frontiers in Earth Science

ORIGINAL RESEARCH
published: 30 April 2021
doi: 10.3389/feart.2021.596860

Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events

Liexing Huang[1,2†], Junfeng Kang[2,4], Mengxue Wan[5,6], Lei Fang[7], Chunyan Zhang[8] and Zhaoliang Zeng[3*†]*

[1] Ganzhou National Territory Spactial Investigation and Planning Research Center, Ganzhou, China, [2] School of Civil and Surveying and Mapping, Jiangxi University of Science and Technology, Ganzhou, China, [3] Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan, China, [4] Department of Geography, University of Connecticut, Storrs, CT, United States, [5] State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing, China, [6] National Joint Research Center for Yangtze River Conservation, Beijing, China, [7] Department of Environmental Science and Engineering, Fudan University, Shanghai, China, [8] Chongqing Wanzhou District Planning and Design Institute, Chongqing, China

OPEN ACCESS

Edited by:
Hiroyuki Murakami,
University Corporation for Atmospheric Research (UCAR), United States

Reviewed by:
Yuanjian Yang,
Nanjing University of Information Science and Technology, China
Yong Xu,
Guangzhou University, China
S. Shamshirband,
Ton Duc Thang University, Vietnam

***Correspondence:**
Junfeng Kang
junfeng.kang@jxust.edu.cn
Zhaoliang Zeng
zhaoliang.zeng@whu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Earth Science

Received: 21 August 2020
Accepted: 04 March 2021
Published: 30 April 2021

Citation:
Huang L, Kang J, Wan M, Fang L, Zhang C and Zeng Z (2021) Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events. Front. Earth Sci. 9:596860. doi: 10.3389/feart.2021.596860

Solar radiation is the Earth's primary source of energy and has an important role in the surface radiation balance, hydrological cycles, vegetation photosynthesis, and weather and climate extremes. The accurate prediction of solar radiation is therefore very important in both the solar industry and climate research. We constructed 12 machine learning models to predict and compare daily and monthly values of solar radiation and a stacking model using the best of these algorithms were developed to predict solar radiation. The results show that meteorological factors (such as sunshine duration, land surface temperature, and visibility) are crucial in the machine learning models. Trend analysis between extreme land surface temperatures and the amount of solar radiation showed the importance of solar radiation in compound extreme climate events. The gradient boosting regression tree (GBRT), extreme gradient lifting (XGBoost), Gaussian process regression (GPR), and random forest models performed better (poor) prediction capabilities of daily and monthly solar radiation. The stacking model, which included the GBRT, XGBoost, GPR, and random forest models, performed better than the single models in the prediction of daily solar radiation but showed no advantage over the XGBoost model in the prediction of the monthly solar radiation. We conclude that the stacking model and the XGBoost model are the best models to predict solar radiation.

Keywords: solar radiation prediction, meteorological factors, machine learning, stacking model, climate extremes model comparison

INTRODUCTION

Solar radiation is the Earth's main source of energy and the amount of solar radiation reaching the Earth's surface is affected by the atmosphere, hydrosphere and biosphere (Budyko, 1969; Islam et al., 2009). Solar radiation also has a vital role in the global climate, and even small changes in the output of energy from the Sun will cause considerable changes in the Earth's climate (Beer et al., 2010; Siingh et al., 2011). Variations

in solar radiation affect global temperatures, global mean sea-level, and compound extreme climate events (Bhargawa and Singh, 2019). Accurate observations and analyses of the temporal and spatial variability of solar radiation are therefore essential in research on solar energy, building materials, and extreme weather and climate events (Garland et al., 1990; Cline et al., 1998; Hoogenboom, 2000; Grant and Tuohimaa, 2004; Wild, 2009; Beer et al., 2010; Besharat et al., 2013; Ohunakin et al., 2015). Many methods have been developed to predict solar radiation, including theoretical parameter models, empirical models, artificial intelligence models, and satellite retrieval data (Iziomon and Mayer, 2002; Mellit, 2008; Lu et al., 2011; Li et al., 2012; Halabi et al., 2018; Makade et al., 2019). Angstrom (1924) and Prescott (1940) first proposed the A–P model, which is widely used to predict solar radiation. Bristow and Campbell (1984) constructed the BCM model by analyzing the relationship between solar radiation and daily maximum and minimum temperatures. Yang et al. (2001) developed a hybrid model (YHM), improving the A–P model by exploring the effects of meteorological parameters and then validating the model's accuracy in Japan. Salazar (2011) compared the YHM and a climatological solar radiation model to estimate the horizontal direct and diffuse components of solar radiation to generate a corrected version of the YHM (CYHM). Gueymard, 2003 selected 19 solar radiation models to investigate solar irradiance predictions, concluding that detailed transmittance models perform better than bulk models. The development of machine learning has inspired many researchers to use machine learning algorithms to develop solar radiation prediction models (Azadeh et al., 2009; Jiang, 2009; Chen et al., 2011; Voyant et al., 2012). Fadare (2009) and Linares-Rodríguez et al. (2011) adopted artificial neural network (ANN) technology to construct solar radiation prediction models to test their predictive ability. Xue (2017) used a back-propagation algorithm to develop a solar radiation prediction model and showed that the predictive accuracy depended on the combination and configuration of the input parameters. Chen et al. (2011) used the support vector machine (SVM) method to construct a solar radiation prediction model and showed that the SVM-based algorithm had a differential predictive accuracy when using different kernel functions. Olatomiwa et al. (2015) and Shamshirband et al. (2016) both optimized the SVM algorithm and achieved good prediction results. Tree algorithms, such as the random forest algorithm and the gradient boosting regression tree (GBRT) algorithm, have been used to construct solar radiation prediction models with encouraging results (Sun et al., 2016; Persson et al., 2017; Fan et al., 2018; Zeng et al., 2020). In recent years, some scholars have carried out the comparative analysis of a variety of machine learning algorithms (Meenal and Selvakumar, 2018; Pang et al., 2020; Shamshirband et al., 2020), and all these works show that the ANN algorithm does not realize

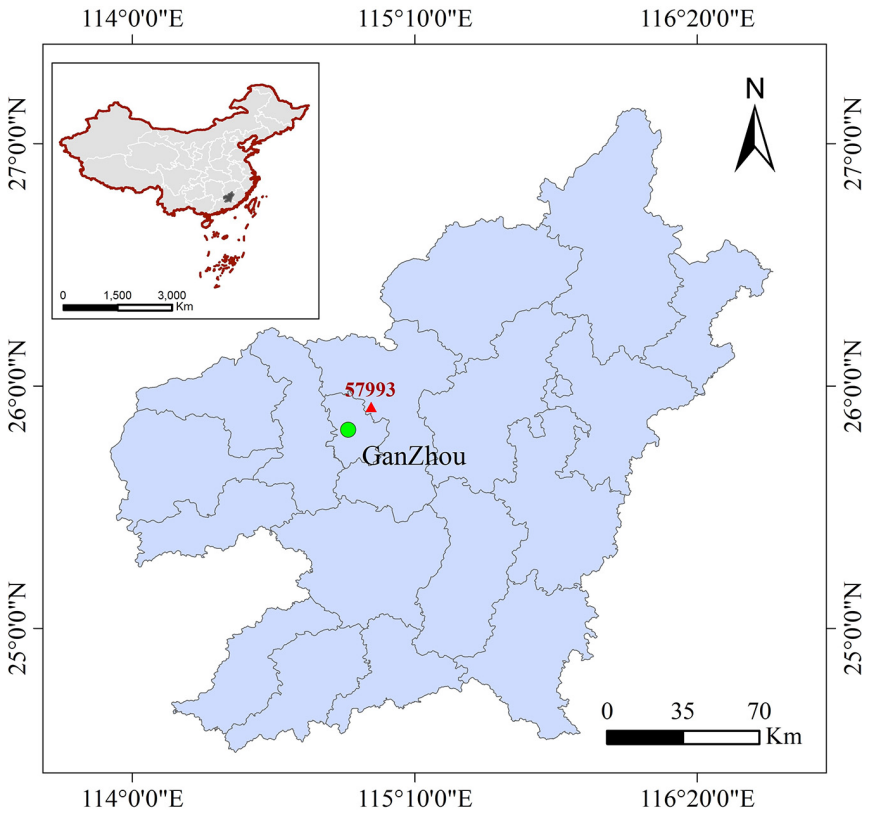

FIGURE 1 | The geographical location of solar radiation monitoring station in Ganzhou County (red triangle).

good prediction results but provides a direction for algorithm improvement. Some studies use deep learning techniques to predict solar radiation. For example, Shamshirband et al. (2019) discuss different types of deep learning algorithms applied in the field of solar, and results show hybrid networks have better performance compared with single networks. Mishra et al. (2020) proposed a short-term solar radiation prediction model using WT-LSTM and achieved good results, showing that deep learning technology has great potential in solar radiation. A CEEMDAN–CNN–LSTM model is proposed by Gao et al. (2020) for hourly multi-region solar irradiance forecasting, and the results present that the model can achieve more accurate prediction performance than other models.

As an investigative technique, machine learning has achieved noteworthy success in many areas, including natural language processing and image recognition (Angra and Ahuja, 2017).

