Energy Conversion and Management 220 (2020) 113111

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Improving solar radiation estimation in China based on regional optimal combination of meteorological factors with machine learning methods

Chuan He[a,b,c], Jiandong Liu[b], Fang Xu[c,d], Teng Zhang[c,d], Shang Chen[c,d], Zhe Sun[c,d], Wenhui Zheng[c,d], Runhong Wang[c,d], Liang He[e], Hao Feng[a,d], Qiang Yu[a], Jianqiang He[a,c,d,f,*]

[a] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Northwest A&F University, Yangling 712100, Shaanxi, China
[b] State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, China
[c] Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, Shaanxi, China
[d] Institute of Water-Saving Agriculture in Arid Areas of China, Northwest A&F University, Yangling 712100, Shaanxi, China
[e] National Meteorological Center, Beijing 100081, China
[f] Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Meteorological Bureau, Xian 710014, Shaanxi, China

ARTICLE INFO

Keywords:
Solar radiation
Machine learning
Climatic zones
Input combination
Meteorological factor

ABSTRACT

The values of global solar radiation are important fundamental data for potential evapotranspiration estimation, solar energy utilization, climate change study, crop growth model, and etc. This research tried to explore the optimal combination of input meteorological factors and the machine learning methods for the estimation of daily solar radiation under different climatic conditions so as to improve the estimation accuracy. Based on the correlation between meteorological factors, different meteorological factor input combinations were established and the support vector machine method was used to estimate global solar radiation at 80 weather stations in four climatic regions of China mainland. The results showed that, the optimal combinations of input meteorological factors were different in the four different climatic zones in China mainland. Three meteorological factors of sunshine hours, extraterrestrial radiation, and air temperature had greater impacts on the solar radiation estimation. Adding the factor of precipitation could obviously improve the estimation accuracy in humid regions, but not remarkably in arid regions. Wind speed had very little influence on solar radiation estimation. The accuracies of machine learning methods were better than the Angstrom-Prescott formula and the multiple linear regression method. Among them, support vector machine and extreme learning machine were more appropriate. In some sites, the root mean square error of support vector machine method was even 20% less than that of the Angstrom-Prescott formula. In general, reasonable division of the areas and establishment of appropriate input combinations of meteorological factors according to the climatic conditions, combined with machine learning methods, can effectively improve the accuracy of solar radiation estimation.

1. Introduction

Global solar radiation is the main source of energy on the earth, as well as the basic driving force for various physical and biological processes on the earth surface [1]. Many natural phenomena on the earth are mainly caused by the difference, transformation, and transportation of solar radiation energy. Global solar radiation is of great importance to many research fields, such as reference evapotranspiration estimation [2], solar energy utilization [3], climate change [4], and crop growth models [5]. However, radiation observation equipment is usually expensive to construct and maintain, which makes solar radiation observation not as easy as sunshine hours, temperature, precipitation, and etc [6]. At present, among the more than 2000 national meteorological stations in China mainland, only about 100 stations have continuous observations of solar radiation. Thus, the limited number of existing observation stations of solar radiation can hardly meet the needs of scientific research and production [7]. Solar energy as a clean energy has been given a full attention [8]. The mainland of China has abundant solar energy resources [9] and Chinese government has also formulated a series of energy policies. Thus, the accurate estimation of global solar radiation is helpful for the development of new energy-related industries in China [10].

* Corresponding author at: State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Northwest A&F University, Yangling 712100, Shaanxi, China.
E-mail address: jianqiang_he@nwsuaf.edu.cn (J. He).

https://doi.org/10.1016/j.enconman.2020.113111
Received 1 April 2020; Received in revised form 10 June 2020; Accepted 13 June 2020

Nomenclature

Variables

R_s	global solar radiation (MJ m^{-2} d^{-1})
R_a	extraterrestrial radiation (MJ m^{-2} d^{-1})
n	sunshine hours (h)
N	maximum possible sunshine duration (h)
T_{max}	maximum temperature (°C)
T_{min}	minimum temperature (°C)
T_{mean}	average temperature (°C)
Δt	diurnal temperature range (°C)
P	Precipitation (mm)
P_t	1 for rainfall > 0; 0 for rainfall < 0
U	wind speed (m s^{-1})
RH	relative humidity (%)
d_r	inverse square of the relative distance earth to sun
ω_s	sunset hour angle (rad)
φ	latitude (rad)
δ	solar declination (rad)
J	day of the year

