Hindawi
Complexity
Volume 2022, Article ID 7495651, 13 pages
https://doi.org/10.1155/2022/7495651

Research Article

Interval Prediction Method for Solar Radiation Based on Kernel Density Estimation and Machine Learning

Meiyan Zhao,[1] Yuhu Zhang,[2] Tao Hu ,[1] and Peng Wang[3]

[1] *School of Mathematical Sciences, Capital Normal University, Beijing 100048, China*
[2] *College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China*
[3] *Faculty of Civil Engineering and Mechanics, Jiangsu University, Zhenjiang, 212013, China*

Correspondence should be addressed to Tao Hu; hutao@cnu.edu.cn

Received 6 July 2021; Revised 21 December 2021; Accepted 27 January 2022; Published 18 February 2022

Academic Editor: Daniele Salvati

Precise global solar radiation (GSR) data are indispensable to the design, planning, operation, and management of solar radiation utilization equipment. Some examples prove that the uncertainty of the prediction of solar radiation provides more value than deterministic ones in the management of power systems. This study appraises the potential of random forest (RF), V-support vector regression (V-SVR), and a resilient backpropagation artificial neural network (Rprop-ANN) for daily global solar radiation (DGSR) point prediction from average relative humidity (RHU), daily average temperature (AT), and daily sunshine duration (SD). To acquire more accurate predictions of DGSR and examine the influence of historical DGSR on the performance of point prediction models, two different model inputs are considered: (1) three meteorological variables and (2) the lags of DGSR and three meteorological variables. Then, two interval prediction methods are developed by introducing the KDE to out-of-bag (OOB), introducing kernel density estimation (KDE) to split conformal (SC) based on the three machine learning models. The two methods for interval prediction are denoted as OOB-KDE and SC-KDE. The mean absolute error (MAE), mean relative error (MRE), and Kendall rank correlation (Kendall) are used to assess the point prediction models. The performance of interval prediction methods is evaluated by the prediction interval coverage probability (PICP), prediction interval normalized average width (PINAW), and coverage width criteria (CWC). The following conclusions are drawn from this study. First, the V-SVR model performs best with the lowest mean absolute error (MAE) of 0.016 and mean relative error (MRE) of 0.001. Second, the lags of DGSR improve the prediction accuracy by about 30%. Third, the OOB-KDE and SC-KDE methods improved the quality of the prediction interval (PI). OOB-KDE improved CWC by 81%, and SC-KDE improved CWC by 99.99%. Fourth, the best interval prediction result is obtained using the SC-KDE method using the V-SVR model. The average difference between its PICP and prediction interval nominal coverage (PINC) is only 3% of the PINC, and its PINAW is less than 0.007.

1. Introduction

Today, 85% of the worldwide energy demand is satisfied using fossil fuels [1]. However, heavy dependence on fossil fuel and limited refinement of nature as well as mismanagement of waste sources has caused environmental crises such as an increase in the average temperature of the Earth's surface caused by greenhouse gases [2]. Moreover, fossil fuels are nonrenewable primary energy sources that cannot satisfy the continuously increasing energy demand [3]. As an alternative, solar energy is a renewable and clean energy source that shows the greatest promise for satisfying the energy demand [4], and it has many useful applications for architectural design, evapotranspiration estimates, agriculture, and atmospheric, land, ocean, and hydrologic models. Precise global solar radiation (GSR) data are indispensable to the design, planning, operation, and management of solar radiation utilization equipment. However, direct GSR measurements using instruments are costly and difficult to perform in multiple regions. Therefore, a GSR prediction method based on common meteorological data demands to be developed [5]. Many GSR prediction methods have been developed thus far. Usually, the models for SR prediction can be classified into numerical weather prediction models, empirical models [6–8], time series models [9, 10], machine learning models [11–15], and deep learning models [16–18].

Numerical weather prediction is based on a physics process to establish a numerical weather prediction model, and then the data of numerical weather prediction is put into the numerical weather prediction model to obtain the prediction value. Numerical weather prediction can usually obtain accurate GSR prediction. However, it depends on various means (such as radar observation and satellite observation) to obtain meteorological data, and the calculation depends on numerical methods, so it is difficult to be used in areas with poor scientific and technological resources. Empirical models are a series of models that estimate solar radiation based on linear or polynomial relationships between solar radiation and specific variables such as sunshine duration and temperature. Empirical models have several limitations, such as different cases presenting different results owing to the high randomness of actual data [19]. Studies have confirmed that empirical models have lower prediction accuracy than machine learning models [20]. Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. In the field of solar radiation, deep learning models are mainly used to deal with big data to meet the requirements of admirable precision and the demands of the operator. With respect to machine learning models, deep learning models pay more computational expense.

In recent years, machine learning models have been used in various applications and have displayed excellent prediction performance. Commonly used machine learning models include artificial neural networks (ANN), random forest (RF), and support vector regression (SVR). Zeng et al. [21] used an RF model to construct a high-density network for the daily global solar radiation (DGSR). They demonstrated that the RF model could accurately predict the DGSR under different climatic and geographical conditions in China. Moreover, they found that the RF method was superior to a decision tree, backpropagation neural network (BPNN), SVR, and multiple linear regression (MLR). Fan et al. [22] assessed the performance of a support vector machine (SVM) and four tree-based soft computing models for predicting the diffuse horizontal solar radiation. They found that the SVM model showed the highest prediction accuracy and the strongest generalization ability among all the models they tested. Sahin et al. [23] investigated the ability of MLR and ANN models to estimate the monthly average GSR in Turkey. They found that the ANN model outperformed the MLR model. Overall, previous studies suggest that the RF, SVR, and ANN models can be used to predict solar radiation. Studies of DGSR prediction mostly used the ε-SVR algorithm [24] and the BPNN algorithm [25]. Backpropagation (BP) is an excellent method that has been widely used for recognizing complex patterns. However, its learning process requires a long time. Resilient BP (Rprop) is a modified version of BP, and it outperforms BPNN in terms of convergence speed, accuracy, and robustness of training parameters [26]. In ε-SVR, a prior of the parameter that represents the regression approximation error must be specified. However, in practice, the error should be minimized. V-support vector regression (V-SVR) is a modified version of ε-SVR that can also serve as a model for SVM to deal with regression issues; it affords the advantage of enabling the automatic adjustment of the width of the ε-tube [27]. In this light, this study compares the DGSR prediction ability of the RF, V-SVR, and Rprop-ANN models.

Most studies of GSR focused on point prediction, but in the field of solar energy, the uncertainty of the prediction of solar radiation is also important. There are many examples of how probabilistic renewable energy predictions (including solar) provide more value than deterministic ones in the management of power systems [28]. The measurement and recording of GSR data inevitably contain random noise, and the noise-related uncertainty makes it difficult to widely apply the point prediction method to solar energy utilization technologies with high risk and low fault tolerance, such as solar photovoltaics. The prediction reliability is important in these applications; in particular, unreliable predictions often result in huge losses. To avoid this problem, the error rate of each prediction should be controllable; interval prediction offers this possibility. Therefore, constructing a reliable prediction interval (PI) for DGSR has great theoretical and practical significance. However, typical approaches for constructing the PI rely too strongly on the model hypothesis or can only guarantee the prediction reliability in an asymptotic case, resulting in low prediction reliability in practical applications. The conformal prediction (CP) framework [29, 30] is a good approach to overcome this problem [31]. CP is a learning method that aims to complement the predictions of traditional learning algorithms with confidence measures, which are called underlying algorithms. A conformal regressor is defined as a combination of a CP and a specific underlying algorithm. It can construct a PI for the inputs of the testing data. Since the CP framework was developed, the reliability of CP has been confirmed by many experiments and applications [32, 33]. Further, CP serves as a practical and effective nonparametric method to construct a PI with high confidence. The out-of-bag (OOB) [34] and split conformal (SC) [35] interval prediction techniques are based on the CP framework. In this study, we improved these two techniques based on kernel density estimation (KDE) to obtain the OOB-KDE and SC-KDE interval prediction algorithms.