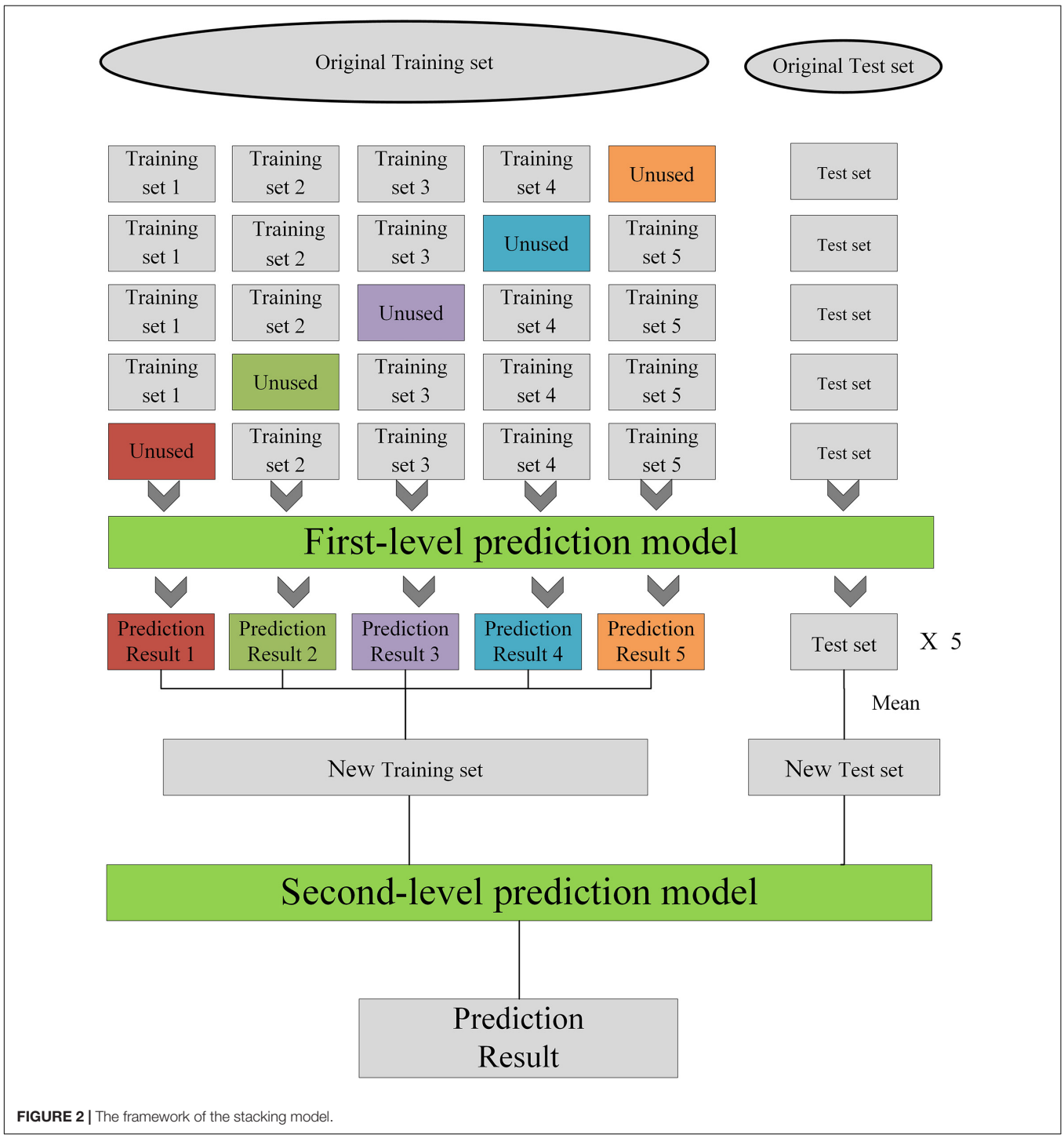

FIGURE 2 | The framework of the stacking model.

The use of machine learning has come to the forefront of the construction of solar radiation models and is a popular direction of research. However, many researchers have focused on the construction of one or several machine learning methods, and there are few in-depth considerations of the differences among these models. Therefore, we used a daily dataset of meteorological elements and basic radiation elements for Ganzhou, China, for the time period 1980–2016 to explore the differences between models of solar radiation prediction. After data processing, we applied the random forest algorithm to selected variables

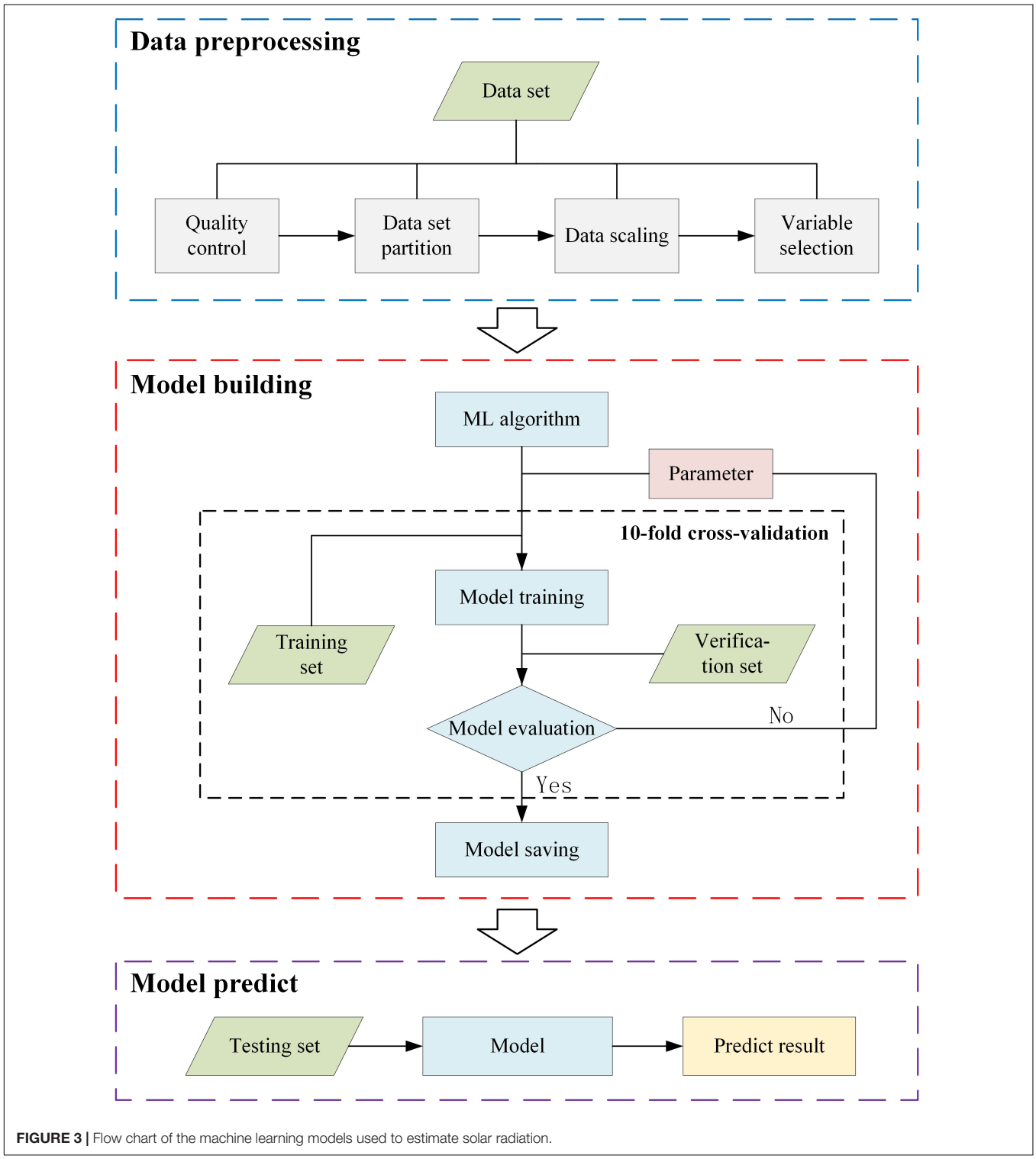

FIGURE 3 | Flow chart of the machine learning models used to estimate solar radiation.

and extracted a monthly dataset based on the daily dataset. We selected 12 machine learning methods to construct a solar radiation prediction model. By comparing the prediction results of these 12 machine learning models, we found the solar radiation prediction models with the best prediction ability. The models with the best prediction ability were then stacked in a linear model. A stacking model was obtained and the predicted results were analyzed.

DATA AND MACHINE LEARNING ALGORITHMS

Study Area and Datasets

Ganzhou city (24.48–30.06° N, 113.57–118.46° E) lies in the south of Jiangxi province in the southern subtropical zone of China and is characterized by a subtropical monsoon climate. It is bordered to the south by Guangdong province, to the east by Fujian province, and to the west by Hunan province. Ganzhou has a mild climate with four distinct seasons and both winter and summer monsoons, with precipitation concentrated in the spring and summer seasons. The annual average temperature is 19.1–20.8°C and the annual rainfall is 1152.2–1554.9 mm. There is a solar radiation monitoring station (No. 57993) in Ganxian County (25.51° N, 114.57° E, 137.5 m above sea-level) (**Figure 1**).

Experimental data were gathered from the China Meteorological Information Center website, including a dataset (V3.0) of daily climate data (temperature, precipitation, air pressure, humidity, temperature, visibility, wind speed, and sunshine duration) from surface stations in China and a daily radiation dataset from Ganzhou's surface solar radiation monitoring station. After referring to relevant research (Will et al., 2013; Mohammadi et al., 2016) and analyzing the quality of the collected data, we selected the data from 1980 to 2016 to estimate solar radiation. The data were selected including the visibility (VIS), the mean relative humidity (RHU-mean), the minimum relative humidity (RHU-min), the mean wind speed (WIN-mean), the mean precipitation (PRE-mean), the mean pressure (PRS-mean), the maximum pressure (PRS-max), the minimum pressure (PRS-min), the sunshine duration (SSD), the mean temperature (TEM-mean), the maximum temperature (TEM-max), the minimum temperature (TEM-min), the mean ground temperature (GST-mean), and the total solar radiation (RAD).

Quality control of the data was essential considering the length of the study period and the inherent errors in the instrument-based observations. We excluded missing and abnormal values in the meteorological data from the final dataset and then applied the requirements for solar radiation data quality control proposed by Younes et al. (2005). In total, 13,100 daily data records and 432 monthly average data records were obtained. The dataset was further divided into training and test sets and then normalized, with the training set

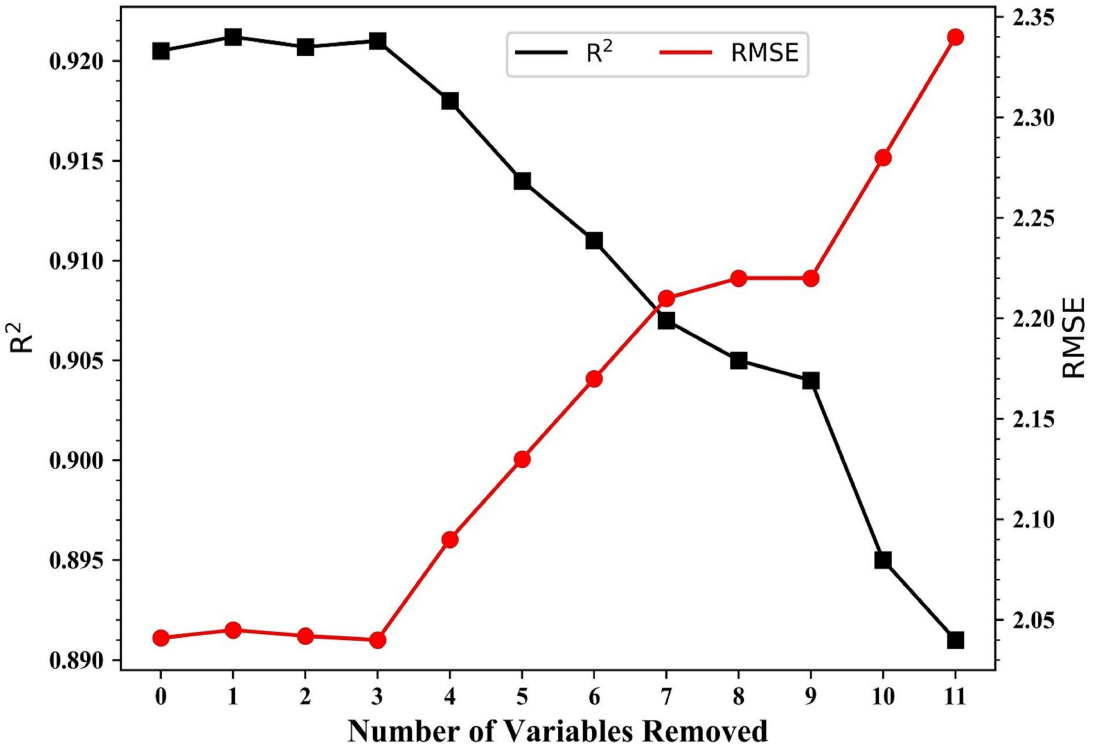

FIGURE 4 | Predictive performance (R^2 and RMSE) of the random forest model during variable selection. Variables were removed in the order PRS-min, PRS-max, RHU-min, PRE-mean, TEM-mean, WIN-mean, TEM-max, TEM-min, RHU-mean, PRS-mean, and VIS.

accounting for 90% and the test set accounting for 10% of all data. Our final sample consisted of 11,790 daily training sets, 1,310 daily test sets, 388 monthly training sets, and 44 monthly test sets.

