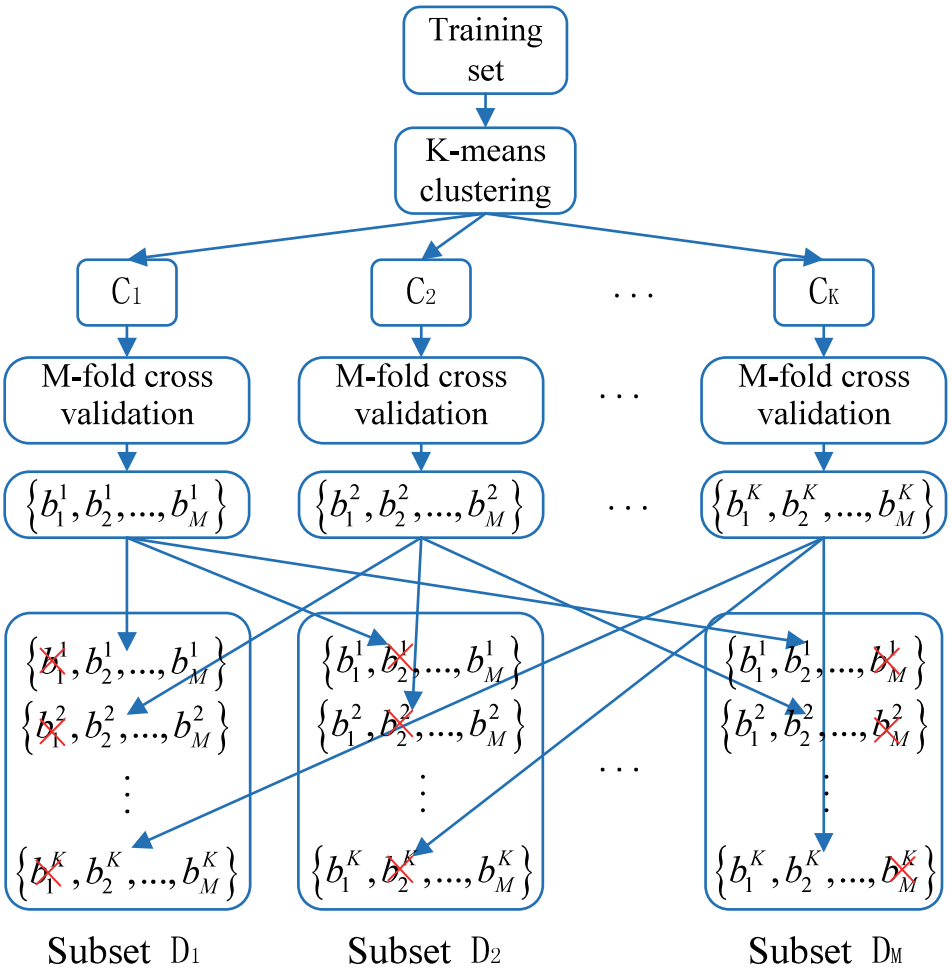

FIG. 2. Schematic diagram of data sampling.

the divided training subsets are prepared for training individual random forests.

Supposing that the training set is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The node of the CART sets a segmentation point s for an attribute variable j of the sample x_i, samples with the input variable greater than s are divided into one part $R_1(j,s) = \left\{x_i | x_i^{(j)} > s\right\}$, otherwise divided into other part $R_2(j,s) = \left\{x_i | x_i^{(j)} < s\right\}$. The partitioned parts are further divided by other different attribute variables, and the samples are divided into m parts according to the node segmentation points, which are, respectively, denoted as R_1, R_2,…,R_m. The corresponding output for each part is defined as c_1, c_2,…,c_m, respectively. The CART could be described as formula (1)

$$f(x) = \sum_{m=1}^{m} c_m I(x \in R_m), \tag{1}$$

where $I(x \in R_m) = \begin{cases} 1(x \in R_m) \\ 0(x \notin R_m) \end{cases}$. The square error of CART is as follows:

$$E = \sum_{x_i \in R_m} \left(y_i - f(x_i)\right)^2 = \sum_{x_i \in R_m} \left(y_i - \sum_{m=1}^{m} c_m I(x \in R_m)\right)^2. \tag{2}$$

When c_m is equal to the average of samples output which belongs to the R_m, the square error is optimal. So $\hat{c}_m = ave(y_i | x_i \in R_m)$.

From the Eq. (2), the optimal output c_m of part R_m can minimize the square error. Traversing all the attribute variables j and the possible segmentation point s in the sample, the $R_m(j,s)$ whose excellent square error is the smallest is defined as part R_m. Similarly, the partitioned parts are further divided and the optimal segmentation variables and segmentation points $(R_m(j,s))$ could be obtained. The final CART model is $f(x) = \sum_{m=1}^{m} \hat{c}_m I(x \in R_m)$.

IV. THE COMBINATION STRATEGY

Ensemble learning can alleviate the poor generalization ability caused by the base learner error, which relies on the ingenious model combination strategy. The Bayesian model combination defines the posterior probability of the model's prediction performance on the validation set as the weight of the model, assigns multiple random forest models with reasonable weights, and selects one best model combination strategy from a plurality of combination strategies. Here, Bayesian model combination has been utilized to generate the random forest combination strategy.

In the Bayesian model averaging, assuming there are n samples in dataset D, and each sample d_i consists of attribute vector x_i and real value y_i. The hypothetical space H contains a finite number of individual hypotheses and h represents an individual hypothesis of hypothetical space. Under the preconditions of hypothetical space H and dataset D, the posterior distribution of y_i is as follows:

$$p(y_i | x_i, D, H) = \sum_{h \in H} p(y_i | x_i, h) p(h | D), \tag{3}$$

where $p(y_i|x_i, D, H)$ is a weighted average of posterior distributions $p(y_i|x_i, h)$ under all individual hypotheses, $p(y_i|x_i, h) = \int p(y_i|\theta_k, h, D) p(\theta_k|h, D) d\theta_k$ is the posterior distribution of y_i under the individual hypothesis h, and θ_k is the parameter vector corresponding to the individual hypothesis h.

The posterior probability $p(h|D)$ of individual hypothesis h under the condition of dataset D can be calculated by equation $p(h|D) = \frac{p(D|h)p(h)}{\sum_{h \in H} p(D|h)p(h)}$. Here $\sum_{h \in H} p(D|h)p(h)$ is a constant, so $p(h|D) \propto p(D|h)p(h)$. $p(D|h) = \int p(D|\theta_k, h) p(\theta_k|h) d\theta_k$ is the integral likelihood estimate of the individual hypothesis h, $p(\theta_k|h)$ is the prior distribution of the vector parameter θ_k corresponding to the individual hypothesis h, and $p(D|\theta_k, h)$ is the likelihood estimation. $p(h)$ is the priori probability of the individual hypothesis h.

In order to ensure that all base learners have a higher predictive performance, there is no difference in base learner parameter initialization, so prior knowledge probability $p(h)$ does not need to be "biased" on any individual hypothesis, so $p(h) = \frac{1}{k}$ (k is the number of individual hypotheses in the hypothetical space).

In the Bayesian model averaging, the calculation of the integral likelihood estimate sets a very high weight on the hypotheses that makes accuracy slightly increased,[24] but this may cause over fitting easily.[26]

In order to alleviate the above condition, the Bayesian model averaging can be modified into the Bayesian model combination (BMC) method. Equation (3) is modified to Eq. (4)

$$p(y_i | x_i, D, H, E) = \sum_{h \in E} p(y_i | x_i, H, e) p(e | D), \tag{4}$$

where e is the individual hypothesis model in the combined model space E. Bayesian model averaging and Bayesian model combination are shown in Figs. 3 and 4 below, respectively.

V. ENSEMBLE LEARNING BASED ON BAYESIAN MODEL COMBINATION

The flow chart of the ensemble learning prediction method based on Bayesian model combination is shown as Fig. 5.