First, this study compares and analyzes the applicability of the RF, V-SVR, and Rprop-ANN models for the point prediction of DGSR. Then we improved these two techniques based on kernel density estimation (KDE) to obtain the OOB-KDE and SC-KDE interval prediction algorithms. It is the first time to apply the improved CP framework methods to solar radiation interval prediction. It is also one of the few studies on the uncertainty of solar radiation prediction. The rest of this paper is organized as follows. Section 2 introduces point prediction models and interval prediction techniques. Section 3 includes a brief description of the dataset and evaluation indicator. In section 4, the development of point prediction models and interval prediction methods is presented in detail. Section 5 analyzes the point and interval prediction results. Finally, Section 6 presents the main conclusions drawn from this study.

2. Methodology

2.1. Point Prediction Model

2.1.1. RF. RF [36, 37] is a popular and highly flexible machine learning model that is robust to data containing noisy or missing values. RF combines classification and regression trees (CART) and bagging. CART [38] is an approach to partition a variable space according to a group of rules embedded in a decision tree, in which each node is partitioned based on a decision rule. Each decision tree comprises decision nodes and leaf nodes. The decision node evaluates the input sample through test functions and passes it to different branches according to the characteristics of the sample. Each decision tree eventually forms a multiple nonlinear combination model. Bagging [39] adopts a bootstrap resampling technique to obtain K subtraining sample sets and further obtains K decision trees. These K decision trees are integrated into the RF to improve the prediction accuracy and generalization ability of the model. Eventually, the outputs of the RF model are obtained using the average of the output of K decision trees.

RF regression is very user-friendly in that only two parameters need to be tuned: the number of trees (n_{tree}) and the number of randomly selected features (mtry). Generally, $n_{\text{tree}} = 500$ is sufficient to ensure the prediction accuracy of the model and the convergence of the algorithm. Setting mtry to one-third of the predictor is usually a good option [40]. In this study, n_{tree} is set to 500, and the grid search method is used to tune $mtry$.

2.1.2. V-SVR. SVM is a supervised machine learning model that can be used to solve classification and regression problems. Unlike traditional machine learning methods that use the empirical risk minimization principle, SVMs use the structural risk minimization principle to reduce the upper bound on the generalization error rather than reducing the local training error. V-SVR [41], which is based on the ε-SVR [42] regression model, is an SVM model that is used to deal with fitting regression problems. ε-SVR uses the principle of structural risk minimization to transform a regression problem into a quadratic programming problem by introducing the ε-insensitive loss function and applying the Lagrange multiplier method. In ε-SVR, a priori of the parameter ε that represents the approximation error of regression must be specified; however, in practice, the error must be minimized. V-SVR overcomes the problem of prior selection in ε-SVR by introducing the parameter ν. This study builds the V-SVR model using the *kernlab R* package.

2.1.3. Rprop-ANN. An ANN is an information processing technique that is inspired by the biological nervous system. It uses an input dataset to approximate a potential nonlinear function. The advantages of ANN techniques are that they do not require mathematical computational knowledge between parameters, are less computationally intensive, and can solve multivariate problems easily. A BPNN [43] is a typical feedforward neural network, which is one of the most widely used types of learning models among ANN models. It comprises an input layer, one or more hidden layers, and an output layer. Figure 1 shows the architecture of BPNN with N input variables and M output variables. The weights between the input layer and the hidden layer are denoted as W_{ij}, and the weights between the hidden layer and the output layer are denoted as W_{jk}.

The major drawbacks of a traditional BPNN are the slow learning speed and tendency to fall into a local minimum. Rprop-ANN overcomes these drawbacks and affords optimal results in terms of accuracy, convergence speed, and robustness of the training parameters [44]. For example, when the partial derivatives of the network weights are the same at adjacent moments, Rprop-ANN updates the network weights along the gradient decline direction. To avoid missing a minimum value, when the network weights have different partial derivative symbols at adjacent moments, the current network weights are returned to the previous ones.

2.2. Interval Prediction Method

2.2.1. KDE. The KDE method is one of the most popular nonparametric density estimation techniques [45]. Let X be a random variable with an absolutely continuous distribution function F. Further, let f be the corresponding density function and $\{X_1, \ldots, X_m\}$ be a sample generated by X. Then, the KDE $\widehat{f}(x)$ of $f(x)$ is defined as

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

where $h\,(h > 0)$ is a bandwidth parameter and K is a kernel function. After the kernel function is determined, the population quantile can be estimated by the KDE method. Let $\widehat{F}_h(x)$ be the KDE of the cumulative distribution function $F(x)$. Then, $\widehat{F}_h(x)$ can be derived as follows:

$$\begin{aligned}\widehat{F}_h(x) &= \int_{-\infty}^{x} \widehat{f}(z)dz = (nh)^{-1}\int_{-\infty}^{x}\sum_{i=1}^{n} K\left(\frac{z - X_i}{h}\right)dz \\ &= (nh)^{-1}\sum_{i=1}^{n}\Gamma\left(\frac{z - X_i}{h}\right),\end{aligned} \tag{2}$$

where Γ is the cumulative distribution function of the kernel function K. Then, for $\alpha \in (0, 1)$, the quantile function $\widehat{Q}_h$ based on KDE is $\widehat{Q}_h(\alpha) = \widehat{F}_h^{-1}(x)$. The kernel functions and bandwidth are the parameters that most strongly influence KDE. However, when the sample size is large enough, the selection of the kernel function has little influence on the estimation results [46]. Bandwidth selection methods include the reference method [47, 48] and least-squares cross-validation [49]. This study uses a Gaussian kernel function and the reference method for bandwidth selection; in other words,

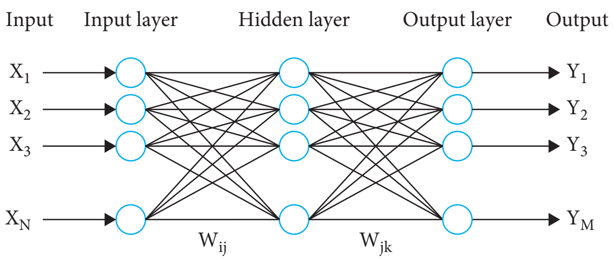


Figure 1: Schematic architecture of BPNN.

$$h = \left(\frac{(2\sqrt{\pi})^{-1}}{1/8\pi^{-1/2}\sigma^{-5}} \right)^{\frac{1}{5}} n^{-\frac{1}{5}} = 1.06\, n^{-\frac{1}{5}}, \tag{3}$$

where σ can be replaced by $\sigma = \min(S, Q/1.34)$ and Q is the interquartile range.

2.2.2. OOB-KDE. The OOB method constructs PIs based on the quantile of the empirical distribution of the prediction error. Suppose that $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a random sample; then, let $(X^{(i)}, Y^{(i)}) = \{(X_j, Y_j | j = 1, \ldots, i-1, i+1, \ldots, n)\}$. For the sample (X_i, Y_i), the OOB method uses $(X^{(i)}, Y^{(i)})$ to build an RF regression model to obtain the prediction of Y_i. Finally, the prediction error for this sample is calculated as $D_i = Y_i - \widehat{Y}_i, i = 1, \cdots, n$, where $\widehat{Y}_i$ is the prediction. Given a new sample (X_{n+1}, Y_{n+1}), the PI of Y_{n+1} constructed using the OOB method is

$$\left[\widehat{Y}_{n+1} + D_{\left[\frac{n,\alpha}{2}\right]}, \widehat{Y}_{n+1} + D_{\left[\frac{n,1-\alpha}{2}\right]} \right], \tag{4}$$

where $D_{[n,\alpha/2]}$ and $D_{[n,1-\alpha/2]}$ are the quantile of the empirical distribution of $[D_1, \ldots, D_n]$ and $\widehat{Y}_{n+1}$ is the prediction of Y_{n+1} obtained using the RF model. The OOB method makes full use of the sample data. More importantly, the coverage of its PI converges to the theoretical interval coverage under regular conditions. Unfortunately, the PI construction method relies on the empirical distribution of the prediction errors; therefore, the calculated quantile is not accurate enough. The KDE method can give an accurate density estimation [50]. The proposed OOB-KDE method is an extension and improvement of the OOB method. For a given regression algorithm $\mathscr{A}$, the OOB-KDE method obtains the prediction error of Y_i using $\mathscr{A}$ and $(X^{(i)}, Y^{(i)})$, estimates the error distribution using the KDE, and eventually constructs the PI of Y_{n+1} using the quantiles $D^*{}_{[n,\alpha/2]}$ and $D^*{}_{[n,1-\alpha/2]}$; that is,

$$\left[\widehat{Y}_{n+1} + D^*{}_{\left[\frac{n,\alpha}{2}\right]}, \widehat{Y}_{n+1} + D^*{}_{\left[\frac{n,1-\alpha}{2}\right]} \right], \tag{5}$$

where $\widehat{Y}_{n+1}$ is the prediction obtained by applying regression algorithm $\mathscr{A}$.