Machine Learning Predictive Algorithms and Stacking Techniques

Machine Learning Algorithms

With the development of machine learning technology, an increasing number of researchers are using machine learning to predict solar radiation. We investigated 12 different machine learning predictive algorithms: multiple linear regression (Baczek et al., 2005; Nathans et al., 2012), the radial basis function neural network (Mahanty and Dutta Gupta, 2004; Li M. et al., 2008), the K-nearest neighbor model (Shen and Chou, 2005; Deng et al., 2016), the decision tree (Brodley and Friedl, 1997; Quinlan, 1999), the back-propagation neural network (Van Ooyen and Nienhuis, 1992; Trappey et al., 2006), the extreme learning machine (Deng et al., 2015; Huang G. et al., 2015), SVM regression (Burges, 1998; Shamshirband et al., 2016), Gaussian process regression (GPR) (Nguyen-Tuong et al., 2009; Ebden, 2015), the GBRT (Zhang and Haghani, 2015; Johnson et al., 2018), adaptive boosting (Adaboost) (Zhu et al., 2006; Li X. et al., 2008; Wang, 2012), extreme gradient lifting (XGBoost) (Nielsen, 2016; Torlay et al., 2017), and random forest (Kapwata and Gebreslasie, 2016; Sun et al., 2016) algorithms. A detailed description of machine learning methods can be found in **Supplementary Text S1**.

Stacking Model

Stacking technology is a general integration algorithm that integrates advanced learners by using multiple lower-level learners to achieve higher performance (Agarwal and Chowdary, 2020). In general, the K-fold cross-validation method is used to train and test these models and then output the prediction results. The prediction results output by each model is then combined into a stacking model, which is built to reduce the generalization errors. The stacking model usually consists of two layers. The first layer is the base learner, and the input is the initial training set. The second layer is trained with the output data from the first layer as the input data and gives the final results.

The steps of the stacking model construction are as **Figure 2**. Each model is trained using five-fold cross-validation. The training set is divided into five parts, and four parts are selected as the training data and one set as the test data. The test data in each of the four training sets is predicted to obtain a prediction result (**a**) and the test set data are predicted by the trained model to obtain the test set prediction result (**b**). After five training runs, the prediction result **a** of each of the five runs is combined into one column as **A** and the prediction result **b** is averaged as **B**. The new datasets **A** and **B** are obtained, in which the number in **A** is the same as the number of training sets, but **A** is one-dimensional data. After constructing **N** single models, **N A** and **N B** are generated, then the **N A** and **N B** data are combined into a new training set and a new test set. A simple linear model is used as the second layer to train using the new training set and test with the new test set.

TABLE 1 | Descriptive statistics of the modeling variables in the training dataset.

	Spring				Summer				Fall				Winter				Year			
	Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std	Max	Min
VIS	14.69	5.64	50.00	1.78	22.97	10.58	65.75	4.90	14.21	5.09	51.75	4.05	12.19	3.23	35.75	2.70	16.02	6.21	65.75	1.78
RHU-mean	78.13	11.00	100.00	34.25	72.41	10.00	97.75	43.00	72.70	10.74	99.00	31.25	74.58	12.42	98.75	27.25	74.46	11.04	100.00	27.25
WIN-mean	1.41	0.80	5.25	0.00	1.62	0.75	5.00	0.00	1.39	0.76	46.00	0.00	1.39	0.81	4.75	0.00	1.45	0.78	5.25	0.00
PRE-mean	62.1	121.8	1067.0	0.00	49.3	130.0	1497.0	0.00	22.8	77.8	1184.0	0.00	23.9	66.0	1010.0	0.00	39.5	98.9	1497.0	0.00
PRS-mean	998.57	5.61	1018.80	994.80	990.42	3.17	1001.00	975.50	1001.10	5.63	1019.10	982.60	1007.96	5.04	1024.40	987.20	999.51	4.86	1024.00	975.50
TEM-mean	19.34	5.87	31.90	2.50	28.64	2.36	34.00	18.10	21.10	5.26	31.90	1.60	9.54	4.33	25.20	−0.90	19.66	4.46	34.00	−0.90
TEM-max	23.75	6.84	36.50	4.80	33.49	3.10	40.00	20.30	25.93	59.10	38.50	2.70	13.95	5.97	31.40	−0.20	24.28	5.46	40.00	−0.20
TEM-min	16.22	5.49	27.40	−0.30	25.16	1.85	30.00	15.60	17.64	5.31	27.80	0.40	6.53	4.13	22.50	−3.90	16.39	4.20	30.00	−3.90
GST-mean	21.31	6.74	39.20	3.60	33.07	4.57	43.20	20.60	24.02	6.50	40.60	3.00	10.77	4.44	27.30	0.10	22.29	5.56	43.20	0.00
SSD	3.47	3.91	126.00	0.00	71.50	4.03	13.00	0.00	5.26	4.02	11.60	0.00	3.26	3.70	10.40	0.00	4.79	3.92	13.00	0.00
RAD	10.90	7.41	28.46	2.00	17.70	6.42	30.48	5.00	12.05	6.14	26.80	3.00	7.36	4.96	20.15	1.00	12.02	6.28	30.48	1.00

Spring: March, April, May; Summer: June, July, August; Fall: September, October, November; Winter: December, January, February. Std, standard deviation; Max, maximum; Min, minimum.

MATERIALS AND METHODS

Prediction of the Flow of Solar Radiation

Our experiment consisted of three parts (**Figure 3**): data preprocessing, model building, and model prediction. The data preprocessing involved four steps: data quality control, dataset partitioning, data scaling, and variable selection. Among them, data quality control, dataset partitioning, and data scaling are described in Section "Study Area and Datasets," and variable selection is described in Section "Variable Selection." The main processes of the model building were as follows: the selection of the machine learning algorithm, parameter selection, model construction, and model saving. We used the 10-fold cross-validation method (Jiang and Wang, 2017) in the parameter selection step. We can get a detailed description of the model building in Section "Model Building." In the model prediction step, the saved model from the model building step was used to predict the solar radiation using the test dataset. Then, we save the predicted results and analysis. The specific experimental steps proceeded as follows:

(1) data collection and data preprocessing;
(2) choose a machine learning algorithm from the 12 algorithms to predict solar radiation;
(3) compare solar radiation predictive ability based on different parameters;
(4) if the best predictive ability is achieved, save the model;
(5) return to step (2) and choose another machine learning algorithm until all 12 algorithms have been subjected to machine learning model building;
(6) input the preprocessing dataset (we prepared datasets on two timescales—daily and monthly—to estimate the solar radiation predictive performance of the 12 machine learning models) and use the 12 saved machine learning models to predict solar radiation and obtain the predicted results;
(7) save predicted results and analyze.

Variable Selection

The variable selection step is important in constructing machine learning models. The current mainstream variable selection algorithms include the genetic algorithm (Huang and Chiu, 2006), the Tabu search (Corazza et al., 2013), particle swarm optimization (Khatibi Bardsiri et al., 2013), and the random forest algorithm (Kapwata and Gebreslasie, 2016). We used the random forest algorithm to select data variables (Zeng et al., 2020). Normalized daily data were used to construct and train the random forest model and to calculate the model's importance. The data preprocessing experiment was intended to verify the importance of variables in a given model and to analyze the impact of changes in the variables on the model's predictive performance. The experiment proceeded as follows:

(1) divide the dataset into a training set and test set after completing the data quality control process;
(2) use the training set to train and save the model, then calculate the correlation coefficient (R^2) and the root mean square error (RMSE) of the saved model;
(3) based on the order of importance of the variables in the model, eliminate the least important variable;
(4) repeat steps (2) and (3) until only two variables remain (the minimum required for calculation).

Figure 4 shows that when the model contained <10 variables, R^2 tended to decrease and the RMSE tended to increase. Between 12 and 10 variables, R^2 reached 0.921 and the RMSE was 2.042 MJ/m^2. With four variables, R^2 decreased sharply from 0.904 to 0.895 and the RMSE decreased from 2.19 to 2.28 MJ/m^2. Therefore, the prediction of solar radiation can achieve the best performance when using 10 variables, then the subsequent model experiments were trained with these 10 variables.

Model Building

Experiments were performed in Python 3.6 using third-party libraries such as Pandas, NumPy, the scikit-learn machine learning library (Sklearn), and the Xgb library. Twelve machine learning algorithms were chosen to build the models. The initial parameter settings of each algorithm were determined according to the algorithm's characteristics. For example, for a neural network model, the number of hidden layers and the number of neurons were determined based on empirical formulas and neural network design principles (Basheer and Hajmeer, 2000). The respective selection ranges of the adjustment parameters and other parameters were then set according to the parameter adjustment methods for different machine learning algorithms. We used Sklearn's GridSearchCV method to select parameters for each of the 12 machine learning models, ultimately saving the best model. The first layer of the stacking model consists of those multiple models with excellent predictive power. The parameters of the first layer model are the parameters selected previously and the second layer is constructed by multiple linear regression. After obtaining the best parameters, the train set was used to train the model and the final model was saved. The time spent training the model is the model construction time, and the final model size is the model memory. When the model was constructed, input the test set was input to get the prediction result.