Implementation steps of the ensemble learning prediction method based on Bayesian model combination are as follows:

1. The original data are normalized by the formula $x^* = \frac{x - \min(x)}{\max(x) - \min(x)}$, and cluster $\{C_1, C_2, \ldots, C_k\}$ is generated by K means clustering, M

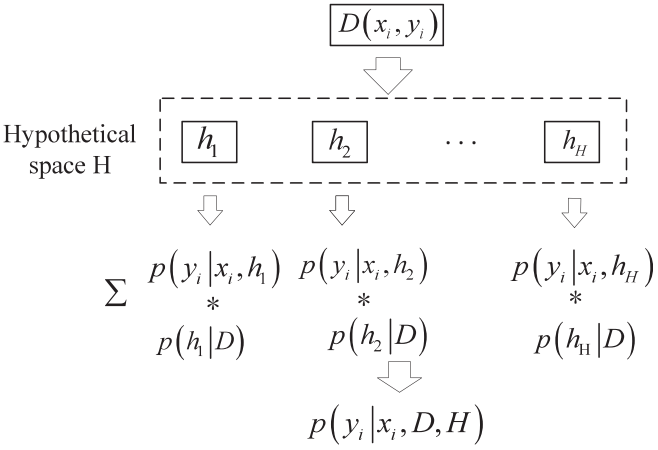

FIG. 3. Bayesian model averaging.

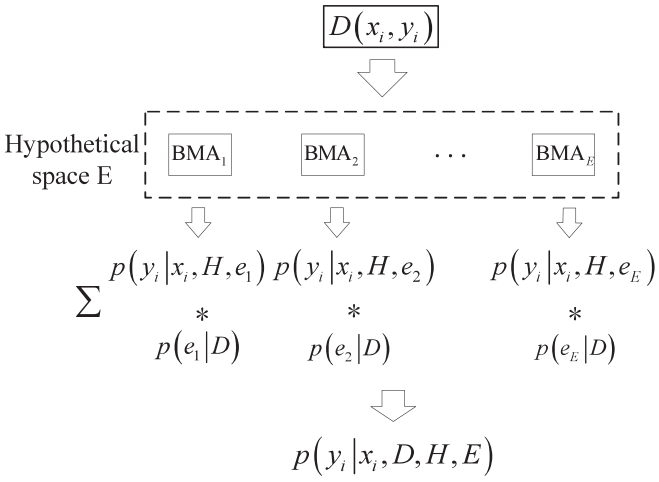

FIG. 4. Bayesian model combination.

fold cross validation is performed on each cluster, and training subsets $\{D_1, D_2, \ldots, D_k\}$ are sequentially generated.
2. Training subsets are used to train base learners (random forest) of ensemble learning.
3. Import validation sets into random forests and output predicted values $(y_1, y_2, \ldots, y_k)$. Assuming that the real output of the verification set is y, a matrix $(y, y_1, y_2, \ldots, y_k)$ is constructed and imported into a Bayesian model combination. Bayesian model combinations can make the optimal model combination strategies $p(y_i|x_i, D, H, E) = \sum_{h\in E} p(y_i|x_i, H, e)p(e|D)$.
4. Import test sets into random forests and output predicted values $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k)$. The final prediction output of the integrated learning algorithm is $p(Y|\bar{y}_k, D, H, E) = \sum_{h\in E} p(y_i|x_i, H, e)p(e|D)$.

VI. SIMULATION STUDY

To test the performance of the proposed method in solar irradiance prediction applications, the American Meteorological Society 2013–2014 Solar Energy Prediction Contest dataset[25] is used to establish prediction experiments.

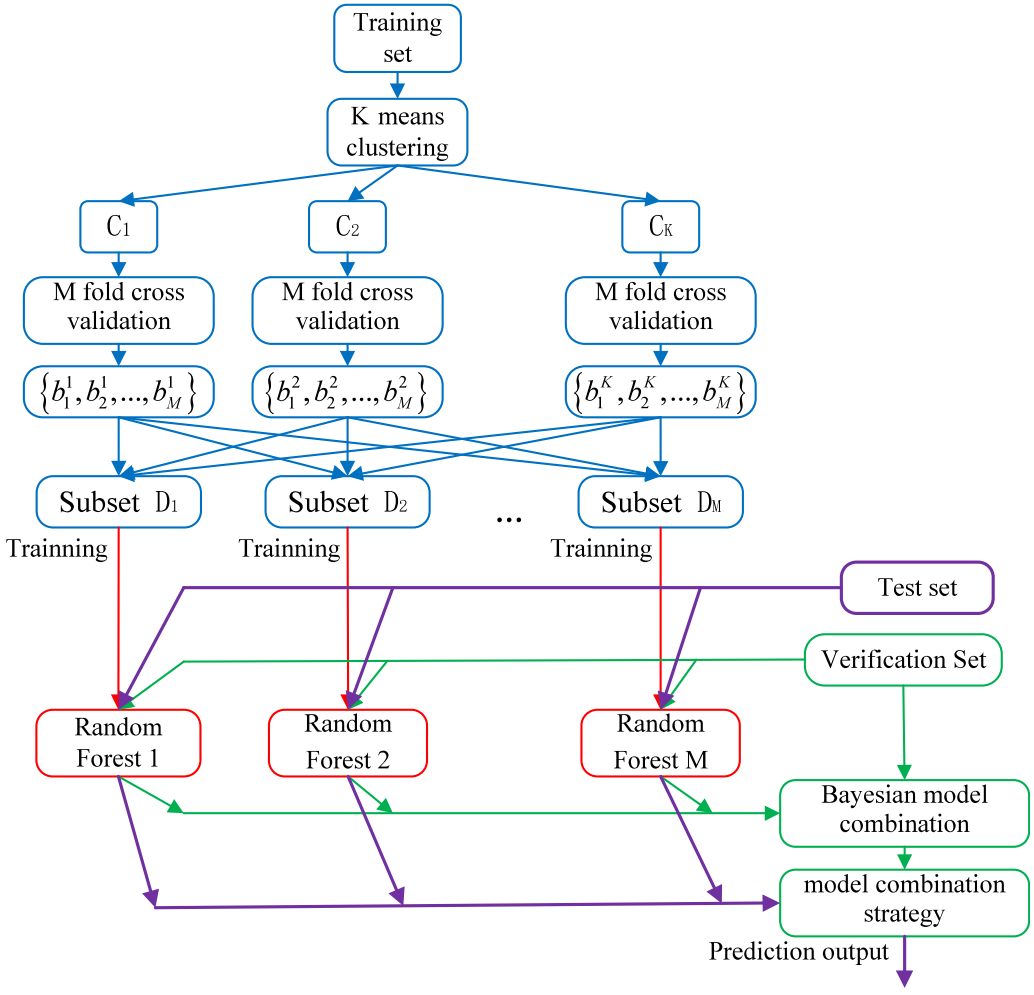

FIG. 5. Ensemble learning based on Bayesian model combination.

In this paper, the meteorological data of the five nearest Global Ensemble Forecast System (GEFS) sites around the Mesonet station are selected as features to predict the solar irradiance of Mesonet stations. For each GEFS site, 15 meteorological attributes are apcp_sfc (3-h accumulated precipitation at the surface, kilogram per square centimeter), dlwrf_sfc (downward long-wave radiative flux average at the surface, watts per square meter), dswrf_sfc (downward short-wave radiative flux average at the surface, watts per square meter), pres_msl (air pressure at mean sea level, pascal), pwat_eatm (precipitable water over the entire depth of the atmosphere, kilogram per square centimeter), spfh_2m (specific humidity at 2 m above ground), tcdc_eatm (total cloud cover over the entire depth of the atmosphere, %), tcolc_eatm (total column-integrated condensate over the entire atmos, kilogram per square centimeter), tmax_2m (maximum temperature over the past 3 h at 2 m above the ground, Kelvin), tmin_2m (minimum temperature over the past 3 h at 2 m above the ground, Kelvin), tmp_2m (current temperature at 2 m above the ground, Kelvin), tmp_sfc (temperature of the surface, Kelvin), ulwrf_sfc (upward long-wave radiation at the surface, watts per square meter), ulwrf_tatm (upward long-wave radiation at the top of the atmosphere, watts per square meter), uswrf_sfc (upward short-wave radiation at the surface, watts per square meter), respectively, which have been applied in this paper. All 15 features are sampled 5 times a day, and the sampling times are UTC 12:00, UTC 15:00, UTC 18:00, UTC 21:00, and UTC 24:00.