2.2.3. SC-KDE. SC is proposed as an interval prediction method based on the CP framework. Its computational cost is a small fraction of that of the full conformal method. This study improves the SC method based on the KDE method and proposes the improved SC-KDE method. Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a training set; then, the SC-KDE algorithm is summarized as follows:

(1) Randomly split training set into two equal-sized subsets I_1 and I_2

(2) Estimate true mean function using I_1; that is, $\widehat{\mu} = \mathscr{A}\{(X_i, Y_i), i \in I_1\}$

(3) Obtain errors from I_2 using $D_i = Y_i - \widehat{\mu}(X_i), i \in I_2$

(4) Estimate the density of errors using KDE and obtain the quantiles $D^*{}_{[n/2,\alpha/2]}$ and $D^*{}_{[n/2,1-\alpha/2]}$

(5) Then, PI of X_{n+1} is $[\widehat{\mu}(X_{n+1}) + D^*{}_{[n/2,\alpha/2]}, \widehat{\mu}(X_{n+1}) + D^*{}_{[n/2,1-\alpha/2]}]$

SC uses $|\widehat{\mu}(X_i) - Y_i|$ to calculate the errors of I_2, and it uses $[\widehat{\mu}(X_{n+1}) - d, \widehat{\mu}(X_{n+1}) + d]$ to construct the PI, where d is the kth smallest value in the errors and $k = (1-\alpha)(n/2+1)$. The SC method constructs a PI based on the same D value; therefore, it is not applicable when the distribution of prediction errors is skewed. The SC-KDE method can effectively overcome this shortcoming by applying the KDE method and introducing two different quantiles. The SC-KDE method uses $Y_i - \widehat{\mu}(X_i)$ to calculate the errors, and it uses the KDE method to estimate the quantiles $D_{\alpha/2}$ and $D_{1-\alpha/2}$ of the error distribution. Finally, the PI is calculated using $[\widehat{\mu}(X_{n+1}) + D^*{}_{[n/2,\alpha/2]}, \widehat{\mu}(X_{n+1}) + D^*{}_{[n/2,1-\alpha/2]}]$.

3. Dataset and Evaluation Indicator

3.1. Dataset. In this study, DGSR and daily meteorological variable observation data from Hami station for January 2009 to December 2016 were used. Hami (93°31′E, 42°49′N) is located in the eastern part of Xinjiang Uyghur Autonomous Region. It has a typical temperate continental arid climate, with many sunny days throughout the year and an average annual sunshine duration of 3358 h, making it one of the regions with the longest sunshine duration in China. The meteorological variables selected in this study include the average relative humidity (RHU), daily average temperature (AT), and daily sunshine duration (SD). The data used undergoes strict revision and quality control to ensure that the data collection time is continuous and complete and that the data availability and accuracy are close to 100%. When the data are collected, the whole dataset is divided into three categories: training set, validation set, and test set. All data used in this study were obtained from the National Meteorological Information Center (see Appendix for further details of the data).

3.2. Evaluation Indicator. In this study, the mean absolute error (MAE), mean relative error (MRE), and Kendall rank correlation (Kendall) are used to evaluate the prediction performance of the RF, V-SVR, and Rprop-ANN models.

The calculation of MAE, MRE, and Kendall is detailed in [51, 52]. The quality of PIs is quantitatively evaluated using the prediction interval coverage probability (PICP), prediction interval normalized average width (PINAW), and coverage width criteria (CWC). PICP and PINAW are used to evaluate the reliability and accuracy of the PI, respectively. A small PINAW value indicates the high accuracy of the PI. However, owing to the contradiction between PICP and PINAW, one of them alone cannot be used to effectively judge the quality of PI. Therefore, CWC is introduced as a tradeoff between PICP and PINAW. CWC is expected to be small for good solutions. The calculation of PICP, PINAW, and CWC is detailed in [53]. Penalty factors in CWC generally take larger values such as 50 [54]. In the present study, the penalty factor was set as 25 in consideration of the magnitude of the CWC value.

4. Model Development

4.1. Model Input and Feature Selection. The dataset selected in this study includes DGSR and three meteorological variables. In order to acquire more accurate predictions of DGSR and examine the influence of historical DGSR on the performance of point prediction models, the two different model inputs are considered. One uses three meteorological variables as model inputs. Another uses the lags of DGSR and meteorological variables as a new model input. For simplicity of representation, the two model inputs are denoted as Input I and Input II. And the combination of the data with respect to Input I and Input II is denoted as original data (OD) and feature selection data (FSD), respectively. The feature selection method to determine the order of lags of the DGSR in Input II is the mean decrease Gini index (MDGI) [55] calculated by RF. Figure 2 is the result of feature selection expressed as a percentage of MDGI. The ranking of the MDGI shows that the first two lags accounted for 83% of all lags and were significantly higher than the other lags. Therefore, the historical information of the DGSR for the first two days is selected to construct the new inputs.

4.2. Point Prediction Model Development. In the present study, 40% of the dataset was used to construct models, 10% was used to tune the parameters of the models, and the remaining 50% was used to test the prediction performance of the models. $mtry$ was used as the tuning parameter for the RF model, and kernel functions were used for the V-SVR model. This study considers the radial basis kernel, polynomial kernel, Laplace kernel, and linear kernel. The tuning parameters of Rprop-ANN include the number of neurons in the hidden layer (Hidden), activation function in the output layer (Active), and threshold (Threshold) used to specify when the model stops training. The optimal parameters of each model are searched by using grid search with the criterion of the minimum MAE value. Figure 3 shows the flowchart of the point prediction of DGSR. Table 1 shows the results of tuning parameters of each machine learning model.

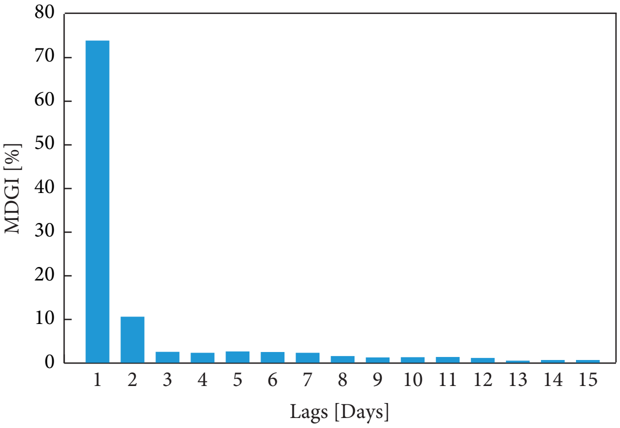


Figure 2: MDGI ranks of 10 lags.