Statistical Metrics

The models were evaluated using four indicators: R^2, RMSE, MAE, and BIAS:

$$R^2 = \frac{\left(\sum_{t=1}^{n}(yo_t-\bar{yo})(ym_t-\bar{ym})\right)^2}{\sum_{t=1}^{n}(yo_t-\bar{yo})^2 \cdot \sum_{t=1}^{n}(ym_t-\bar{ym})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(yo_t-ym_t)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\left|\left(yo_t-ym_t\right)\right|\right| \quad (3)$$

$$BIAS = \left(yo_t-ym_t\right) \quad (4)$$

where n indicates the amount of data, ym_t is the predicted solar radiation, yo_t is the observed solar radiation, and $\bar{ym}$ and $\bar{yo}$ represent the average of the predicted and observed results, respectively.

If R^2 is close to 1, then the observed and predicted values are closely correlated. The closer the RMSE/MAE values are to 0, the better the predicted value fits the observed value. A combination of metrics, including, but not limited to, the RMSE and MAE, are often required to assess the performance of the model.

RESULTS

Description and Selection of Variables

The average annual range of the RAD was 1–30.48 MJ/m^2, with a mean value of 12.02 MJ/m^2 and a standard deviation of 6.28 MJ/m^2 (**Table 1**). The annual mean (standard deviation) values were VIS 16.02 (6.21) km, RHU-mean 74.46 (11.04)%, WIN-mean 1.45 (0.78) m/s, PRE-mean 39.5 (98.9) mm, PRS-mean 999.51 (4.86) hPa, TEM-mean 19.66 (4.46)°C, TEM-max 24.28 (5.46)°C, TEM-min 16.39 (4.2)°C, GST-mean 22.29 (5.56)°C, and SSD 4.79 (3.92) h. Apart from the RHU-mean, PRE-mean, and PRS-mean, the mean values of the variables were highest in summer, followed by spring and autumn, and were lowest in winter. **Supplementary Figure 1** shows the annual maximum GST-mean and the corresponding solar radiation from 1980 to 2016. The trend of GTS-max and the corresponding solar radiation values were generally consistent and increased with the solar radiation, confirming the importance of solar radiation in compound climate extreme events (Ohunakin et al., 2015).

Figure 5 shows the importance of the input variables as predictors in the final random forest model. SSD was identified as the most critical variable, followed in descending order by GST-mean, VIS, PRS-mean, RHU-mean, TEM-min, TEM-max, WIN-mean, TEM-mean, PRE-mean, RHU-min, PRS-max, and PRS-min. The importance of SSD was 85%, which agrees with the results of earlier studies (Chen et al., 2013; Suehrcke et al., 2013; Zeng et al., 2020). The importance of GST-mean was 6% and the importance of all other variables was <5%.

Predictive Performance for Daily Solar Radiation

Figure 6 shows the performance of the 12 machine learning models in predicting solar radiation for the given daily dataset. The statistical results show that most of the machine learning models used to predict solar radiation yielded satisfactory results. The R^2 values of the 12 machine learning models ranged from 0.838 to 0.925. The GBRT, GPR, XGBoost, and random forest models were the best machine learning models to predict solar radiation with R^2 values of 0.925, 0.923, 0.922, and 0.921,

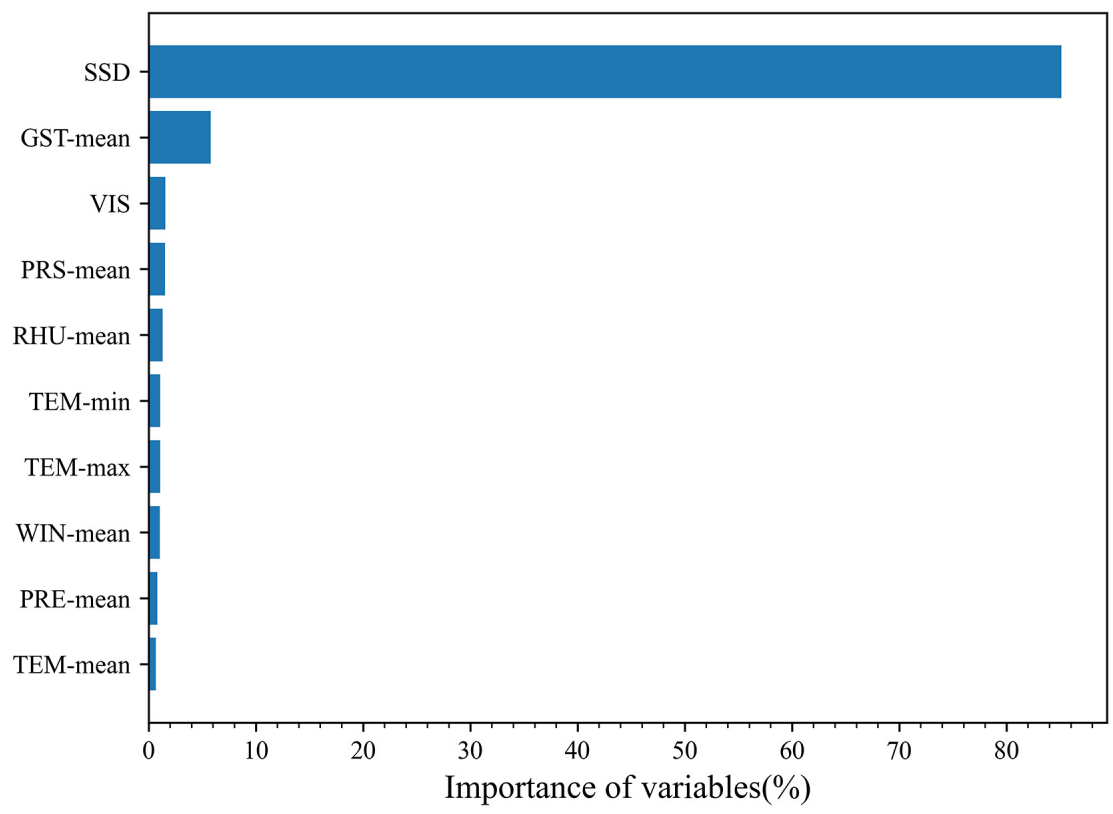

FIGURE 5 | Importance of variables in predicting solar radiation using the random forest model.

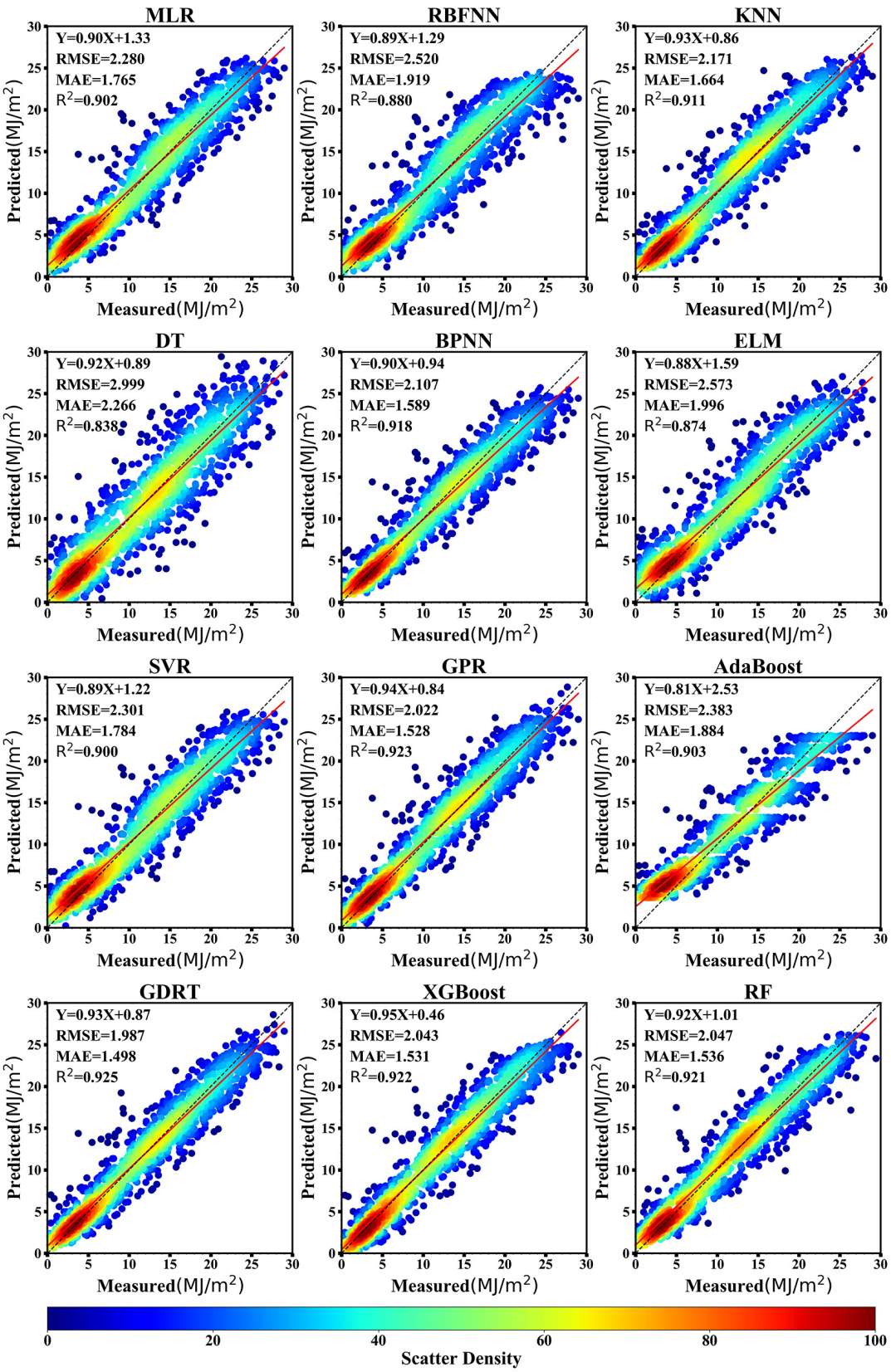

FIGURE 6 | Scatter plots of the cross-validation results for 12 machine learning models in predicting daily solar radiation at Ganzhou from 1980 to 2016.

respectively. The R^2 values of the extreme learning machine and decision tree models were 0.874 and 0.838, respectively, which indicated that these models had the poorest precision for the prediction of solar radiation. The RMSE values of the 12 machine learning models were in the range 1.987–2.999 MJ/m^2. The RMSE value of the GBRT model was the lowest (1.987 MJ/m^2),