Abbreviations

MPZ	mountain plateau zone
SMZ	subtropical monsoon zone
TMZ	temperate monsoon zone
TCZ	temperate continental zone
SVM	support vector machine
GBDT	gradient boosting decision tree
MARS	multivariate adaptive regression spline
ELM	extreme learning machine
A-P	Angstrom-Prescott formula
R^2	decision coefficient
RMSE	root mean square error
AICc	akaike's information corrected criterion
$RMSE_{reduction}$	RMSE reduction value of SVM compared to A-P formulate

Constants

a, b	empirical coefficients
G_{sc}	= 0.082 MJ m^{-2} min^{-1}, solar constant

In order to solve the problem of insufficient observations of solar radiation, previous researchers usually used empirical models [11], machine learning models [12], and satellite-based methods [13] to estimate global solar radiation. The empirical model and machine learning model are more commonly used in practice because of their low cost and high estimation accuracy [14]. Over the past several decades, scientists in various countries have established different empirical models to estimate global solar radiation, including sunshine-hour-based models, temperature-based models, and models combining various meteorological factors [15]. According to previous studies, empirical models based on sunshine duration were generally better than the models based on temperature or other single meteorological factors [16]. The Angstrom-Prescott formula, which links the relative sunshine hours with the clear sky index, is the most widely used estimation method in the world. Thereafter, the later empirical models of solar radiation estimation were more or less based on the transformation of the Angstrom-Prescott model or the introduction of other meteorological factors [17]. In addition, there were many studies to calibrate and validate the Angstrom-Prescott model in different parts of the world [18]. However, traditional empirical models were not able to deal with the complex non-linear relationship between variables and other abnormal conditions [19]. In recent years, machine learning methods have been widely used in many fields with the development of computer technology [20]. In terms of solar radiation estimation, empirical models were not able to completely meet the different needs since meteorological data were always incomplete and unavailable in targeted regions. Research by Tymvios et al. [21] showed that the accuracy of the artificial neural network (ANN) method was better than the Angstrom formula. Chen et al. [22] compared the support vector machine (SVM) method with other empirical models and found that the error of the SVM method was smaller in solar radiation estimation based on temperature data. Thus, machine learning methods have become a promising way for solar radiation estimation due to its high accuracy and flexible combination of input variables.

When estimating solar radiation based on other common meteorological data and machine learning methods, it was necessary to select several relevant meteorological factors as model inputs. Chen and Li [23] used SVM to estimate solar radiation with the inputs of sunshine hours, temperature, relative humidity, and vapor pressure. It was found that the combination of sunshine hours and temperature had the highest estimation accuracy, while the input combination without sunshine hours had poor accuracy. The results of the research by Fan et al. [24] showed that the estimation accuracy of solar radiation by the input of sunshine duration, temperature, and precipitation was better than the results of using only sunshine duration. Meenal and Selvakumar [25] identified month, latitude, maximum temperature and sunshine hours as the most influential and relative humidity as the least influential input parameters in solar radiation estimation. In addition, some relevant studies showed that the inclusion of precipitation could help improve the estimation accuracy of solar radiation. Both the amount of rainfall (mm) and the binary form of rainfall event (1 for rainfall and 0 for no rainfall) were widely used [26]. The above research showed that different input combinations of meteorological factors might have great impacts on the estimation results. Yadav and Chandel [27] proposed that the prediction accuracy of neural network model depended on input parameter combination, training algorithm and architecture configuration, which also illustrated the importance of selecting appropriate input meteorological factors. Due to the different climate conditions in different regions, the correlation between local meteorological factors and global solar radiation was different. The input combinations of the best meteorological factors may also differ between regions, and the accuracy obtained with the same method of radiation estimation was also different. Alizamir et al. [28] used six machine learning models to estimate solar radiation. With the same method, the estimation errors of the Turkish site were larger than that of the US site. China has a vast area and complex internal climate [29]. However, previous studies usually focused on a specific region in China [30] or generally took the whole country as a single region [31], which did not reflect the differences within China mainland, So it is necessary to divide China mainland into several different regions based on the climatic conditions and then conduct the relevant research.

Current research focused mainly on the improvement of radiation estimation methods, while few studies concentrated on the selection of meteorological factors required for the models [32]. If there was only one set of input combinations, it obviously could not meet the estimation requirements under different meteorological conditions. It is still unclear what are the optimal combinations of input meteorological factors and the best machine learning methods for solar radiation estimation in different climatic regions of China. Based on climatic characteristics, this study divided China mainland into four different climatic zones, and applied different methods to estimate daily solar radiation for each of them. The objectives were to (1) explore the correlations among different meteorological factors in different climatic zones of China mainland; (2) to obtain the optimal combinations of

input meteorological factors required for the solar radiation estimation based on machine learning methods in different climatic regions of China mainland; and (3) to assess the estimation accuracies of different machine learning methods based on the optimal combinations of input meteorological factors determined above in China mainland. Finally, the estimation accuracy of daily global solar radiation will be improved in China mainland.