4.3. Interval Prediction Method Development. To examine the performance of the proposed interval prediction technique, three different values for the prediction interval nominal coverage (PINC) have been used: 0.8, 0.9, and 0.95. Because the OOB and OOB-KDE methods do not specify the training set ratio (TSR), three different TSR are used: 0.2, 0.5, and 0.8. Considering the limitations of the SC and SC-KDE methods regarding the data segmentation ratio, only the case of an equal segmentation ratio is considered in both methods. During the construction of the PI, the initial value selection of the Rprop-ANN model is seen to influence the constructed PI. To overcome this problem, we construct PIs of the DGSR five times and use their average as the final metric result. Figure 4 shows the flowchart of the interval prediction of the DGSR.

5. Results and Discussion

5.1. Point Prediction. Table 2 illustrates the three machine learning models results.

First, we compare the prediction performance of three machine learning models in the OD. As we can see from Table 2, the V-SVR has the best capability of predicting the DGSR. The statistical indicator of V-SVR used for performance evaluation indicates the lowest values of MAE and MRE and highest values of Kendall when compared to RF and Rprop-ANN model. The order of magnitude of MAE and MRE for the V-SVR model is around one-tenth of those for the RF and Rprop-ANN models. Both Rprop-ANN and V-SVR provide better results compared to the RF model.

Next, we consider the prediction performance of three models for the FSD. According to the table, the prediction performance of the FSD for each model is similar to the OD. The model performance decreases in the order of the V-SVR model, Rprop-ANN model, and RF model.

Finally, we compare the ability of the OD and FSD to predict the DGSR. Compared with the OD, the prediction accuracy of the models constructed using the FSD is improved significantly. For the RF, V-SVR, and Rprop-ANN models, the MAE reduced by around 33%, 36%, and 52%, respectively, and the MRE reduced by around 33%, 50%, and 50%, respectively. Further, Kendall of the RF and Rprop-

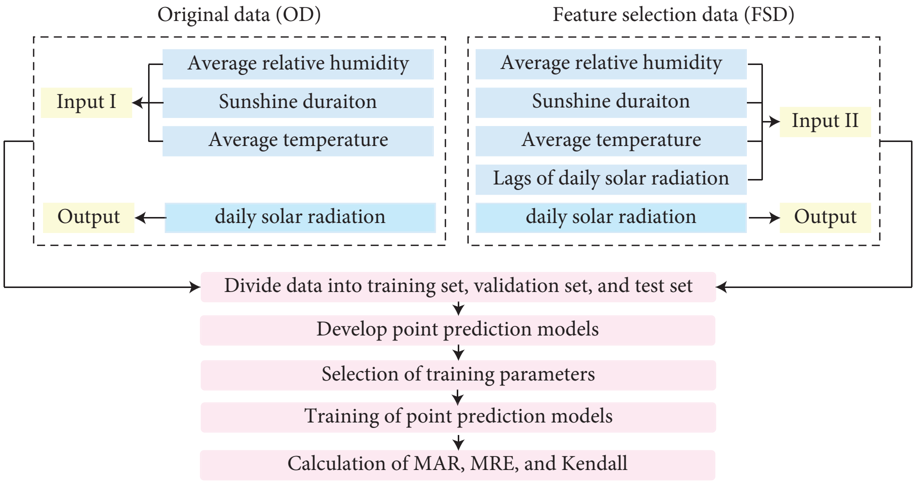

Original data (OD)
Feature selection data (FSD)
Input I
Average relative humidity
Sunshine duraiton
Average temperature
Output
daily solar radiation
Average relative humidity
Sunshine duraiton
Average temperature
Lags of daily solar radiation
daily solar radiation
Input II
Output
Divide data into training set, validation set, and test set
Develop point prediction models
Selection of training parameters
Training of point prediction models
Calculation of MAR, MRE, and Kendall

Figure 3: Flowchart of point prediction of DGSR.

Table 1: Tuning parameters of each model.

Model	Tuning parameters	OD	FSD
RF	*mtry*	3	5
V-SVR	Kernel	Polynomial	Linear
Rprop-ANN	Hidden	1	1
	Threshold	0.01	0.01
	Active	Tanh	Tanh

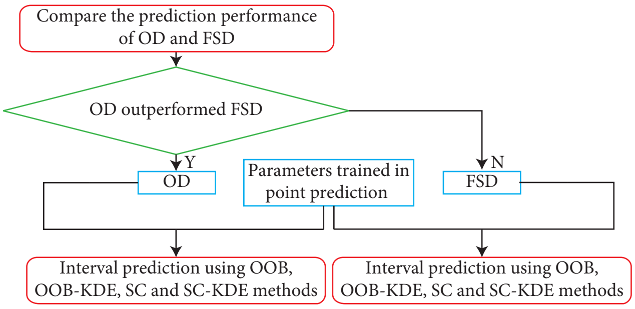

Compare the prediction performance of OD and FSD
OD outperformed FSD
Y
N
OD
Parameters trained in point prediction
FSD
Interval prediction using OOB, OOB-KDE, SC and SC-KDE methods
Interval prediction using OOB, OOB-KDE, SC and SC-KDE methods

Figure 4: Flowchart of interval prediction of DGSR.

Table 2: Error statistics of different models for DGSR prediction.

	OD			FSD		
	RF	V-SVR	Rprop-ANN	RF	V-SVR	Rprop-ANN
MAE	0.365	0.025	0.121	0.245	0.016	0.058
MRE	0.015	0.002	0.014	0.010	0.001	0.007
Kendall	0.986	1	0.997	0.993	1	0.999

ANN models is higher than that of the OD. These analyses clearly indicate that FSD is more powerful in explaining the variation mechanism of DGSR and provides higher prediction accuracy.

To compare the predictive capability of each model explicitly and visually on the OD and FSD, Figure 5 is a Taylor diagram comparing predicted DGSR from RF, V-SVR, and Rprop-ANN; Figure 6 illustrates the comparisons of the predicted DGSR (red) verse measured data (blue) for the RF, V-SVR, and Rprop-ANN models in 2016. The error of V-SVR and Rprop-ANN is lower than RF, and the correlation of V-SVR and Rprop-ANN is higher than RF in Figure 5. This proves that the V-SVR and Rprop-ANN models perform much better than the RF model. The predicted values of SVR and ANN models follow the measured data favorably in Figure 6, indicating that the prediction accuracy is high. RF model can effectively reflect DGSR fluctuations in spring and winter but cannot fit the DGSR in summer and autumn. At the same time, it can be seen that the FSD improves the fitting performance of the RF model to a certain extent from both Figures 5 and 6.

Figure 7 shows scatter plots of the predicted DGSR verse measured data as well as their MAE for all testing phases. All scatter dots of the V-SVR and Rprop-ANN models are seen to be close to the red line that represents $y = x$, indicating that these models accurately predict not only the test samples for 2016 but also the complete test set. The scatter plots of MAE show that the V-SVR model provides the highest prediction accuracy. The scatter plot of the RF model clearly indicates that it underestimates the DGSR when the real DGSR is large.

The annual mean value of variables in the test set (Table 3) shows that the annual mean value of DGSR in 2016 increased significantly by around 22%. However, other variables did not increase significantly, indicating that, for RF models, the existing variables are insufficient to explain the sudden increase of DGSR in 2016.