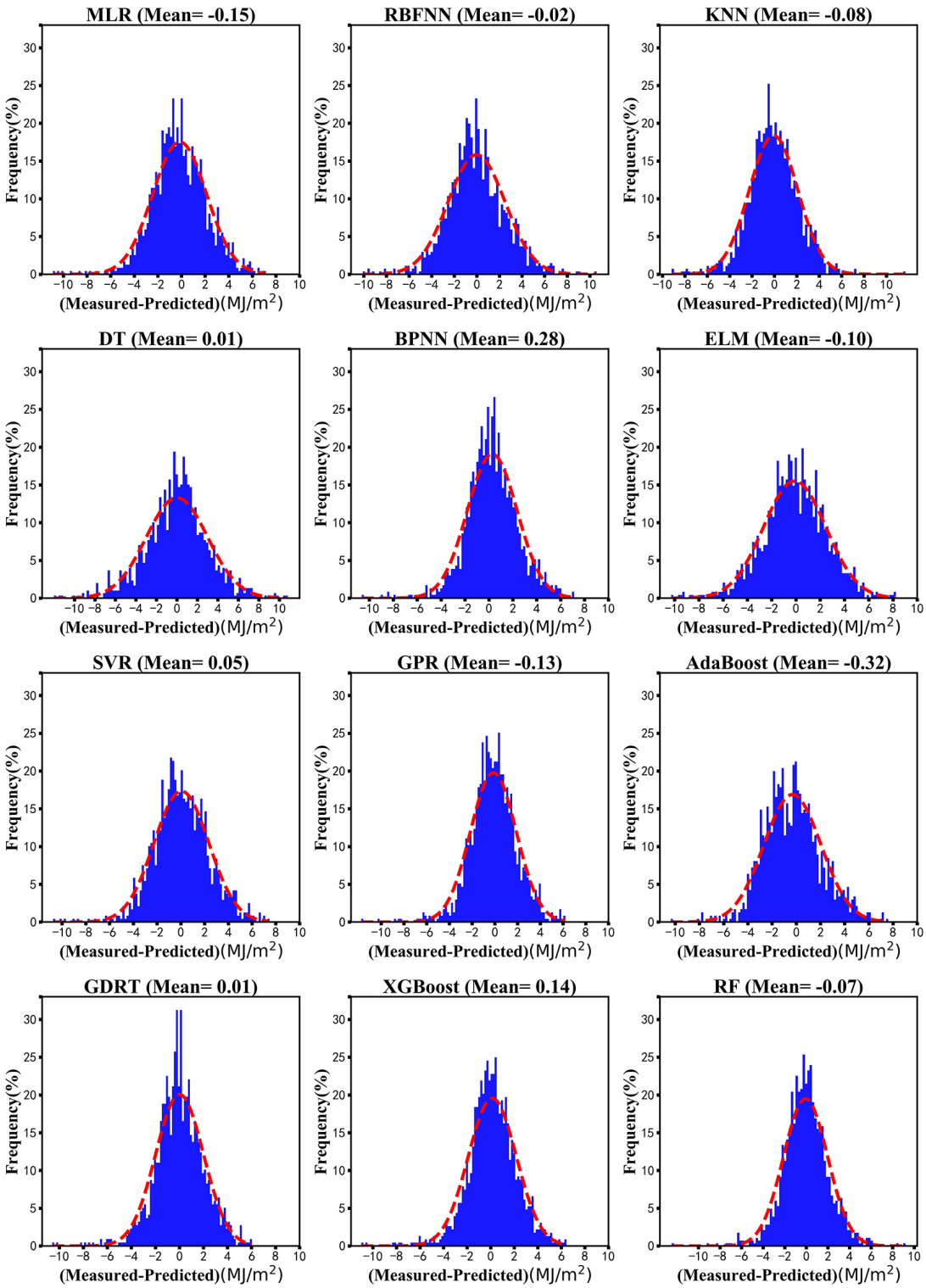

FIGURE 7 | Deviation distribution of machine learning models in predicting daily solar radiation at Ganzhou from 1980 to 2016.

indicating that this model was the best for predicting solar radiation. By contrast, the RMSE value of the decision tree model was the largest (2.999 MJ/m^2), suggesting that this model was the poorest predictor of solar radiation. The MAE values of the 12 machine learning models ranged from 1.498 to 2.266 MJ/m^2, with the GBRT model returning the smallest value (MAE = 1.498 MJ/m^2), meaning that the deviation between the predicted and measured values was also the smallest. The MAE value of the decision tree model was the largest (MAE = 2.266 MJ/m^2), demonstrating that this model had the largest prediction bias. The MAE values of the other machine learning models were both <2.0 MJ/m^2.

Figure 7 shows distribution maps of the daily deviation probability to further explore the distribution of the deviation of solar radiation prediction for the 12 machine learning models. The results showed that the bias of the GBRT and the decision tree models both were 0.01 MJ/m^2, followed by the RBNN model (−0.02 MJ/m^2). The bias of the AdaBoost model for solar radiation prediction was −0.32 MJ/m^2. The deviation values of most models were mainly distributed between −6 and +6 MJ/m^2, whereas those of the decision tree and extreme learning machine models were mainly distributed between −8 and +8 MJ/m^2. **Table 2** shows the number of deviation values that fell within the range ±2 MJ/m^2 in the prediction of solar radiation for the 12 models. The deviation in solar radiation prediction for the GBRT, GPR, XGBoost, and random forest models each exceeded 940, compared with only 734 for the decision tree model.

The prediction results from the daily value data indicate that the GBRT, XGBoost, GPR, and random forest models had a relatively good predictive ability, whereas the extreme learning machine and decision tree models performed poorly. The random forest model had the longest construction time, followed by the GBRT and the GPR models; the XGBoost model had the shortest construction time. This is related to the model principle—for example, to obtain better training results, the random forest model needs more CART-based models, which increases the training time. By contrast, XGBoost uses parallel processing to increase the operational speed and therefore requires less time.

TABLE 2 | Statistics for the amount of daily data for each model deviation within ±2 8 MJ/m^2.

Model	Number of data points	Percentage
Multiple linear regression	861	65.7
Radial basis function neural network	804	61.4
K-nearest neighbor	894	68.2
Decision tree	734	56.0
Back-propagation neural network	935	71.4
Extreme learning machine	768	58.6
Support vector machine regression	846	64.5
Gaussian process regression	941	71.8
AdaBoost	794	60.6
Gradient boosting regression tree	956	73
XGBoost	950	72.5
Random forest	945	72.1

Predictive Performance for Monthly Solar Radiation

Figure 8 presents a scatter plot of the monthly predicted and measured values for different models. The R^2 values for the 12 machine learning models ranged from 0.900 to 0.944 and were > 0.9 for all models. The XGBoost model gave the best prediction result, with an R^2 value of 0.944; the GPR (R^2 = 0.941), GBRT (R^2 = 0.938), and random forest (R^2 = 0.936) models also demonstrated a good prediction performance. The K-nearest neighbor (R^2 = 0.900) and decision tree (R^2 = 0.901) models gave relatively poor prediction results. The RMSE of each model fell between 1.131 and 1.580 MJ/m^2. The XGBoost model returned the lowest RMSE of 1.131 MJ/m^2, reflecting the highest precision of all the models. The decision tree model had the lowest precision (RMSE = 1.580 MJ/m^2). The MAE values for all models ranged from 0.870 to 1.174 MJ/m^2. The MAE of the XGBoost model was the smallest (MAE = 0.870 MJ/m^2), indicating that the predicted value was close to the observed value.

For the monthly average data, **Figure 9** shows the largest deviation in the RBNN model (bias 0.88 MJ/m^2), followed by random forest (bias −0.02 MJ/m^2) and SVM regression (bias 0.08 MJ/m^2) models and the lowest deviation in the GBRT model (bias −0.01 MJ/m^2). In contrast with the deviation in the daily data, the monthly average prediction bias of most models was positive, although the decision tree, GBRT, and random forest models showed a negative deviation. According to the monthly mean deviation probability distribution, the main distribution interval of the model deviation was within ±4. **Table 3** gives the statistical results for the monthly data with a predicted deviation between −2 and +2 MJ/m^2, with 37 data points in the random forest model and 40 data points in the GBR model.

The XGBoost, GPR, GBRT, and random forest models showed better predictive ability on the monthly average data, whereas the K-nearest neighbor and decision tree models performed poorly. When the amount of data is small, the XGBoost, GPR, GBRT, and random forest models are all built very quickly, but the XGBoost model is the fastest with the highest prediction accuracy. Besides, XGBoost has strong anti-overfitting and generalization abilities. This is advantageous for the construction of the monthly radiation value in models with a small number of data points, which is an advantage over the other machine learning models. The XGBoost model is therefore recommended when there is only a small number of data points.

Predictive Performance of the Stacking Model

The XGBoost, GPR, GBRT, and random forest single models showed excellent prediction capabilities. These four models were therefore used as the first layer model and multiple linear regression was used as the second layer model to build a stacking model. **Figures 10A,B** show the predicted results and bias probability distributions. **Figure 10A** shows that the R^2 of the stacking model is 0.929, the RMSE is 1.940 MJ/m^2, and the MAE is 1.457 MJ/m^2. Compared with the 12 single models, the stacking model has the highest R^2 value, but the lowest RMSE