The dataset contains total daily incoming solar energy at Mesonet sites and the meteorological data of the Global Ensemble Forecast System (GEFS). The goal of this dataset is to discover which statistical and machine learning techniques provide the best short-term predictions of solar energy production based on numerical meteorological data. The forecasting time horizon of the solar irradiance prediction is day-ahead; it should be noted that the meteorological data used for solar irradiance prediction is day-ahead, which is the prediction meteorological data of the GEFS site for the next day.

The experimental platform is 64-bit Windows 10, 8 GB RAM, 64-bit Matlab 2016b. The BMC calculates and improves the model combination strategy relying on the R-BMS (R package-Bayesian Model Sampling) imported by Matlab 2016b.

The unit of solar energy in the dataset is MJ$\times$m^{-2} or J$\times$m^{-2}, manuscript remains uniform, where 1 MJ$\times$m^{-2} = 0.28 KWh$\times$m^{-2}.

A. Performance indicators

The MAE (Mean Absolute Error), MSE (Mean Square Error), AER (Average Error Rate), RS (rate of success), MAPE (Mean Absolute Percentage Error), R^2 (R-square), and R^2_{adj} (adjusted R-square) have been introduced as performance indices, as described in formulas (5)–(10)

$$Er = \frac{|Y_{pre} - Y_{real}|}{Y_{pre}}, \tag{5}$$

$$AER = \frac{\sum_{i=1}^{n} Er(i)}{n}, \tag{6}$$

$$RS(0.1) = \frac{num(Er < 0.1)}{n}, \tag{7}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Y_{real} - Y_{pre}}{Y_{real}}\right|, \tag{8}$$

$$R^2 = 1 - \frac{\sum (Y_{real} - Y_{pre})^2}{\sum \left(Y_{real} - \frac{1}{n}\sum_{i=1}^{n} Y_{pre}\right)^2}, \tag{9}$$

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}, \tag{10}$$

where Y_{pre} is the prediction output, Y_{real} is the real data, Er is the error rate for each sample, and AER is the average error rate. n is the total number of samples in the test set, and num is the number of samples whose error rate is less than 0.1. p is the number of meteorological variables.

The AER reflects the average value of prediction error. RS reflects the reliability of the prediction method. $MAPE$ is an indicator of accuracy in statistics. R^2 and R^2_{adj} reflect the degree to which the prediction value fit the real value. R^2_{adj} eliminates the impact of the samples and attributes. The closer R^2 and R^2_{adj} are to 1, the closer the prediction value is to the real value.

B. Diversity of the training subset

Normalize the raw meteorological data to $[-1, 1]$, and then perform a 10 means clustering operation on the training set to divide the training set into 10 clusters $\{C_1, C_1, \ldots, C_{10}\}$. The dlwrf_sfc, dswrf_sfc, and tmp_sfc of the samples are applied to establish a three-dimensional coordinate; the samples in the training set are distributed in the coordinate as shown in Fig. 6.

The divided clusters are, respectively, subjected to tenfold cross validation, and 9 different packages are, respectively, introduced into each training subset. The distribution of samples under different weather conditions in the training subset is shown in Fig. 7.

Since clustering is followed by tenfold cross validation, the size of each training subset is 90% of the training set. According to distribution of samples in training subsets, the sampling process does not

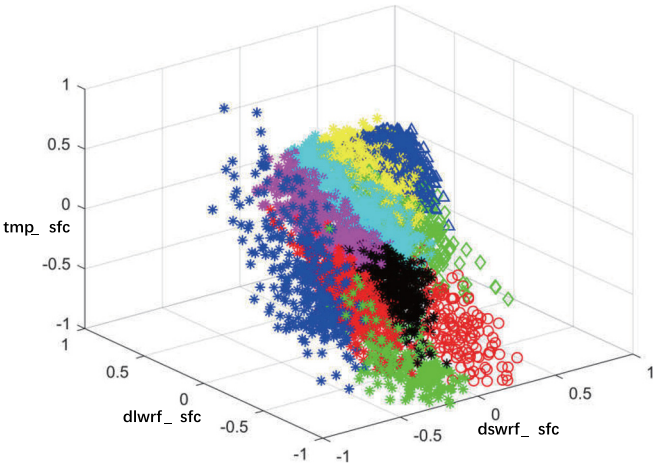

FIG. 6. Distribution of the training set.

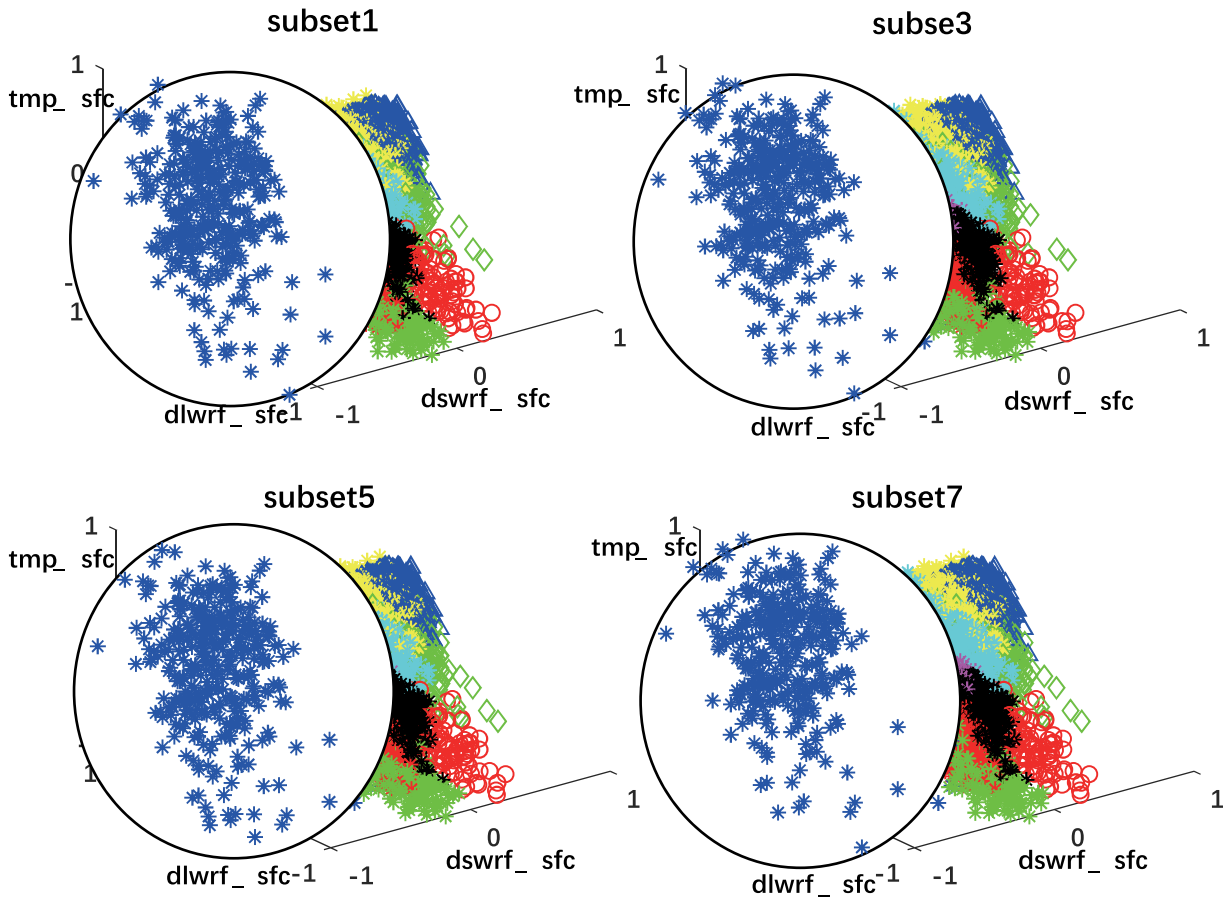

FIG. 7. Distribution of training subsets (4 typical training subsets are given to illustrate the distribution).

affect the sample distribution of different weather conditions. On the other hand, the cross validation operation of the clusters also increases the diversity of training subsets (as samples of the marked area), and correspondingly increases the sample perturbation of the base learners.