5.2. Interval Prediction. In this section, we present and compare the results of PIs. Table 4 shows the PICP, PINAW, and CWC of PIs obtained by OOB and OOB-KDE for different values of PINC and for different TSR. The PICP of the PIs of the RF and V-SVR models obtained using the OOB-KDE method is higher than that obtained using the OOB method for all combinations of PINC and TSR. The PICP of the Rprop-ANN model obtained using the OOB is higher than that obtained using the OOB-KDE under only a few conditions. These results proved that the OOB-KDE

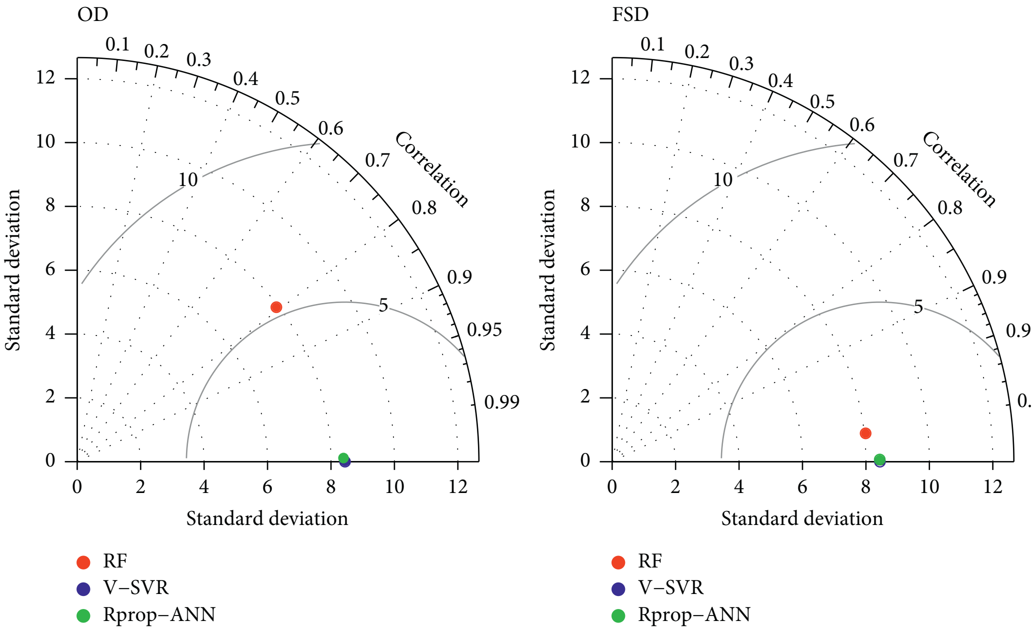
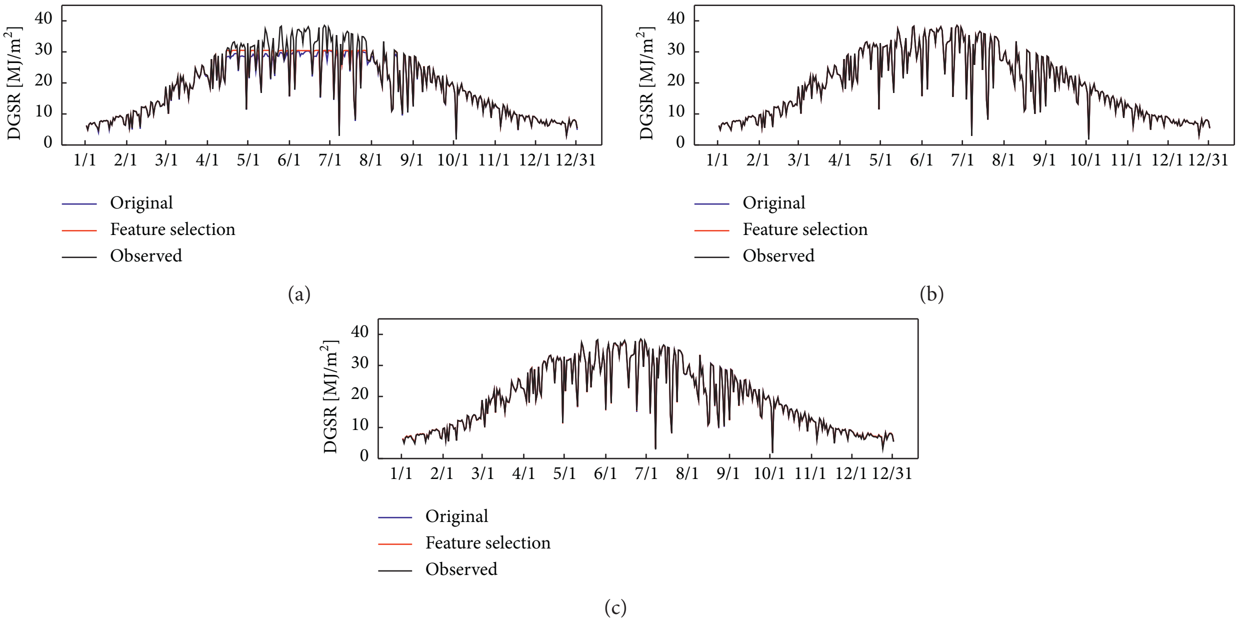


Figure 5: Taylor diagram comparing predicted DGSR from RF, V-SVR, and Rprop-ANN on OD and FSD.



(a)

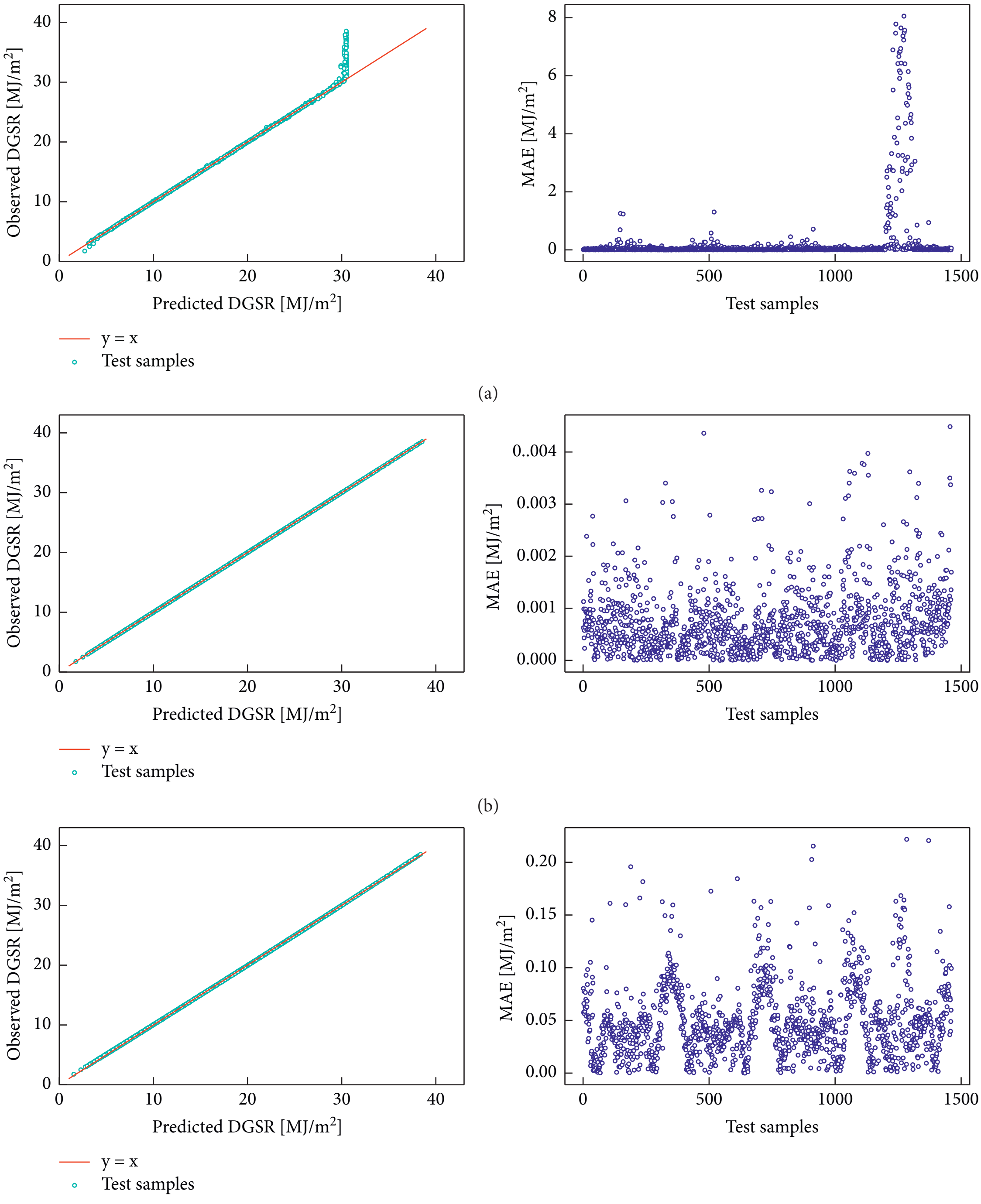
(b)

(c)

Figure 6: Fitting curve plot between measured and predicted DGSR.

method effectively improves the reliability of the PI. For all TSR and for all PINC values, the PINAW of all PIs obtained using the OOB-KDE method is higher than that obtained using the OOB method. However, it is caused by the inherent contradiction between PICP and PINAW. The composite metric CWC confirms that the PI of the OOB-KDE method was significantly better than that of the OOB method, with an average CWC reduction of approximately 81%.