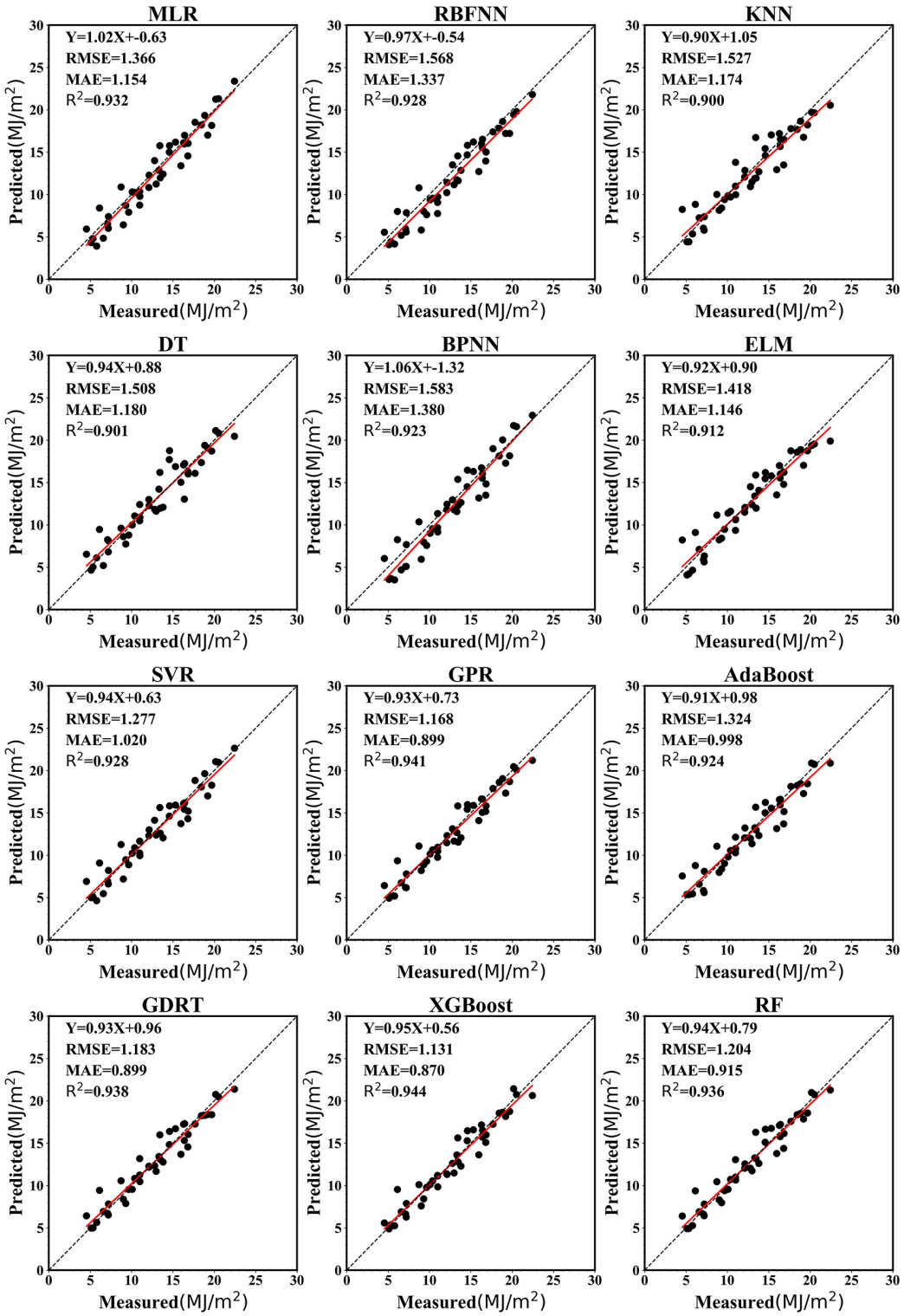

FIGURE 8 | Scatter plots of the results of machine learning models in predicting monthly average solar radiation at Ganzhou from 1980 to 2016.

and MAE. **Figure 10B** shows that the average deviation of the stacking model is 0 MJ/m^2 and the deviation of the distribution is more uniform than that of the single models. The stacking model predicts 74.8% of the data with a bias distribution in [−2, 2]. The stacking model has a better prediction ability for the daily data than the single models. **Figure 10C** shows that

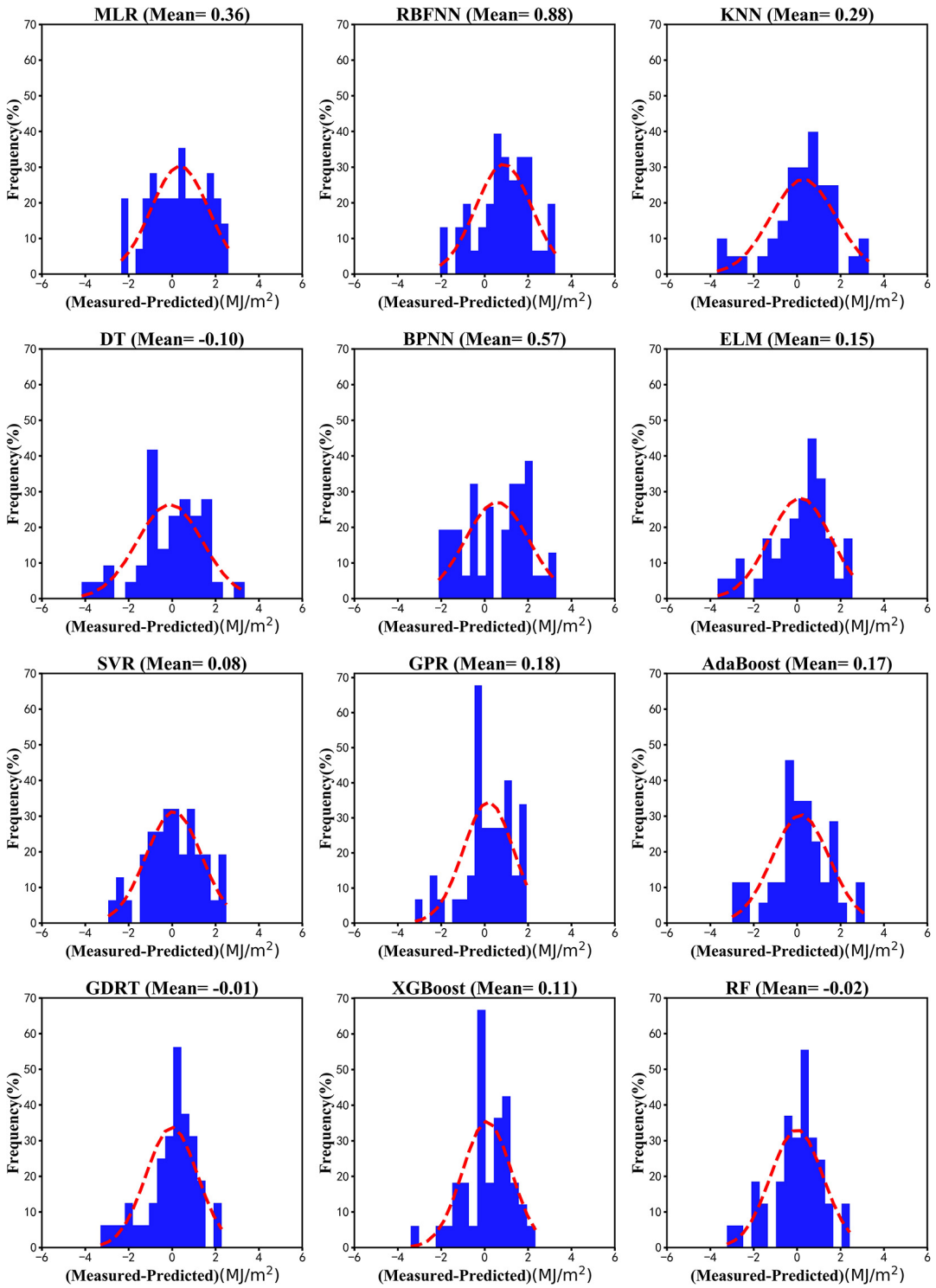

FIGURE 9 | Deviation distribution of machine learning models in predicting the monthly average solar radiation at Ganzhou from 1980 to 2016.

the R^2 value of the stacking fusion model is 0.943, the RMSE is 1.142 MJ/m^2, and the MAE is 0.884 MJ/m^2, all lower than the XGBoost model (R^2 0.944, RMSE 1.131 MJ/m^2, and MAE 0.870 MJ/m^2). **Figure 10D** shows that the average value of the stacking deviation of the stacking model is 0.13 MJ/m^2 and there are only 39 deviations between [2, −2]. The stacking model has no advantage over the XGBoost model in terms of construction time.

TABLE 3 | Statistics for the amount of monthly data in each model deviation within ±2.

Model	Number of data points	Percentage
Multiple linear regression	35	79.5
Radial basis function neural network	36	82
K-nearest neighbor	36	82
Decision tree	38	86.4
Back-propagation neural network	34	77.3
Extreme learning machine	35	79.5
Support vector machine regression	36	82
Gaussian process regression	40	90.9
AdaBoost	37	84.1
Gradient boosting regression tree	38	77.3
XGBoost	40	90.9
Random forest	37	84.1

DISCUSSION

Many studies have compared the ability of machine learning algorithms to predict solar radiation (**Supplementary Table 1**). Moreno et al. (2011) used an ANN and generalized regression to build models separately, positing that an ANN has the same predictive power as generalized regression. Yang et al. (2014) applied ANN-SVM, SVM, and ANN to construct separate models, giving a model performance in the order ANN-SVM > SVM > ANN. Wang et al. (2016) compared the MLP, RBNN, and GRNN models and noted RBNN > GRNN > MLP in terms of performance. We used daily and monthly data to predict the performance of 12 machine learning models and showed that the GBRT, GPR, XGBoost, and random forest models had better prediction capabilities than the other models. We also combined the XGBoost, GBRT, GPR, and random forest models using stacking technology. The performance of the stacking model in predicting the daily solar radiation set was better than that of the 12 single models, but the performance using the monthly dataset gave no advantage over the XGBoost model.

We found that the input of a small measured value of solar radiation returned a large predicted output value, whereas the input of a large value of solar radiation returned a small predicted output value after machine learning processing. This phenomenon may be linked to data that were relatively concentrated and contained fewer, but higher, measured values. The data scaling method greatly influences the performance of machine learning models (Huang J. et al., 2015; García et al., 2016). Normal processing methods include no processing, normalization, standardization, and regularization. We adopted four different data processing methods to build 12 different machine learning models with daily or monthly data. The results are shown in **Supplementary Tables 2, 3**.