2. Materials and methods

In this study, China mainland was divided into four different climatic regions, namely the mountain plateau zone (MPZ), the subtropical monsoon zone (SMZ), the temperate monsoon zone (TMZ) and the temperate continental zone (TCZ) based on local temperature, precipitation, latitude, and longitude (Fig. 1) [33]. The average altitudes of the four climatic zones were 4236 m, 611 m, 288 m, and 912 m above sea level, respectively. TCZ is an arid region with an average annual precipitation of 193 mm; SMZ is a humid region with an average annual precipitation of 1360 mm; TMZ and MPZ have average annual precipitations of 591 and 460 mm, respectively.

2.1. Dataset

The meteorological data of a total of 80 meteorological stations were collected for the four different climatic regions of China mainland (Fig. 1 and Appendix A), including maximum temperature (T_{max}), minimum temperature (T_{min}), average temperature (T_{mean}), precipitation (P), wind speed (U), relative humidity (RH), and daily global solar radiation (R_s). Daily extraterrestrial radiation (R_a) and maximum possible sunshine duration (N) were calculated with site latitude, solar constant, solar declination, and date of the year [34]. The variable Δt was the diurnal temperature range, which was defined as the difference between the daily maximum and minimum temperatures. The variable P_t was a piecewise function of rainfall event, where $P_t = 1$ when $P > 0$ and $P_t = 0$ when $P = 0$. The meteorological data were obtained from the National Meteorological Information Center of China Meteorological Administration. The incomplete and abnormal data were deleted from the dataset.

2.2. Solar radiation estimation models

In this study, we investigated six different types of solar radiation estimation models (Models 1–6), which belonged to three categories of empirical formulas, machine learning methods, and multiple linear regression method.

(1) Angstrom-Prescott formula (Model 1)

The Angstrom-Prescott (A-P) empirical model, proposed by Angstrom [35] and further revised by Prescott [36], is the most widely used model of global solar radiation estimation (Eq. (1)).

$$\frac{R_s}{R_a} = a + b\frac{n}{N} \tag{1}$$

where R_a and N were calculated with the method recommended by FAO [34] (Eqs. (2)–(6)).

$$R_a = (24 \times 60/\pi)G_{sc}d_r(\omega_s \sin\varphi \sin\delta + \cos\varphi \cos\delta \sin\omega_s) \tag{2}$$

$$d_r = 1 + 0.033\cos(2\pi \times J/365) \tag{3}$$

$$\delta = 0.409\sin(2\pi \times J/365 - 1.39) \tag{4}$$

$$\omega_s = \arccos(-\tan\varphi \tan\delta) \tag{5}$$

$$N = 24 \times \omega_s/\pi \tag{6}$$

where R_s is the daily global solar radiation, MJ m^{-2} d^{-1}; R_a is the daily extraterrestrial radiation, MJ m^{-2} d^{-1}; a and b are the empirical coefficients; n is the sunshine duration, h; N is the maximum possible sunshine duration, h; $G_{sc} = 0.082$ MJ m^{-2} min^{-1} is the solar constant; d_r is the inverse square of the relative distance earth to sun; ω_s is the sunset hour angle, rad; φ is the latitude, *rad*; δ is the solar declination, rad; and J is the day of the year.

(2) Support vector machine (Model 2)

The support vector machine (SVM) algorithm proposed by Vapnik [37] is a supervised machine learning method for data analysis and pattern recognition, and it has been widely employed for solar radiation estimation [38]. The SVM is based on the principle of structural risk