To further explore the quality of the PIs obtained using the OOB-KDE method for different TSR, the variations of the PICP, PINAW, and CWC are shown in Figures 8–10, respectively. Figure 8 shows that the PICPs of the RF and V-SVR models have an obvious positive correlation with the PINC, indicating that the PI is reasonable to a certain extent. The PICP of the Rprop-ANN model does not show a linear trend related to the PINC owing to the mechanism used to generate the results. Figure 9 shows that the acuity of the PIs is optimal at a training set ratio of 0.2 and worst at a TSR of 0.8. Figure 10 shows that the quality of the PI obtained using the OOB-KDE method for the RF and Rprop-ANN models is negatively correlated with the TSR, and the optimal PI for



(a)

(b)

(c)

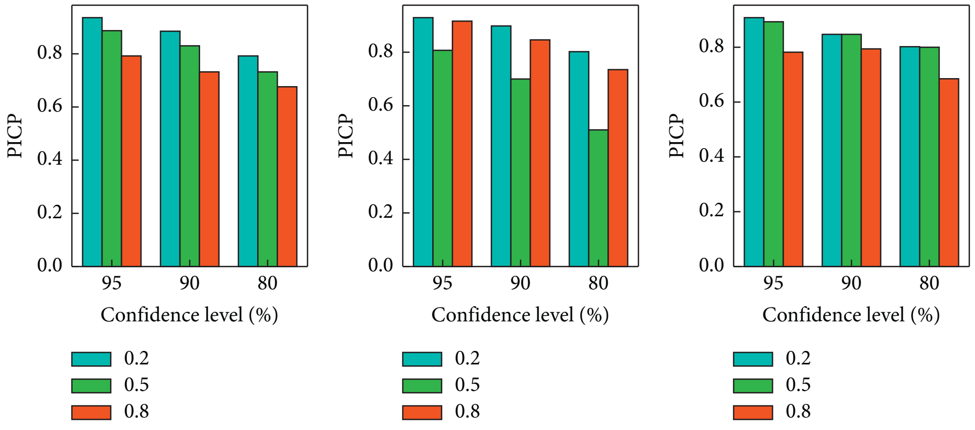
Figure 7: Scatter plots and MAEs of test samples.

Table 3: Annual mean values of variables in test set.

Year	AT (°C)	RHU (%)	SD (h)	DGSR (MJ/m^2)
2013	10.81	42.56	9.75	16.33
2014	10.38	40.45	9.79	16.24
2015	11.47	45.08	9.22	15.55
2016	11.66	44.23	9.08	19.50

Table 4: Comparison of PIs obtained using OOB and OOB-KDE methods.

			0.95			0.9			0.8		
TSR	Method	Model	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC
0.2	OOB	RF	0.890	0.297	1.628	0.771	0.168	4.394	0.662	0.118	3.835
		V-SVR	0.919	0.005	0.016	0.880	0.004	0.011	0.770	0.003	0.009
		Rprop-ANN	0.903	0.754	3.196	0.840	0.583	3.196	0.746	0.413	2.006
	OOB-KDE	RF	0.936	0.415	1.004	0.885	0.282	0.692	0.792	0.179	0.398
		V-SVR	0.929	0.006	0.016	0.898	0.005	0.01	0.820	0.004	0.006
		Rprop-ANN	0.908	0.790	3.048	0.893	0.610	1.337	0.782	0.446	1.145
0.5	OOB	RF	0.823	0.134	3.34	0.771	0.110	2.877	0.654	0.063	2.487
		V-SVR	0.769	0.003	0.28	0.675	0.003	0.835	0.492	0.002	4.419
		Rprop-ANN	0.739	0.153	30.048	0.809	0.336	3.605	0.854	0.242	0.305
	OOB-KDE	RF	0.887	0.182	1.061	0.830	0.118	0.797	0.732	0.078	0.505
		V-SVR	0.807	0.004	0.147	0.700	0.003	0.448	0.510	0.003	4.227
		Rprop-ANN	0.872	0.464	3.725	0.847	0.352	1.676	0.794	0.248	0.536
0.8	OOB	RF	0.810	0.112	3.821	0.764	0.076	2.353	0.550	0.039	20.242
		V-SVR	0.897	0.004	0.019	0.824	0.003	0.023	0.726	0.003	0.022
		Rprop-ANN	0.801	0.327	13.888	0.811	0.247	2.533	0.651	0.181	7.687
	OOB-KDE	RF	0.812	0.115	3.738	0.767	0.080	2.304	0.676	0.050	1.16
		V-SVR	0.916	0.004	0.013	0.846	0.003	0.015	0.735	0.003	0.018
		Rprop-ANN	0.750	0.305	45.571	0.802	0.245	3.084	0.685	0.174	3.258

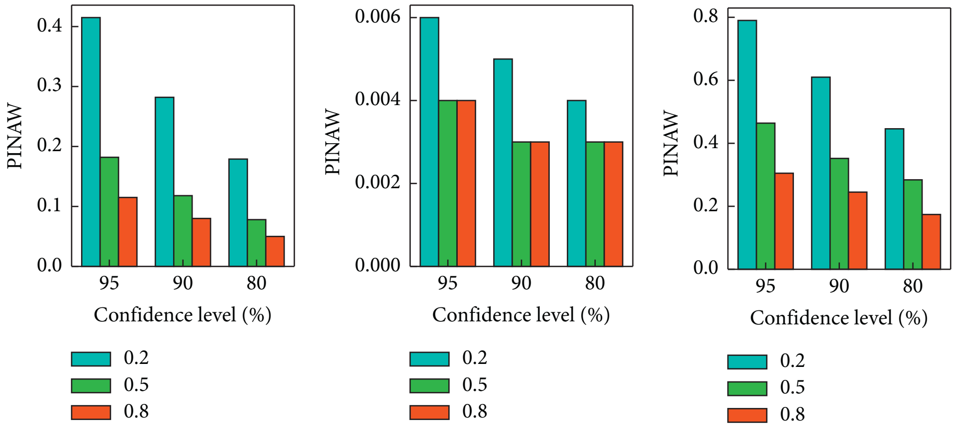


(a)

(b)

(c)

Figure 8: Histogram of PICP for all prediction models.