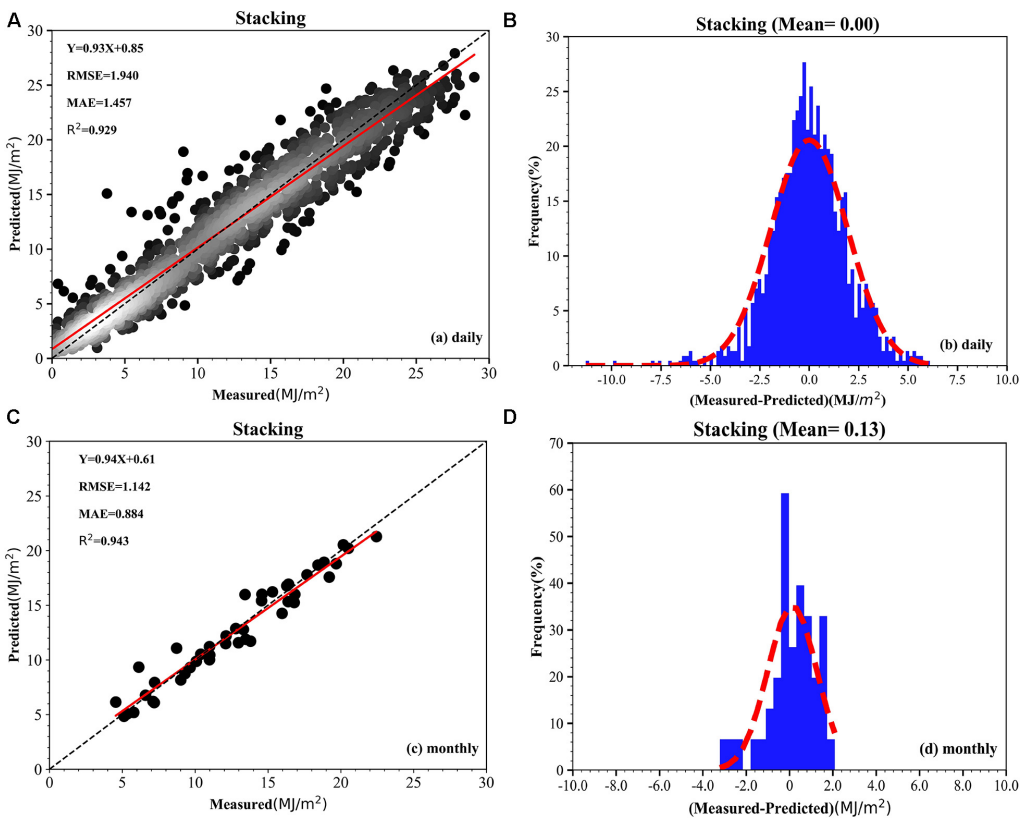

FIGURE 10 | (A–D) Scatter plot and distribution of deviation of the prediction results of the stacking model.

CONCLUSION

We performed data preprocessing and variable selection based on meteorological elements and solar radiation data from 1980 to 2016 for Ganzhou station, China. Then, 12 machine learning models were developed using Sklearn and the Xgb library. By comparing and evaluating the predictive ability of the 12 machine learning models using R^2, the RMSE, the MAE and BIAS indices, the XGBoost, GPR, GBRT, and random forest models were selected as the first layer, and multiple linear regression was selected as the second layer to construct a stacking model to predict solar radiation.

Using the random forest algorithm to select the variables, the SSD was identified as the most important variable. The time series of the annual maximum GST-mean and the corresponding solar radiation value from 1980 to 2016 showed that the maximum GTS-max increases with the solar radiation, which confirms the importance of solar radiation in compound extreme climate events. The GBRT, XGBoost, random forest, and GPR models performed better than the other models for the daily and monthly datasets. The GBRT model had the best predictive ability for the daily datasets, whereas the XGBoost model had the best predictive ability for the monthly datasets. The random forest model had the longest construction time, followed by the GBRT and GPR models, whereas the XGBoost model had the shortest construction time. This phenomenon is related to the principles of the models.

The prediction ability of the stacking model was improved in the daily solar radiation prediction model, but the monthly model performed poorly, which may be related to too little monthly training data. We concluded that the XGBoost model is the best solar radiation value prediction model, although when the amount of data is large, we suggest using the stacking fusion or XGBoost model to build the model.

DATA AVAILABILITY STATEMENT

All meteorological data were obtained from the China Meteorological Data Service Center (CMDC, http://data.cma.cn/en/?r=data/index&cid=6d1b5efbdcbf9a58), which requires an authorized log-in or via off-line data processing and product tailoring services. Specifically, daily observations are found at http://data.cma.cn/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0.

AUTHOR CONTRIBUTIONS

JK and ZZ conceived the idea of the study. LH analyzed the data and wrote the initial draft of the manuscript. The remaining authors contributed to refining the ideas, carrying out additional analyses, and finalizing this manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Program of China (Grant No. 2016YFC0803105), National Natural Science Foundation of China (Grant No. 41301423), China Postdoctoral Science Foundation (Grant No. 2018M641926), Projects supported by the National Fund for Study Abroad (Grant No. 201808360065), and Jiangxi Provincial Department of Education Science and Technology Research Projects (Grant No. GJJ150661).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2021.596860/full#supplementary-material

REFERENCES

Agarwal, S., and Chowdary, C. R. (2020). A-stacking and A-bagging: adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Syst. Appl.* 146:113160. doi: 10.1016/j.eswa.2019.113160

Angra, S., and Ahuja, S. (2017). "Machine learning and its applications: a review," in *Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017*, (Piscataway, NJ: IEEE), 57–60. doi: 10.1109/ICBDACI.2017.8070809

Angstrom, A. (1924). Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. *Q. J. R. Meteorol. Soc.* 50, 121–126. doi: 10.1002/qj.49705021008

Azadeh, A., Maghsoudi, A., and Sohrabkhani, S. (2009). An integrated artificial neural networks approach for predicting global radiation. *Energy Convers. Manag.* 50, 1497–1505. doi: 10.1016/j.enconman.2009.02.019

Baczek, T., Wiczling, P., Marszałł, M., Heyden, Y. V., and Kaliszan, R. (2005). Prediction of peptide retention at different HPLC conditions from multiple linear regression models. *J. Proteome Res.* 4, 555–563. doi: 10.1021/pr049780r

Basheer, I. A., and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *J. Microbiol. Methods* 43, 3–31. doi: 10.1016/S0167-7012(00)00201-3

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., et al. (2010). Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science* 329, 834–838. doi: 10.1126/science.1184984

Besharat, F., Dehghan, A. A., and Faghih, A. R. (2013). Empirical models for estimating global solar radiation: a review and case study. *Renew. Sustain. Energy Rev.* 21, 798–821. doi: 10.1016/j.rser.2012.12.043

Bhargawa, A., and Singh, A. K. (2019). Solar irradiance, climatic indicators and climate change – An empirical analysis. *Adv. Space Res.* 64, 271–277. doi: 10.1016/j.asr.2019.03.018

Bristow, K. L., and Campbell, G. S. (1984). On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agric. For. Meteorol.* 31, 159–166. doi: 10.1016/0168-1923(84)90017-0

Brodley, C. E., and Friedl, M. A. (1997). Decision tree classification of land cover from remotely sensed data. *Rem. Sens. Environ.* 61, 399–409. doi: 10.1016/S0034-4257(97)00049-7

Budyko, M. I. (1969). The effect of solar radiation variations on the climate of the Earth. *Tellus* 21, 611–619. doi: 10.3402/tellusa.v21i5.10109

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Discov. Data Min. Knowl. Discov.* 2, 121–167.

Chen, J. L., Li, G. S., and Wu, S. J. (2013). Assessing the potential of support vector machine for estimating daily solar radiation using sunshine

duration. *Energy Convers. Manag.* 75, 311–318. doi: 10.1016/j.enconman.2013.06.034

Chen, J. L., Liu, H. B., Wu, W., and Xie, D. T. (2011). Estimation of monthly solar radiation from measured temperatures using support vector machines – A case study. *Renew. Energy* 36, 413–420. doi: 10.1016/j.renene.2010.06.024

Cline, D. W., Bales, R. C., and Dozier, J. (1998). Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling. *Water Resour. Res.* 34, 1275–1285. doi: 10.1029/97WR03755

Corazza, A., Di Martino, S., Ferrucci, F., Gravino, C., Sarro, F., and Mendes, E. (2013). Using tabu search to configure support vector regression for effort estimation. *Empir. Softw. Eng.* 18, 506–546. doi: 10.1007/s10664-011-9187-3

Deng, C. W., Huang, G. B., Xu, J., and Tang, J. X. (2015). Extreme learning machines: new trends and applications. *Sci. China Inf. Sci.* 58, 1–16. doi: 10.1007/s11432-014-5269-3

Deng, Z., Zhu, X., Cheng, D., Zong, M., and Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing* 195, 143–148. doi: 10.1016/j.neucom.2015.08.112

Ebden, M. (2015). Gaussian processes: a quick introduction. *arXiv* [Preprint] arXiv:1505.02965,

Fadare, D. A. (2009). Modelling of solar energy potential in Nigeria using an artificial neural network model. *Appl. Energy* 86, 1410–1422. doi: 10.1016/j.apenergy.2008.12.005

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., et al. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Convers. Manag.* 164, 102–111. doi: 10.1016/j.enconman.2018.02.087

Gao, B., Huang, X., Shi, J., Tai, Y., and Zhang, J. (2020). Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks. *Renew. Energy* 162, 1665–1683. doi: 10.1016/j.renene.2020.09.141

García, S., Luengo, J., and Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge Based Syst.* 98, 1–29. doi: 10.1016/j.knosys.2015.12.006

Garland, F. C., Garland, C. F., Gorham, E. D., and Young, J. F. (1990). Geographic variation in breast cancer mortality in the United States: a hypothesis involving exposure to solar radiation. *Prev. Med. (Baltim).* 19, 614–622. doi: 10.1016/0091-7435(90)90058-R

Grant, W. B., and Tuohimaa, P. (2004). Geographic variation of prostate cancer mortality rates in the United States: Implications for prostate cancer risk related to vitamin D [3] (multiple letters). *Int. J. Cancer* 111, 470–471. doi: 10.1002/ijc.20220

Gueymard, C. A. (2003). Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment. *Sol. Energy* 74, 355–379. doi: 10.1016/S0038-092X(03)00195-6

Halabi, L. M., Mekhilef, S., and Hossain, M. (2018). Performance evaluation of hybrid adaptive neuro-fuzzy inference system models for predicting monthly global solar radiation. *Appl. Energy* 213, 247–261. doi: 10.1016/j.apenergy.2018.01.035

Hoogenboom, G. (2000). Contribution of agrometeorology to the simulation of crop production and its applications. *Agric. Forest Meteorol.* 103, 137–157. doi: 10.1016/S0168-1923(00)00108-8

Huang, G., Huang, G. B., Song, S., and You, K. (2015). Trends in extreme learning machines: a review. *Neural Netw.* 61, 32–48. doi: 10.1016/j.neunet.2014.10.001

Huang, J., Li, Y. F., and Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Inf. Softw. Technol.* 67, 108–127. doi: 10.1016/j.infsof.2015.07.004

Huang, S. J., and Chiu, N. H. (2006). Optimization of analogy weights by genetic algorithm for software effort estimation. *Inf. Softw. Technol.* 48, 1034–1045. doi: 10.1016/j.infsof.2005.12.020

Islam, M. D., Kubo, I., Ohadi, M., and Alili, A. A. (2009). Measurement of solar energy radiation in Abu Dhabi, UAE. *Appl. Energy* 86, 511–515. doi: 10.1016/j.apenergy.2008.07.012

Iziomon, M. G., and Mayer, H. (2002). Assessment of some global solar radiation parameterizations. *J. Atmos. Solar Terrestrial Phys.* 64, 1631–1643. doi: 10.1016/S1364-6826(02)00131-1

Jiang, G., and Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognit.* 69, 94–106. doi: 10.1016/j.patcog.2017.03.025