Due to the randomness of random sampling, the sampling process is inevitably biased. To compare the proposed sampling method and random sampling, sampling experiments were built and the distribution of samples was monitored. The sampling process consists of 5113 samples with a sampling rate of 90%. The 10 subsets, divided in the sampling process, have been utilized to evaluate the sampling performance.

The range of samples R is defined in terms of solar irradiance. Suppose the dataset contains M samples in R and the subset contains N samples in R, then the performance of sampled samples is defined as $y = \frac{N}{M} \times 100\%$. The results are given in Figs. 8 and 9.

As shown in Figs. 8 and 9, lines with different colors represent different subsets. Due to the clustering and cross validation, the proposed sampling method is more convergent than random sampling under different solar irradiance (the sampling rate of the proposed sampling method is more close to 90%, with much smaller fluctuation).

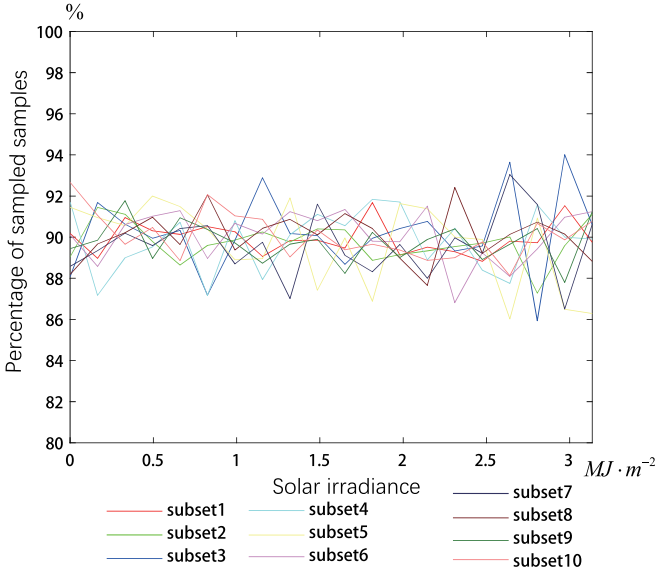

FIG. 8. Sampling rate of the proposed sampling method under different solar irradiances.

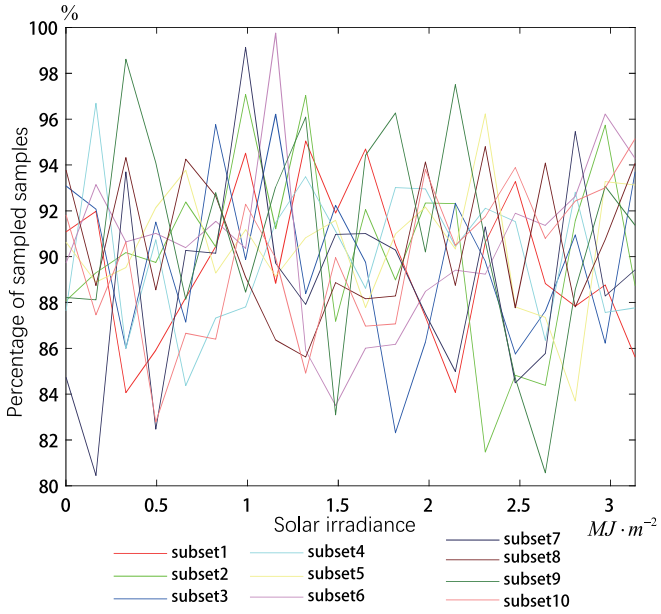

FIG. 9. Sampling rate of the random sampling method under different solar irradiances.

C. Model error estimation and performance testing

Random forests estimate model errors out of bag (OOB). Due to the attribute disturbance, the random forest will converge to a lower generalization error only when there is a certain amount of individual CART. Importing the training set, the relationship between the OOB error and the number of CART is shown in Fig. 10.

As can be seen from Fig. 10, when the number of CART reaches 200, the OOB error of random forests tends to decrease slowly. In order to make the random forest have a lower error and save the computational cost, the number of CART is set to 200.

The solar irradiance of the HOBA (A Mesoscale station of the United States) Mesoscale station and the meteorological data of the surrounding GEFS stations from 1994 to 2007 (5113 samples) are used to test the proposed method. In the chronological order, the first 4000 samples are used as the training set, the next 500 samples are defined as the verification set, and the samples 4501–4550 are defined as the test set. The experimental results are as shown in Fig. 11.

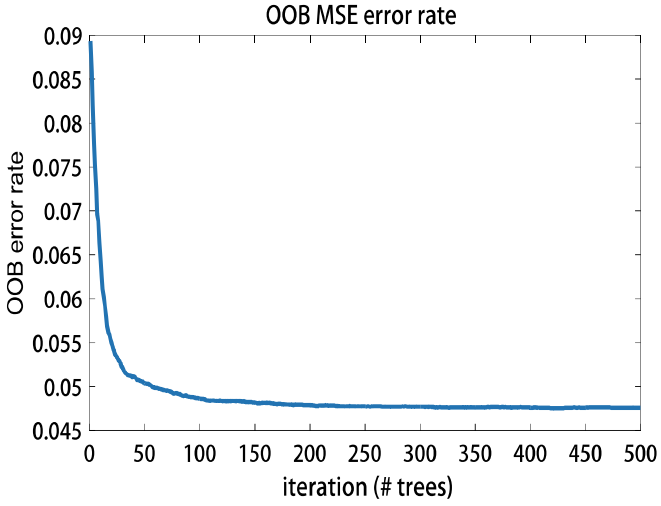

FIG. 10. Error estimation of the random forest.

Figure 11(a) depicts the real values and prediction values of 50 samples, while Fig. 11(b) illustrates the *MAE* of 50 samples predictions, of which $MAE = \frac{1}{m}\sum_{i=1}^{m} |y_{pre}^{(i)} - y_{real}^{(i)}|, (m = 1)$. Figure 11(c) shows the distribution of the prediction error rate (*Er*).

In Fig. 11(a), the long blue line indicates the average of solar irradiance under fine weather conditions. The samples near the blue line have a better prediction performance, indicating that the proposed method has capability for solar irradiance prediction under fine weather conditions.

The samples in the blue dotted box of Fig. 11(a) show slightly larger errors. These samples are far away from the blue line, indicating that these samples have a low solar irradiance and complex meteorological conditions which causes a big interference. Although the sample in the third dotted box is closer to the blue line, the first two samples of this sample appear to fluctuate greatly, indicating that fluctuating meteorological conditions have a significant impact on the prediction of solar energy. Continuously fine weather conditions are of great benefit for accurate prediction, such as the samples framed by the red box, where their MAEs are kept within 1 $MJ \times m^{-2}$.