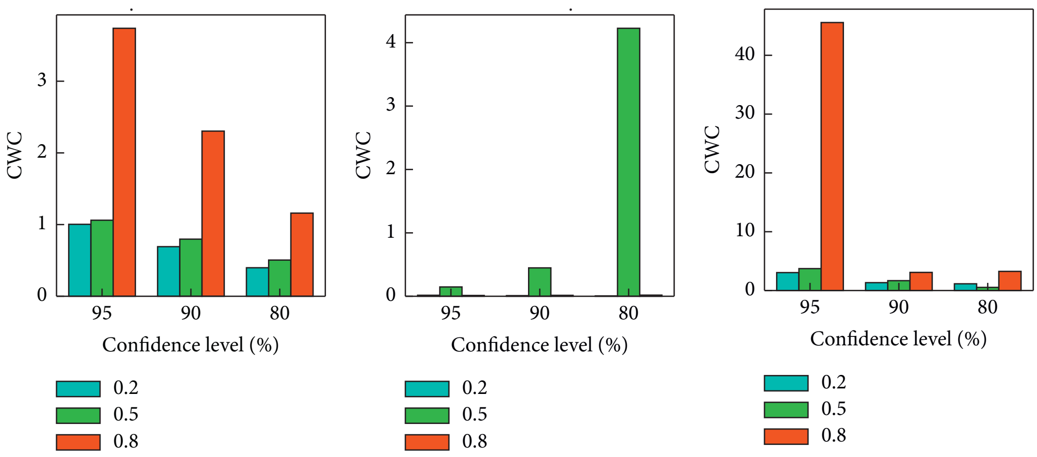


(a)

(b)

(c)

Figure 9: Histogram of PINAW for all prediction models.



(a) (b) (c)

Figure 10: Histogram of CWC for all prediction models.

Table 5: Comparison of PIs obtained using SC and SC-KDE methods.

		0.95			0.9			0.8		
Method	Model	PICP	PINAW	CWC	PICP	PINAW	CWC	PICP	PINAW	CWC
	RF	0.476	0.039	5463.329	0.451	0.038	2849.344	0.401	0.031	665.993
SC	V-SVR	0.476	0.002	280.171	0.451	0	22.495	0.400	0.002	44.055
	Rprop-ANN	0.476	0.020	2801.707	0.451	0.110	8248.102	0.400	0.137	3017.763
	RF	0.950	0.254	0.508	0.901	0.153	0.302	0.808	0.104	0.189
SC-KDE	V-SVR	0.956	0.002	0.004	0.908	0.004	0.007	0.814	0.003	0.005
	Rprop-ANN	0.953	0.012	0.023	0.920	0.015	0.024	0.806	0.001	0.002

each model mostly occurs at a TSR of 0.2. Notably, the V-SVR model performs best in terms of the acuity of the PI; its PINAW values are all below 0.006 (Table 5).

Table 5 The indicators of PIs as obtained using the SC and SC-KDE methods. These results indicate that the PI obtained using the SC method has very poor reliability, and its PICP is approximately 44% lower than the target PINC. Compared with the PI obtained using the SC method, that obtained using the SC-KDE method is greatly improved. PICP increases by an average of 1.01 times, the order of magnitude of CWC is significantly reduced, and the PINAW of the Rprop-ANN model decreases by around 75% on average. In addition, the PICP of the PI obtained using the SC-KDE method is quite close to the target PINC with an average difference of around 0.007, indicating that the SC-KDE method produces a highly reliable PI. In terms of the PIs obtained using the SC-KDE method, the V-SVR model is found to provide the best results, with the highest PINAW and CWC values of 0.007 being far lower than those of the other two machine learning models. It not only ensures the reliability of the PI but also achieves the highest accuracy.

6. Conclusion

In the case of solar radiation prediction, the machine learning model has excellent performance. RF, V- SVR, and Rprop-ANN are the most popular machine learning models. Most of the previous research on solar radiation prediction focused on point prediction. However, the uncertainty of the prediction of solar radiation is also important. Many examples prove that probabilistic renewable energy predictions (including solar) provide more value than deterministic ones in the management of power systems. In this study, the RF, V-SVR, and Rprop-ANN models are used to predict the DGSR and to compare the ability of the OD and FSD to explain its variation mechanism. In addition, the improved OOB-KDE and SC-KDE interval prediction methods are proposed for predicting the interval of GSR. Comparisons are performed with the OOB and SC methods. To demonstrate the performance of the PIs obtained using the OOB-KDE and SC-KDE methods, various combinations of PINC and TSR are considered. The following conclusions can be drawn from this study:

(1) The point prediction results worsen in the following order: V-SVR model, Rprop-ANN model, and RF model. The V-SVR and Rprop-ANN models can accurately describe DGSR fluctuations; however, the RF model shows poor-fitting ability in the summer and autumn.

(2) Compared with the OD, the FSD is more beneficial for explaining the change mechanism of DGSR. For the RF, V-SVR, and Rprop-ANN models, the MAE decreased by 33%, 33%, and 36%, respectively, and the MRE decreased by 33%, 50%, and 50%, respectively.

(3) The OOB-KDE and SC-KDE methods improved the quality of the PI. The CWC of the PIs obtained using the OOB-KDE method is around 81% lower than that obtained using the OOB method, and the CWC of the PIs obtained using the SC-KDE method is around 99.99% lower than that obtained using the SC method.

(4) The best interval prediction result is obtained using the SC-KDE method with the V-SVR model. The average difference between its PICP and PINC is only 3% of the PINC, and its PINAW is less than 0.007.

The FSD shows better performance for the variation mechanism of the DGSR because DGSR data contains strong time-lag effects and is correlated with historical data. The proposed interval prediction technique improves the performance of the PI by estimating the actual error distribution using the KDE method. In conclusion, this study realizes accurate point prediction of the DGSR using the V-SVR model and realizes highly reliable interval prediction of the DGSR using the SC-KDE method. Our study contributes greatly to the research and modeling of point and interval prediction of solar radiation. And it will be a benefit to the efficient and full use of solar energy. In the future, more meteorological variables should be used to predict solar radiation. Our study only considers three machine learning models; therefore, the consideration of more machine learning models is also a direction for future research. As mentioned above, probabilistic renewable energy predictions provide more value in the management of power systems. In view of this, the applicability of the OOB-KDE and SC-KDE methods in other renewable energies' interval prediction fields can be explored.

Data Availability

The daily global solar radiation data and daily meteorological variable observation data used in this study were obtained from the National Meteorological Information Center of China. Requests for access to these data should be made to Ysuhuzhang (e-mail: zhang_yuhu@163.com). The dataset used in this paper can be downloaded at https://github.com/zmy987/solar-radiation-dataset.git and can be seen in https://github.com/zmy987/solar-radiation dataset/blob/main/solar%20radiation.csv.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (Z210003), National Nature Science Foundation of China (Grant no. 12171328), the State Key Program of National Natural Science Foundation of China (Grant no. 12031016), the Beijing Talent Foundation Outstanding Young Individual Project, and Academy for Multidisciplinary Studies Academy for Multidisciplinary Studies of Capital Normal University. Yuhu Zhang's study was supported by the Capital Normal University Multidisciplinary Studies Project, China (00719530012012 and 00719530012010).

References

[1] A. E. Gürel, M. Abulut, and Y. Bien, "Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation," *Journal of Cleaner Production*, vol. 277, Article ID 122353, 2020.

[2] S. Shamshirband, T. Rabczuk, and K.-W. Chau, "A survey of deep learning techniques: application in wind and solar energy resources," *IEEE Access*, vol. 7, pp. 164650–164666, 2019.

[3] D. Guijo-Rubio, A. M. Durán-Rosal, P. A. Gutiérrez et al., "Evolutionary artificial neural networks for accurate solar radiation prediction," *Energy*, vol. 210, Article ID 118374, 2020.

[4] Z. E. Mohamed, "Using the artificial neural networks for prediction and validating solar radiation," *Journal of the Egyptian Mathematical Society*, vol. 27, no. 1, p. 47, 2019.

[5] R. Meenal and A. I. Selvakumar, "Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters," *Renewable Energy*, vol. 121, pp. 324–343, 2018.

[6] K. Namrata, S. P. Sharma, and S. B. L. Seksena, "Empirical models for the estimation of global solar radiation with sunshine hours on horizontal surface for Jharkhand (India)," *Applied Solar Energy*, vol. 52, no. 3, pp. 164–172, 2016.

[7] A. Janaki and P. K. Narayan, "Estimation of global solar radiation using empirical model on meteorological parameters at simara airport, bara, Nepal," *Journal of the Institute of Engineering*, vol. 14, no. 1, pp. 143–150, 2018.