Jiang, Y. (2009). Computation of monthly mean daily global solar radiation in China using artificial neural networks and comparison with other empirical models. *Energy* 34, 1276–1283. doi: 10.1016/j.energy.2009.05.009

Johnson, N. E., Bonczak, B., and Kontokosta, C. E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmos. Environ.* 184, 9–16. doi: 10.1016/j.atmosenv.2018.04.019

Kapwata, T., and Gebreslasie, M. T. (2016). Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province,South Africa. *Geospat. Health* 11, 251–262. doi: 10.4081/gh.2016.434

Khatibi Bardsiri, V., Jawawi, D. N. A., Hashim, S. Z. M., and Khatibi, E. (2013). A PSO-based model to increase the accuracy of software development effort estimation. *Softw. Qual. J.* 21, 501–526. doi: 10.1007/s11219-012-9183-x

Li, M., Tian, J., and Chen, F. (2008). Improving multiclass pattern recognition with a co-evolutionary RBFNN. *Pattern Recognit. Lett.* 29, 392–406. doi: 10.1016/j.patrec.2007.10.019

Li, M. F., Fan, L., Liu, H. B., Wu, W., and Chen, J. L. (2012). Impact of time interval on the ångström-Prescott coefficients and their interchangeability in estimating radiation. *Renew. Energy* 44, 431–438. doi: 10.1016/j.renene.2012.01.107

Li, X., Wang, L., and Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Eng. Appl. Artif. Intell.* 21, 785–795. doi: 10.1016/j.engappai.2007.07.001

Linares-Rodríguez, A., Ruiz-Arias, J. A., Pozo-Vázquez, D., and Tovar-Pescador, J. (2011). Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks. *Energy* 36, 5356–5365. doi: 10.1016/j.energy.2011.06.044

Lu, N., Qin, J., Yang, K., and Sun, J. (2011). A simple and efficient algorithm to estimate daily global solar radiation from geostationary satellite data. *Energy* 36, 3179–3188. doi: 10.1016/j.energy.2011.03.007

Mahanty, R. N., and Dutta Gupta, P. B. (2004). Application of RBF neural network to fault classification and location in transmission lines. *IEE Proc. Gener. Transm. Distrib.* 151, 201–212. doi: 10.1049/ip-gtd:20040098

Makade, R. G., Chakrabarti, S., and Jamil, B. (2019). Prediction of global solar radiation using a single empirical model for diversified locations across India. *Urban Clim.* 29:100492. doi: 10.1016/j.uclim.2019.100492

Meenal, R., and Selvakumar, A. I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renew. Energy* 121, 324–343. doi: 10.1016/j.renene.2017.12.005

Mellit, A. (2008). Artificial Intelligence technique for modelling and forecasting of solar radiation data: a review. *Int. J. Artif. Intell. Soft Comput.* 1:52. doi: 10.1504/ijaisc.2008.021264

Mishra, M., Byomakesha Dash, P., Nayak, J., Naik, B., and Kumar Swain, S. (2020). Deep learning and wavelet transform integrated approach for short-term solar PV power prediction. *Meas. J. Int. Meas. Confed.* 166:108250. doi: 10.1016/j.measurement.2020.108250

Mohammadi, K., Shamshirband, S., Petkoviæ, D., and Khorasanizadeh, H. (2016). Determining the most important variables for diffuse solar radiation prediction using adaptive neuro-fuzzy methodology; Case study: city of Kerman, Iran. *Renew. Sustain. Energy Rev.* 53, 1570–1579. doi: 10.1016/j.rser.2015.09.028

Moreno, A., Gilabert, M. A., and Martínez, B. (2011). Mapping daily global solar irradiation over Spain: a comparative study of selected approaches. *Sol. Energy* 85, 2072–2084. doi: 10.1016/j.solener.2011.05.017

Nathans, L., Oswald, F. L., and Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *Pract. Assesment Res. Eval.* 17, 1–19. doi: 10.3102/00346543074004525

Nguyen-Tuong, D., Seeger, M., and Peters, J. (2009). Model learning with local Gaussian process regression. *Adv. Robot.* 23, 2015–2034. doi: 10.1163/016918609X12529286896877

Nielsen, D. (2016). *Tree Boosting With XGBoost: Why Does XGBoost Win Every Machine Learning Competition?* Ph. D. Thesis. Trondheim: Norwegian University of Science and Technology. doi: 10.1111/j.1758-5899.2011.00096.x

Ohunakin, O. S., Adaramola, M. S., Oyewola, O. M., Matthew, O. J., and Fagbenle, R. O. (2015). The effect of climate change on solar radiation in Nigeria. *Sol. Energy* 116, 272–286. doi: 10.1016/j.solener.2015.03.027

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petkoviæ, D., and Sudheer, C. (2015). A support vector machine-firefly algorithm-based model for global solar radiation prediction. *Sol. Energy* 115, 632–644. doi: 10.1016/j.solener.2015.03.015

Pang, Z., Niu, F., and O'Neill, Z. (2020). Solar radiation prediction using recurrent neural network and artificial neural network: a case study with comparisons. *Renew. Energy* 156, 279–289. doi: 10.1016/j.renene.2020.04.042

Persson, C., Bacher, P., Shiga, T., and Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees. *Sol. Energy* 150, 423–436. doi: 10.1016/j.solener.2017.04.066

Prescott, J. A. (1940). Evaporation from a water surface in relation to solar radiation. *Trans. R. Soc. South Aust.* 61, 114–118. Available online at: https://ci.nii.ac.jp/naid/10025613338/en/

Quinlan, J. R. (1999). Simplifying decision trees. *Int. J. Hum. Comput. Stud.* 51, 497–510. doi: 10.1006/ijhc.1987.0321

Salazar, G. A. (2011). Estimation of monthly values of atmospheric turbidity using measured values of global irradiation and estimated values from CSR and Yang Hybrid models. *Study case: Argentina. Atmos. Environ.* 45, 2465–2472. doi: 10.1016/j.atmosenv.2011.02.048

Shamshirband, S., Mohammadi, K., Tong, C. W., Zamani, M., Motamedi, S., and Ch, S. (2016). A hybrid SVM-FFA method for prediction of monthly mean global solar radiation. *Theor. Appl. Climatol.* 125, 53–65. doi: 10.1007/s00704-015-1482-2

Shamshirband, S., Mosavi, A., Rabczuk, T., Nabipour, N., and Chau, K. W. (2020). Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. *Eng. Appl. Comput. Fluid Mech.* 14, 805–817. doi: 10.1080/19942060.2020.1773932

Shamshirband, S., Rabczuk, T., and Chau, K. W. (2019). A survey of deep learning techniques: application in wind and solar energy resources. *IEEE Access.* 7, 164650–164666. doi: 10.1109/ACCESS.2019.2951750

Shen, H., and Chou, K. C. (2005). Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* 334, 288–292. doi: 10.1016/j.bbrc.2005.06.087

Siingh, D., Singh, R. P., Singh, A. K., Kulkarni, M. N., Gautam, A. S., and Singh, A. K. (2011). Solar activity, lightning and climate. *Surv. Geophys.* 32, 659–703. doi: 10.1007/s10712-011-9127-1

Suehrcke, H., Bowden, R. S., and Hollands, K. G. T. (2013). Relationship between sunshine duration and solar radiation. *Sol. Energy* 92, 160–171. doi: 10.1016/j.solener.2013.02.026

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., et al. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Convers. Manag.* 119, 121–129. doi: 10.1016/j.enconman.2016.04.051

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform.* 4, 159–169. doi: 10.1007/s40708-017-0065-7

Trappey, A. J. C., Hsu, F. C., Trappey, C. V., and Lin, C. I. (2006). Development of a patent document classification and search platform using a back-propagation network. *Expert Syst. Appl.* 31, 755–765. doi: 10.1016/j.eswa.2006.01.013

Van Ooyen, A., and Nienhuis, B. (1992). Improving the convergence of the back-propagation algorithm. *Neural Netw.* 5, 465–471. doi: 10.1016/0893-6080(92)90008-7

Voyant, C., Muselli, M., Paoli, C., and Nivet, M. L. (2012). Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. *Energy* 39, 341–355. doi: 10.1016/j.energy.2012.01.006

Wang, L., Kisi, O., Zounemat-Kermani, M., Salazar, G. A., Zhu, Z., and Gong, W. (2016). Solar radiation prediction using different techniques: model evaluation and comparison. *Renew. Sustain. Energy Rev.* 61, 384–397. doi: 10.1016/j.rser.2016.04.024

Wang, R. (2012). AdaBoost for feature selection, classification and its relation with SVM, a review. *Phys. Proc.* 25, 800–807. doi: 10.1016/j.phpro.2012.03.160

Wild, M. (2009). Global dimming and brightening: a review. *J. Geophys. Res. Atmos.* 114:D00D16. doi: 10.1029/2008JD011470

Will, A., Bustos, J., Bocco, M., Gotay, J., and Lamelas, C. (2013). On the use of niching genetic algorithms for variable selection in solar radiation estimation. *Renew. Energy* 50, 168–176. doi: 10.1016/j.renene.2012.06.039

Xue, X. (2017). Prediction of daily diffuse solar radiation using artificial neural networks. *Int. J. Hydrogen Energy* 42, 28214–28221. doi: 10.1016/j.ijhydene.2017.09.150

Yang, H. T., Huang, C. M., Huang, Y. C., and Pai, Y. S. (2014). A weather-based hybrid method for 1-day ahead hourly forecasting of PV power output. *IEEE Trans. Sustain. Energy* 5, 917–926. doi: 10.1109/TSTE.2014.2313600

Yang, K., Huang, G. W., and Tamai, N. (2001). Hybrid model for estimating global solar radiation. *Sol. Energy* 70, 13–22. doi: 10.1016/S0038-092X(00)00121-3

Younes, S., Claywell, R., and Muneer, T. (2005). Quality control of solar radiation data: present status and proposed new approaches. *Energy* 30, 1533–1549. doi: 10.1016/j.energy.2004.04.031

Zeng, Z., Wang, Z., Gui, K., Yan, X., Gao, M., Luo, M., et al. (2020). Daily global solar radiation in China estimated from high-density meteorological observations: a random forest model framework. *Earth Space Sci.* 7:e2019EA001058. doi: 10.1029/2019EA001058

Zhang, Y., and Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* 58, 308–324. doi: 10.1016/j.trc.2015.02.019

Zhu, J., Zou, H., Rosset, S., and Hastie, T. (2006). Multi-class AdaBoost. *Stat. Interface* 2, 349–360. doi: 10.4310/SII.2009.v2.n3.a8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer YY declared a past co-authorship with one of the authors ZZ to the handling editor.