The sample in the yellow dotted ellipse of Fig. 11(a) has a very low solar irradiance, suggesting bad weather conditions. However, the proposed solar prediction method still guarantees stable prediction performance. Although its prediction error is less than $MJ \times m^{-2}$, the evaluation of this sample is not dominant when calculating the error rate due to its low solar irradiance [as the sample in the dotted ellipse of Fig. 11(c)].

It can be drawn from the distribution graph that the error rate (*Er*) of more than 70% samples are less than or equal to $\pm 2.5\%$, and the error rate of other samples is less than or equal to $\pm 7.5\%$ except for extreme weather conditions.

D. Establishment of experiments

The settings of RF are consistent with RFs in BMC-EL.

1. ANN

Applying the ANN method in reference,[11] the neural network is reproduced in comparison tests. Where the trainParam.epochs and trainParam.goal are set to 5000 and 0.00001, respectively, the learning rate is set to 0.1.

2. K-means-RBF

K-means-RBF[13] is introduced for comparison tests, where the parameter are fixed as: the density coefficient $\psi = 0$, overlap coefficient $\varepsilon = 1$, and cluster radius $\alpha = 1$. The iteration of the gradient descent algorithm is fixed as 5000.

3. SVM

The SVM[7] have been carried for comparison tests. Both cost and gama are set to 1. The epsilon-SVR (Support Vactor Regerssion) model and the RBF kernel function are used here, where the loss function epsilon = 0.01.

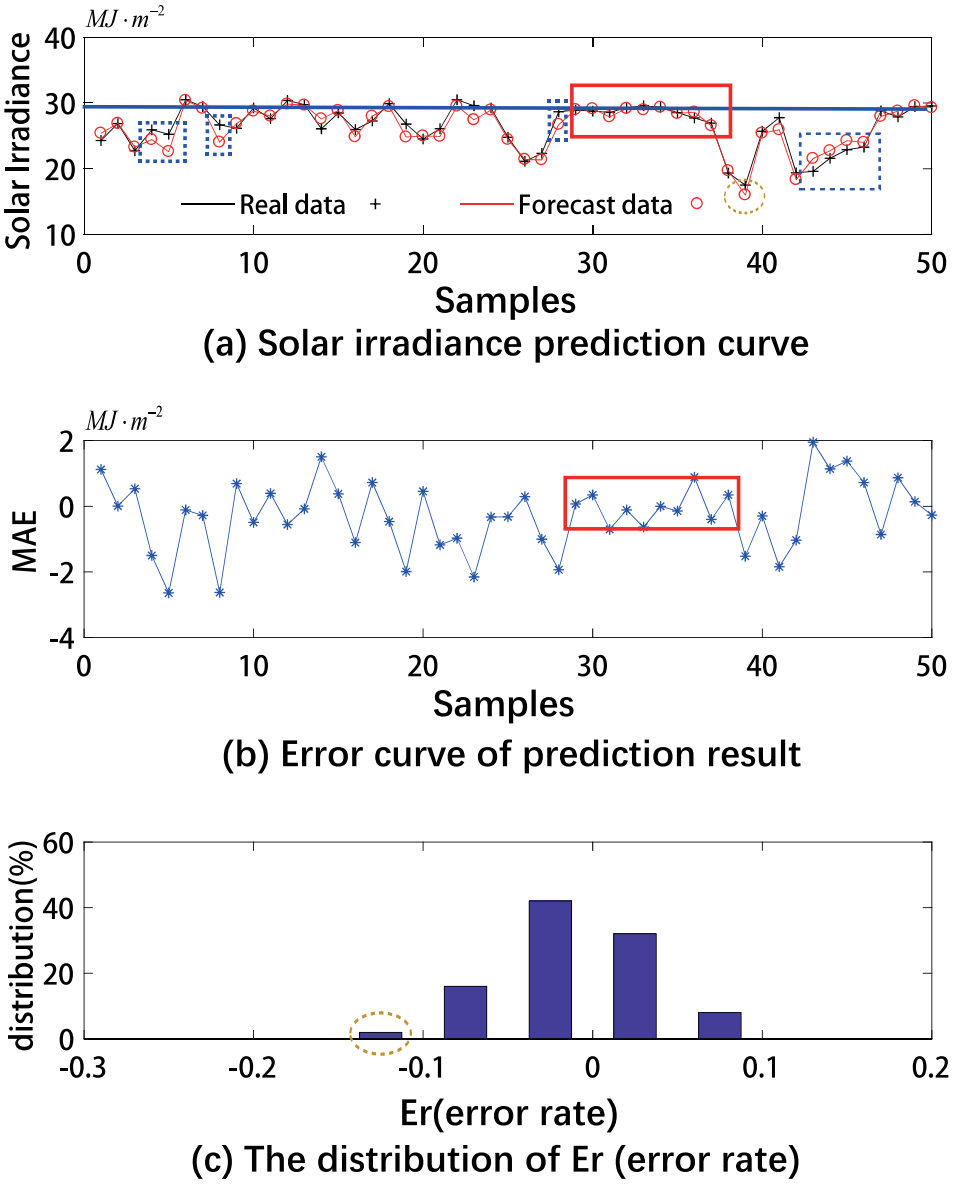

FIG. 11. Prediction performance of the proposed method.

4. Multikernel-SVM

The Multikernel-SVM[9] is also presented in comparison tests. Its parameters are the same as SVM. Multikernel is defined as formula (8)

$$k(x,y) = 0.15 \times \left[x^T y + c\right] + 0.15 \times \left[\left(ax^T y + c\right)^d\right] + 0.5 \times e^{-\gamma||x-y||^2} + 0.2 \times e^{-\frac{||x-y||^2}{2\sigma^2}}. \quad (11)$$

E. Experimental results and observations

The solar irradiance of the HOBA Mesoscale station and the meteorological data of the surrounding GEFS stations are defined as the dataset. The samples from January 1, 1994 to December 31, 2004 are defined as a training set. The samples from January 1, 2005 to December 31, 2006 are defined as a validation set. The samples of 2007 are utilized as a test set for solar energy prediction experiments.

Four representative months (February, May, August, and November) are selected to demonstrate the prediction output as well as prediction error, as shown in Figs. 12–15.

As shown in Fig. 12(a), the solar irradiance of the samples in the red box is low indicating complex meteorological conditions which may causes interference to solar irradiance prediction. Different prediction methods can basically track the trend of solar irradiance on these harsh samples, and the prediction curves of SVM and Multikernel-SVM show large spikes (errors). Although SVM and Multikernel-SVM have better prediction accuracy on two samples in the red dotted ellipse in Fig. 12(a); other tow spikes occur on the blue

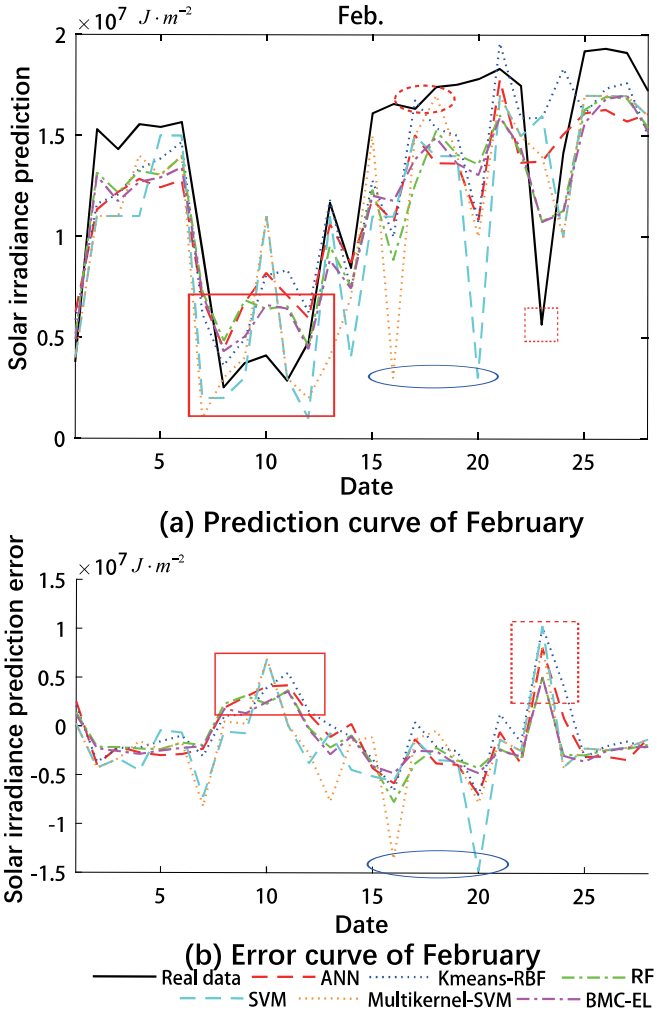

FIG. 12. Prediction performance of different methods in February.