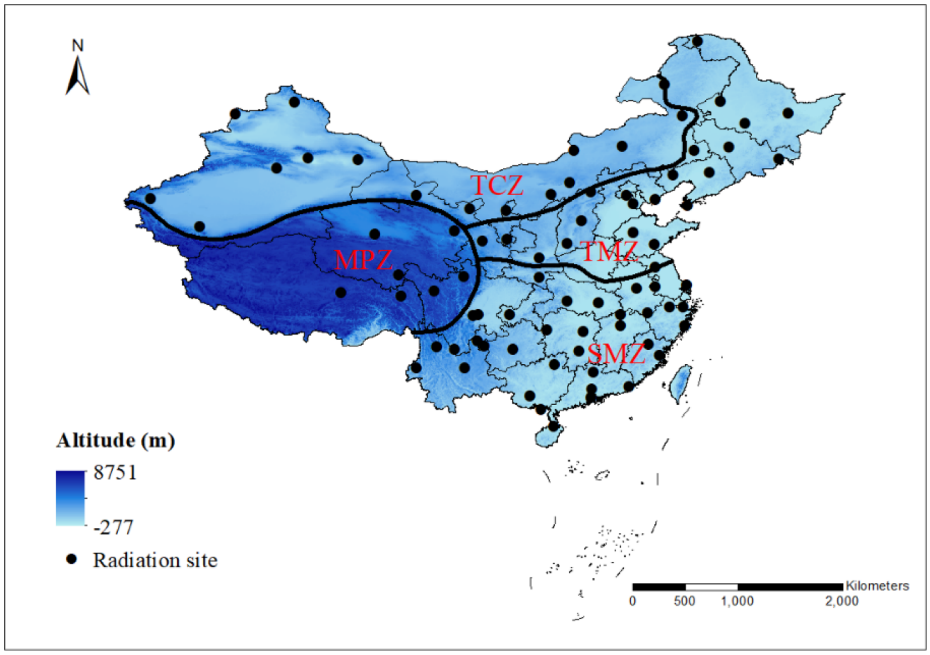

Fig. 1. Distribution of the 80 national meteorological stations (black dots) that have long-term continuous observations of solar radiation in the mainland of China. The whole mainland was divided into four different climatic zones, where the acronym of MPZ represents the mountain plateau zone, SMZ the subtropical monsoon zone, TMZ the temperate monsoon zone, and TCZ the temperate continental zone, respectively. The same below.

minimization, that is, the empirical risk is minimized and the confidence interval is also small, so it has a good generalization for future samples. Therefore, this method can better solve the problems of small samples, nonlinearity and high dimensionality, and is often used for identification and prediction. In this study, the '*kernlab*' [39] package in R language was used to conduct the SVM-based solar radiation estimation.

(3) Gradient boosting decision tree (Model 3)

The gradient boosting decision tree (GBDT) algorithm proposed by Friedman [40] is an integrated decision tree model based on the boosting algorithm. The boosting algorithm generates a base learner based on the residuals from the previous training. Based on boosting, the GBDT establishes a new decision tree in the gradient direction of residual reduction. By generating the optimal tree set, the overall prediction performance of the GBDT model is improved, which is beneficial to handle imbalanced event duration data [41]. In recent years, the GBDT algorithm has also been used in solar radiation estimation [42]. In this study, the '*gbm*' [43] package in R language was used to conduct the GBDT-based solar radiation estimation.

(4) Multivariate adaptive regression spline (Model 4)

The multivariate adaptive regression spline (MARS) algorithm developed by Friedman [44] is a regression analysis method that has a strong generalization ability and specializes in processing high-dimensional data. The method uses the tensor product of the spline function as the basis function, and is divided into three steps: forward process, backward pruning process and model selection. Its advantages are that it can process large amounts of data and high-dimensional data, and has fast calculation and accurate model. Currently, the MARS algorithm has also been used in estimations of evaporation [45] and solar radiation [46]. In this study, the '*earth*' [47] package in R language was used to conduct the MARS-based solar radiation estimation.

(5) Extreme learning machine (Model 5)

The extreme learning machine (ELM) algorithm is a type of machine learning system or method based on feedforward neural networks, which is suitable for supervised and unsupervised learning problems. The ELM was proposed by Huang et al. [48] of Nanyang Technological University and published in the IEEE International Joint Conference. The ELM model consisted of three layers: an input layer, a hidden layer, and an output layer. The biggest feature of ELM is that it is faster than traditional learning algorithms under the premise of ensuring learning accuracy. The ELM algorithm has also been used in solar radiation estimation [49]. In this study, the '*elmNNRcpp*' [50] package in R language was used to conduct the ELM-based solar radiation estimation.

(6) Multiple linear regression (Model 6)

The multiple linear regression (MLR) is a common method to study the relationship between multiple independent variables and a dependent variable. It is also widely used in solar radiation estimation research [51]. The basic task of multiple linear regression analysis is to establish multiple linear regression equations for dependent variables and multiple independent variables based on actual observations (Eq. (7)).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad (7)$$

where, β_0, β_1, β_2, β_3 and β_k, are the regression coefficients; x_1, x_2, x_3 and x_k the correlation parameters; and k is the number of correlation parameters.