[8] J. Fan, L. Wu, F. Zhang et al., "Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China," *Renewable and Sustainable Energy Reviews*, vol. 100, pp. 186–212, 2019.

[9] A. Morenomunoz, J. D. L. Rosa, R. Posadillo, and V. Pallares, "Short term forecasting of solar radiation," in *Proceedings of the 2008 33rd IEEE Photovoltaic Specialists Conference*, San Diego, CA, USA, May 2008.

[10] G. Reikard, "Predicting solar radiation at high resolutions: a comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.

[11] J. Piri, S. Shamshirband, D. Petkovic, C. W. Tong, and M. H. U. Rehman, "Prediction of the solar radiation on the Earth using support vector regression technique," *Infrared Physics & Technology*, vol. 68, pp. 179–185, 2014.

[12] R. C. Deo and M. Şahin, "Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 828–848, 2017.

[13] T. R. Ayodele, A. S. O. Ogunjuyigbe, A. Amedu, and J. L. Munda, "Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms," *Renewable Energy Focus*, vol. 29, no. JUN, pp. 78–93, 2019.

[14] H. Sun, D. Gui, B. Yan et al., "Assessing the potential of random forest method for estimating solar radiation using air pollution index," *Energy Conversion and Management*, vol. 119, pp. 121–129, 2016.

[15] Y. Cao, A. Raise, A. Mohammadzadeh, S. Rathinasamy, S. S. Band, and A. Mosavi, "Deep learned recurrent type-3 fuzzy system: application for renewable energy modeling/ prediction," *Energy Reports*, vol. 7, pp. 8115–8127, 2021.

[16] E. C. Akbaba, E. Yce, and B. G. Akinoglu, "Deep learning algorithm applied to daily solar irradiation estimations," in *Proceedings of the 2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pp. 1–4, Rabat, Morocco, December 2018.

[17] M. Zakroum, M. Ghogho, M. Faqir, and M. A. Ahajjam, "Deep Learning for Inferring the Surface Solar Irradiance from Sky Imagery," 2018, https://arxiv.org/abs/1812.09793.

[18] G. Zhang, S. S. Band, C. Jun et al., "Solar radiation estimation in different climates with meteorological variables using Bayesian model averaging and new soft computing models," *Energy Reports*, vol. 7, pp. 8973–8996, 2021.

[19] M. Hou, T. Zhang, F. Weng, M. Ali, N. Al-Ansari, and Z. Yaseen, "Global solar radiation prediction using hybrid online sequential extreme learning machine model," *Energies*, vol. 11, no. 12, p. 3415, 2018.

[20] N. Samuel Chukwujindu, "A comprehensive review of empirical models for estimating global solar radiation in Africa," *Renewable and Sustainable Energy Reviews*, vol. 78, pp. 955–995, 2017.

[21] Z. Zeng, Z. Wang, K. Gui et al., "Daily global solar radiation in China estimated from high-density meteorological observations: a random forest model framework," *Earth and Space Science*, vol. 7, 2020.

[22] J. Fan, W. Xiukang, F. Zhang, X. Ma, and L. Wu, "Predicting daily diffuse horizontal solar radiation in various climatic rons of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data," *Journal of Cleaner Production*, vol. 248, pp. 119–264, 2019.

[23] M. Sahin, Y. Kaya, and M. Uyar, "Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data," *Advances in Space Research*, vol. 51, no. 5, pp. 891–904, 2013.

[24] S. Belaid and A. Mellit, "Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate," *Energy Conversion and Management*, vol. 118, pp. 105–118, 2016.

[25] A. Marzo, M. Trigo-Gonzalez, J. Alonso-Montesinos et al., "Daily global solar radiation estimation in desert areas using daily extreme temperatures and extraterrestrial radiation," *Renewable Energy*, vol. 113, no. dec, pp. 303–311, 2017.

[26] C.-S. Chen and J.-M. Lin, "Applying Rprop neural network for the prediction of the mobile station location," *Sensors*, vol. 11, no. 4, pp. 4207–4230, 2011.

[27] L. Hu and X.-L. Che, "A nu-support vector regression based system for grid resource monitoring and prediction," *Acta Automatica Sinica*, vol. 36, no. 1, pp. 139–146, 2010.

[28] I. M. Galvn, J. Huertas-Tato, F. J. Rodrguez-Bentez, C. Arbizu-Barrena, D. Pozo-Vzquez, and R. Aler, "Evolutionary- based prediction interval estimation by blending solar radiation forecasting models using meteorological weather types," *Applied Soft Computing*, vol. 109, p. 107531, 2021.

[29] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer-Verlag, New York, NY, USA, 2005.

[30] V. Balasubramanian, S. Ho, and V. Vovk, *Conformal Prediction for Reliable Machine Learning: Theory, Adap- Tations and Applications*, Elsevier, Amsterdam, 2014.

[31] C. Kath and F. Ziel, "Conformal prediction interval estimations with an application to day-ahead and intraday power markets," *International Journal of Forecasting*, 2020.

[32] H. Papadopoulos, "Cross-conformal prediction with ridge regression," *Statistical Learning and Data Sciences*, Springer, in *Proceedings of the International Symposium on Statistical Learning and Data Sciences*, pp. 260–270, April 2015.

[33] H. Boström, H. Linusson, T. Löfström, and U. Johansson, *Evaluation of a Variance-Based Nonconformity Measure for Regression Forests*, Springer International Publishing, New York, NY, USA, pp. 75–89, 2016.

[34] H. Zhang, J. Zimmerman, D. Nettleton, and D. J. Nordman, "Random forest prediction intervals," *American Statian*, vol. 74, pp. 1–20, 2019.

[35] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] L. Breiman, "Statistical modeling: the two cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199–215, 2001.

[38] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees (CART)*, Wadsworth International Group, New York, NY, USA, 1984.

[39] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[40] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 23, no. 23, pp. 12–18, 2002.

[41] B. Schlkopf, P. Bartlett, A. Smola, and R. Williamson, "Support vector regression with automatic accuracy control," in *Proceedings of the 8th International Conference on Artificial Neural Network*, pp. 111–116, Skovde, Sweden, September 1998.

[42] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[44] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 586–591, San Francisco, CA, USA, March 1993.

[45] K. Taaffe, B. Pearce, and G. Ritchie, "Using kernel density estimation to model surgical procedure duration," *International Transactions in Operational Research*, vol. 28, no. 1, pp. 401–418, 2018.

[46] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability and Its Applications*, vol. 14, no. 1, pp. 153–158, 1969.

[47] D. W. Scott and S. R. Sain, "Multidimensional density estimation," *Handbook of Statistics*, vol. 24, no. 24, pp. 229–261, 2005.

[48] R. J. Karunamuni and T. Alberts, "On boundary correction in kernel density estimation," *Statistical Methodology*, vol. 2, no. 3, pp. 191–212, 2005.

[49] A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, no. 32, pp. 353–360, 1984.

[50] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, Hoboken, NJ, USA, 1992.

[51] Y. Zhang, W. Li, Q. Chen, X. Pu, and L. Xiang, "Multi-models for SPI drought forecasting in the north of haihe river basin, China," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 10, pp. 2471–2481, 2017.

[52] D. Qiu, T. Wang, Q. Ye et al., "A deformation prediction approach for supertall building using sensor monitoring system," *Journal of Sensors*, vol. 2019, no. 29, 12 pages, Article ID 9283584, 2019.

[53] F. Liu, C. Li, Y. Xu, G. Tang, and Y. Xie, *A New Lower and Upper Bound Estimation Model Using Gradient Descend Training Method for Wind Speed Interval Prediction*, Wind Energy, Oklahoma,OK,USA, 2020.

[54] H. Liu, *Time Series Hybrid Intelligence Identification, Modelling and Prediction*Science Press, Beijing, China, 2020.

[55] A. Liaw and M. Wiener, "Classification and regression by random forest," *Forest@*, vol. 23, pp. 18–22, 2001.