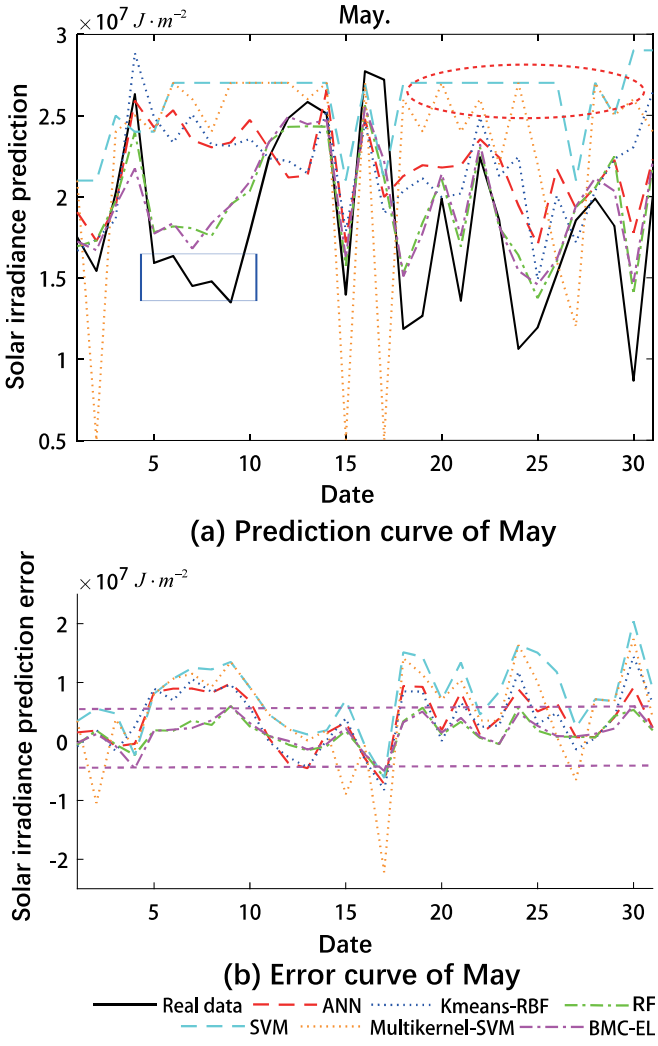

FIG. 13. Prediction performance of different methods in May.

ellipse-framed samples. The prediction curves of SVM and Multikernel-SVM show large spikes, respectively, although we have maximally alleviated over-fitting and under-fitting. The solar irradiance of the sample in the red dotted box of Fig. 12(a) returns to normal after a sharp decay, where the prediction outputs of different methods show a large deviation. The error directions of the different prediction methods are consistent in the red dotted box of Fig. 12(b), and the peak amplitude of the BMC-EL method is the smallest. The output curve of the BMC-EL on the harsh samples is closest to the real curve.

As shown in Fig. 13(a), solar irradiance fluctuated sharply in May, which caused obstacles to solar prediction. Although numerous trials have been carried out for experiment result fitting, SVM and Multikernel-SVM deviate significantly from the real value, as shown in the red ellipse. In the blue box-framed samples, the output curves of RF and BMC-EL are closer to the real curve, which shows that ensemble learning is more advantageous on harsh samples and BMC-EL has a better performance. With the error curve shown in Fig. 13(b), the error curve boundary of the BMC-EL in the third graph is also a learner.

Solar irradiance is abundant in August. As shown in Fig. 14, continuously fine weather conditions are of great benefit for accurate prediction, such as samples in the blue box of Fig. 14, which is consistent with the analysis of Part E. Consistent with the previous, such as samples in the red dotted box of Fig. 14(b), BMC-EL maintains stable prediction performance in samples with large fluctuations.

As shown in Fig. 15, since Bayesian model combination chooses the best model combination strategy from hypothetical space, BMC-EL can guarantee the stable prediction precision in some harsh prediction samples, and thus, BMC greatly improves the reliability of the prediction.

A cartesian coordinate is established, the abscissa indicates the measured value of solar irradiance (real value), and the vertical axis

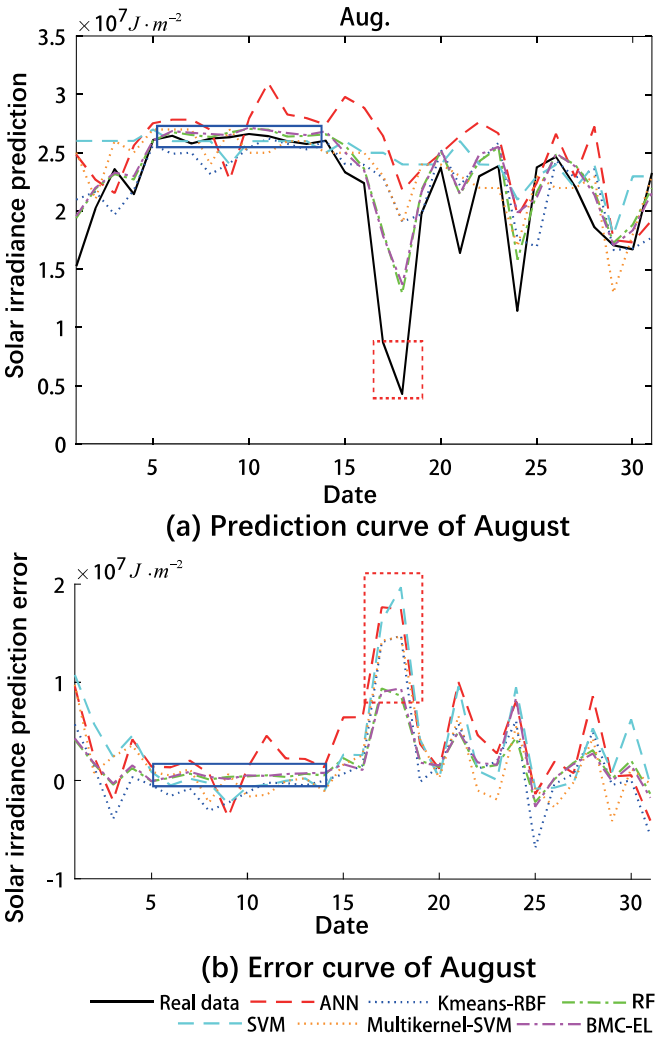

FIG. 14. Prediction performance of different methods in August.

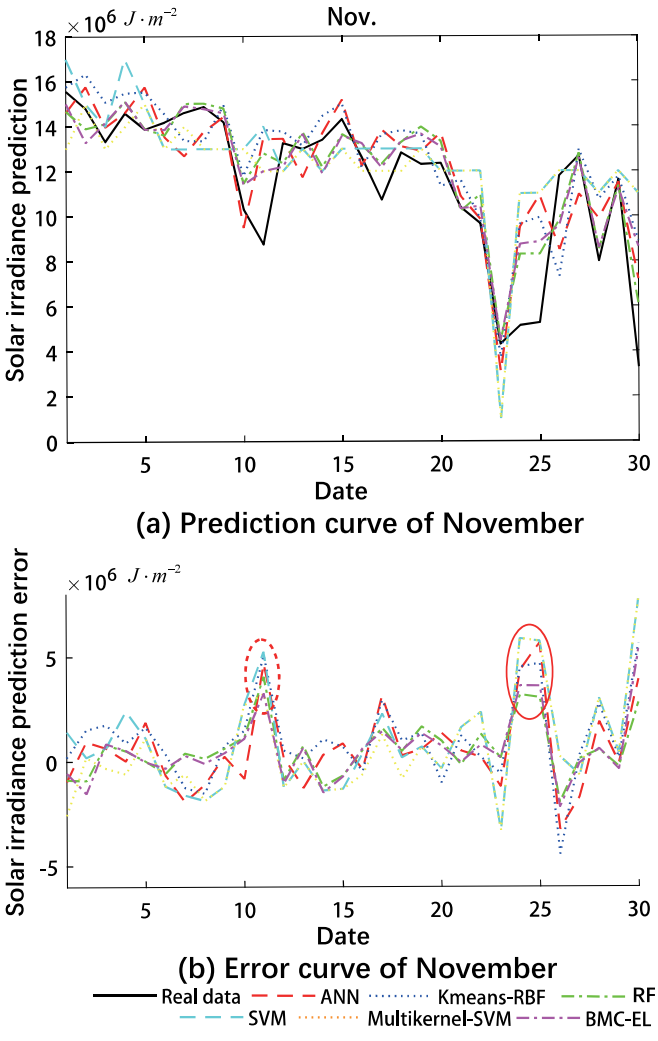

FIG. 15. Prediction performance of different methods in November.

indicates the prediction value. The line $y = x$, defined as the baseline, indicating that the predicted value is equal to the real value. The prediction results of the remaining eight months are plotted in the cartesian coordinate, as shown in Fig. 16, the closer the point is to the baseline, the more accurate the prediction result is. The red dotted line divides the scatterplot into three parts, which, respectively, indicate that solar irradiance is scarce, fairish and rich.

In the solar irradiance prediction simulation experiment, the samples corresponding to different prediction methods fall near the baseline. We have performed abundant tries to adjust the experimental parameters as much as possible and tried several times to ensure the optimal performance of the different prediction methods.

In the solar-rich part, meteorological conditions receive less interference, so samples of different prediction methods are centralized around the baseline. In the solar-scarce area, the meteorological environment is complex and the interference is large, and the samples of different prediction methods have large deviations. However, the scatterplot distribution of RF and BMC-EL methods is the most harmonious. Since Bayesian model combination chooses the best model combination strategy from hypothetical space, the scatterplot of the BMC-EL method is the most centralized.

The performance indicators of different prediction methods throughout the year are shown in Table I, of which the proposed BMC-EL has the best performance.

The prediction performance of random selection and interpolation methods (Gaussian mixture model and Catmull-Rom cubic splines) has also been chosen as datum. In the comparison experiment, the annual average Mean Absolute Error (MAE) of random normal selection is 9.20 MJ×m^{-2}, the annual average MAE of the Gaussian mixture model is 4.02 MJ×m^{-2}, and the annual average MAE of Catmull-Rom cubic splines is 3.61 MJ×m^{-2}. However, these mathematical strategies have not achieved a good performance.

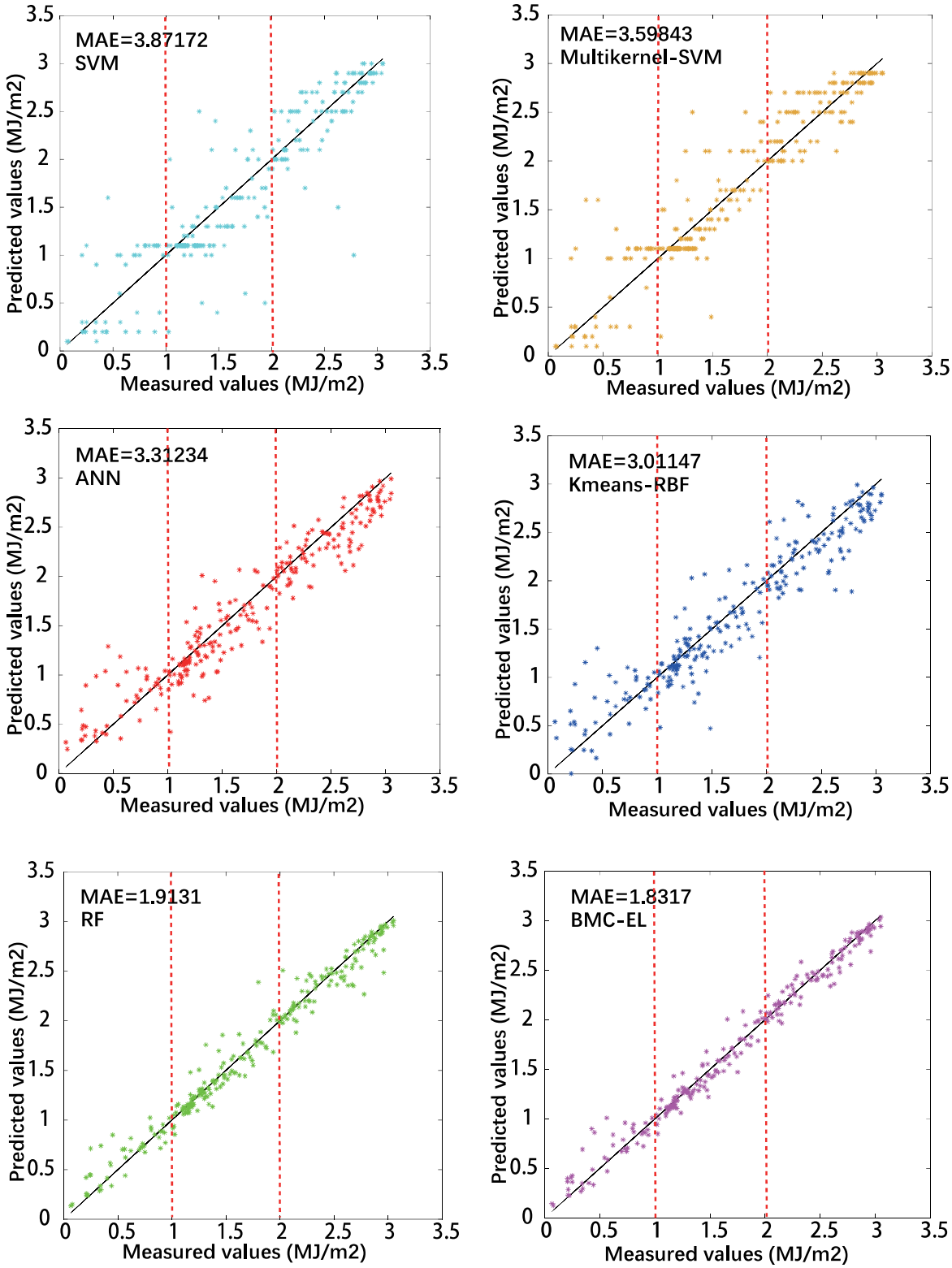

FIG. 16. Scatter plot of solar irradiance prediction for the remaining eight months.

TABLE I. The performance indicators of different prediction methods.

	MSE	MAE	RS	AER
BMC-EL	6.79×10^{12}	1.8628	53.87%	0.2085
RF	7.53×10^{12}	1.9002	51.42%	0.21
ANN	1.95×10^{13}	3.2448	36.10%	0.31796
K-means-RBF	1.78×10^{13}	3.01484	40.08%	0.2966
Multikernel-SVM	2.95×10^{13}	3.62957	44.30%	0.3468
SVM	3.27×10^{13}	3.88253	40.18%	0.3738

	MAPE	R^2	R^2_{adj}
BMC-EL	20.85	0.9661	0.9640
RF	21	0.9166	0.9145
ANN	31.796	0.8853	0.8781
K-means-RBF	29.66	0.8663	0.8580
Multikernel-SVM	34.68	0.8609	0.8522
SVM	37.38	0.8186	0.8073

In summary, the random forest prediction method based on Bayesian model combination has good prediction performance and reliability in solar irradiance prediction, which can achieve accurate and stable prediction under different weather conditions.

F. Calculation cost

In the solar irradiance prediction experiment, the training set contains 4382 samples and the test set contains 365 samples. The average training time and prediction time of various prediction methods are shown in Table II.

Since the BMC-EL combines multiple RF models, its training time and test time are longer than other algorithms. Although the BMC-EL increases the computational cost, the output has much higher reliability and accuracy, which is well worth.

VII. CONCLUSION

An ensemble learning method based on Bayesian model combination is proposed for solar irradiance prediction, which aims to improve the reliability of solar power application. First, the proposed data sampling process not only ensures uniform sampling, but also improves the diversity of the training subset, which is able to improve the diversity of the base learner. Second, multiple training subsets are utilized to train the individual random forests in the ensemble learning. So far, the diversity of ensemble learning comes from the diversity of training subsets. After that, the novel Bayesian model combination strategy expends hypothesis space E on Bayesian model averaging, which is used to construct a combination for base learners. It formulates a combination strategy based on the performance of each base learning on the verification set by choosing the best model combination strategy from hypothetical space, and thus effectively improves the performance of the model. In solar energy prediction experiments, BMC-EL significantly reduces the uncertainty of a single learner by adding Bayesian model combination and increases the reliability of the result. The experimental results show that the proposed prediction method has a better prediction accuracy and reliability, which can be utilized to estimate dynamic fluctuating solar power. Therefore, the proposed method provides a basis for accurate estimation of solar power, which can promote further development of the whole power system.

TABLE II. The computational time of different prediction methods.

	Training process	Prediction process
BMC-EL	23.827 s	0.631 s
RF	3.115 s	0.143 s
ANN	5.231 s	0.155 s
K-means-RBF	6.614 s	0.171 s
Multikernel-SVM	10.167 s	0.152 s
SVM	10.035 s	0.164 s

We provide a novel analytical method for solving prediction and data analyzing problems that are prevalent in the renewable energy fields. In the future, we will introduce more meteorological data and geographic attributes to predict solar irradiance, and will explore more potential applications in the field of energy data analysis.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under No. 61773282 and was partially supported by a grant of the Hong Kong Polytechnic University (H-ZG3K). In addition, Jian-Fang Chang would like to thank, in particular, the patience, care, and support from Xing Liu over the past years.

REFERENCES

[1]A. Mellit and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy," Sol. Energy **84**, 807–821 (2010).
[2]Y. Wu and J. Wang, "A novel hybrid model based on artificial neural networks for solar radiation prediction," Renewable Energy **89**, 268–284 (2016).
[3]I. Maity and S. Rao, "Simulation and pricing mechanism analysis of a solar-powered electrical microgrid," IEEE Syst. J. **4**, 275–284 (2010).
[4]M. K. Das, K. C. Jana, and A. Sinha, "Performance evaluation of an asymmetrical reduced switched multi-level inverter for a grid-connected PV system," IET Renewable Power Gener. **12**, 252–263 (2018).
[5]J. Lin, "Potential impact of solar energy penetration on PJM electricity market," IEEE Syst. J. **6**, 205–212 (2012).
[6]U. Akram, M. Khalid, and S. Shafiq, "Optimal sizing of a wind/solar/battery hybrid grid-connected microgrid system," IET Renewable Power Gener. **12**, 72–80 (2018).
[7]F. Jiang, H. Liu, and X. Yang, "Short-term solar radiation prediction based on SVM with similar data," in Renewable Power Generation Conference (2014).
[8]T. Li, W. Guo, and L. I. Mingjia, "Parameter identification of Hammerstein ARMAX model based on APSO-WLSSVM algorithm," China Sci. Paper **2**, 136–142 (2018).
[9]S. Alam, M. Kang, J. Y. Pyun, and G. R. Kwon, "Performance of classification based on PCA, linear SVM, and multi-kernel SVM," in Eighth International Conference on Ubiquitous and Future Networks (2016).
[10]Y. Zhou, X. Cui, Q. Hu, and J. Yuan, "Improved multi-kernel SVM for multi-modal and imbalanced dialogue act classification," IEEE 2015 International Joint Conference on Neural Networks, 1–8 (2015).
[11]K. M. Rabbi, I. Nandi, A. S. Saleh, F. Faisal, and S. Mojumder, "Prediction of solar irradiation in Bangladesh using artificial neural network (ANN) and data

mapping using GIS technology," in Development in the in Renewable Energy Technology (2016).
[12]M. Anamika, "Prediction and efficiency evaluation of solar energy resources by using mixed ANN and DEA approaches," in PES General Meeting: Conference and Exposition (2014).
[13]A. K. Yadav, H. Malik, and S. S. Chandel, "ANN based prediction of daily global solar radiation for photovoltaics applications," in India Conference (2015).
[14]H. D. Chiang, L. G. Chen, R. P. Liu, and N. Dong, "Group-based chaos genetic algorithm and non-linear ensemble of neural networks for short-term load forecasting," IET Gener. Transm. Distrib. **10**, 1440–1447 (2016).
[15]H. Baili and Y. F. Li, "Online reliability prediction of energy systems with wind generation," in IEEE International Midwest Symposium on Circuits and Systems (2016).
[16]J. Vedelsby and A. Krogh, "Neural network ensembles, cross validation, and active learning," in International Conference on Neural Information Processing Systems **7**(10), 231–238 (1995).
[17]V. Nourani, G. Elkiran, J. Abdullahi, and A. Tahsin, "Multi-region modeling of daily global solar radiation with artificial intelligence ensemble," Nat. Resour. Res. **1**, 1–22 (2019).
[18]S. Sun, S. Wang, G. Zhang, and J. Zheng, "A decomposition-clustering-ensemble learning approach for solar radiation forecasting," Sol. Energy **163**, 189–199 (2018).
[19]H. Li, Y. Cui, Y. Liu, W. Li, Y. Shi, C. Fang, H. Li, T. Gao, L. Hu, and Y. Lu, "Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells," IEEE Access **6**, 1 (2018).
[20]Y. Yao, X. Wang, Y. Li, and T. Wang *et al.*, "Spatiotemporal pattern of gross primary productivity and its covariation with climate in China over the last thirty years," Global Change Biol. **24**(1), 184–196 (2018).
[21]W. Zhang, A. Maleki, M. A. Rosen, and J. Liu, "Sizing a stand-alone solar-wind-hydrogen energy system using weather forecasting and a hybrid search optimization algorithm," Energy Convers. Manage. **180**, 609–621 (2019).
[22]D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," Mach. Learn. **75**, 245–248 (2009).
[23]J. Wan, A. Canedo, and M. A. A. Faruque, "Functional model-based design methodology for automotive cyber-physical systems," IEEE Syst. J. **11**, 1–12 (2017).
[24]K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez, "Turning Bayesian model averaging into Bayesian model combination," in International Joint Conference on Neural Networks (2011).
[25]See https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest for "AMS 2013–2014 solar energy prediction contest, forecast daily solar energy with an ensemble of weather models."
[26]B. Clarke, "Comparing Bayes model averaging and stacking when model approximation error cannot be ignored," J. Mach. Learn. Res. **4**, 683–712 (2003).