2.3. Determination of the optimal combination of input meteorological factors and radiation estimation models

Firstly, the correlations among meteorological factors were calculated based on obtained meteorological data, and then the importance of meteorological factors in the process of solar radiation estimation was evaluated. Next, according to the correlations among meteorological factors and solar radiation in different climatic regions of China, relevant meteorological factors were added successively to form different combinations of input meteorological factors. Third, SVM was used to estimate R_s in different climate regions. Fourth, the errors between the estimated and the observed values of R_s were compared and the optimal combinations of input meteorological factors were obtained for different climatic regions. The SVM method was implemented through the '*kernlab*' packages in R language [52]. The coefficients of the Angstrom-Prescott model and the multiple linear regression method were obtained using the least squares method in R language. Finally, different estimation methods (Models 1–6) were used to estimate the global solar radiation in the four different climate zones, and the estimation errors were calculated based on the optimal combination of the above input meteorological factors. The average error of all weather stations in each climatic region was calculated as the general error to evaluate different solar radiation estimation methods. The method with the minimum error was chosen as the best estimation method in this region.

For the sake of brevity, this study only took the SVM method as an example to introduce the main procedures in R_s estimation with the machine learning method as follows.

(1) Standardize meteorological data at each single site (Eq. (8))

$$x_n = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (8)$$

where x_n is the normalized data; x_i is the raw data; x_{max} and x_{min} are the maximum and minimum values of the raw data.

(2) Determine the ranges of key parameters in the SVM method through the trial and error approach and then use the grid search approach to select the optimal parameter values, while the remaining parameters were set to default values.
(3) Conduct stratified sampling for each site according to the distribution of Rs; use the five folds cross validation method to calculate the error of each single site; randomly divide the data into five completely separated parts, so-called five folds. For each fold of the five folds (one fold contains 20% of the total data), the remaining 80% data were used to construct the model, while its own 20% data were used to evaluate the model constructed.
(4) Repeat the processes of model training and evaluating five times or conduct five different rounds of trainings and tests; calculate the mean value of the five evaluation indicators to evaluate the performance of the model constructed.
(5) Calculate the average error of all of the related weather stations in a given climatic region to represent the general estimation error of daily global solar radiation in this region under a certain combination of input meteorological factors. The combination with the least error was chosen as the optimal combination of input meteorological factors for the region.

2.4. Statistical analysis

Four common statistics were used to evaluate the accuracy and consistency of different estimation models of daily global solar radiation, including the decision coefficient (R^2, Eq. (9)), the root mean square error (RMSE, Eq. (10)), the Akaike's information corrected criterion (AICc, Eq. (11)), and the relative error of RMSE between Angstrom-Prescott formula and SVM method ($RMSE_{reduction}$, Eq. (12)).

Generally, the closer R^2 is to 1, the higher the model fit; the smaller the RMSE, the smaller the model deviation; AIC was developed by Japanese statistician Hirotsugu Akaike [53] and was mainly used for the trade-off between fitness and complexity of the model [54]. AICc is a modified algorithm proposed by McQuarrie and Tsai [55] on the basis of AIC and has been used to select the optimal model identified by its minimum value. In order to better describe the difference between the RMSE for different methods, this study proposed the concept of $RMSE_{reduction}$, which was essentially the relative error of RMSE between A-P method and SVM method. The larger the $RMSE_{reduction}$ value, the higher the accuracy of the SVM method compared to the Angstrom-Prescott formula.

$$R^2 = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2} \tag{10}$$

$$AICc = \ln\frac{RSS}{n-k} + \frac{n+k}{n-k-2} \tag{11}$$

$$RMSE_{reduction} = \frac{|RMSE_{SVM} - RMSE_{A-P}|}{RMSE_{A-P}} \times 100\% \tag{12}$$

where Y_i is R_s value on the i-th day calculated by the model; X_i is the measured R_s value on the i-th day; $\bar{Y}$ is the average of Y_i; $\bar{X}$ is the average of X_i; n is the data sample size; k and RSS are the number of parameters and the residual sum of squares from the model, respectively; $RMSE_{A\text{-}P}$ is the RMSE value of the Angstrom-Prescott formula; and $RMSE_{SVM}$ is the RMSE value of the support vector machine method.

3. Results

The correlation and replaceability of meteorological factors in the process of global solar radiation estimation were explored, the prediction accuracy of different meteorological factor input combinations and different machine learning algorithms were evaluated.

3.1. Correlations among meteorological factors in different climatic zones

The correlations among different meteorological factors and the correlations between single meteorological factor and R_s were analyzed for the four different climatic regions of China mainland (Fig. 2). The relationships between daily global solar radiation R_s and the input

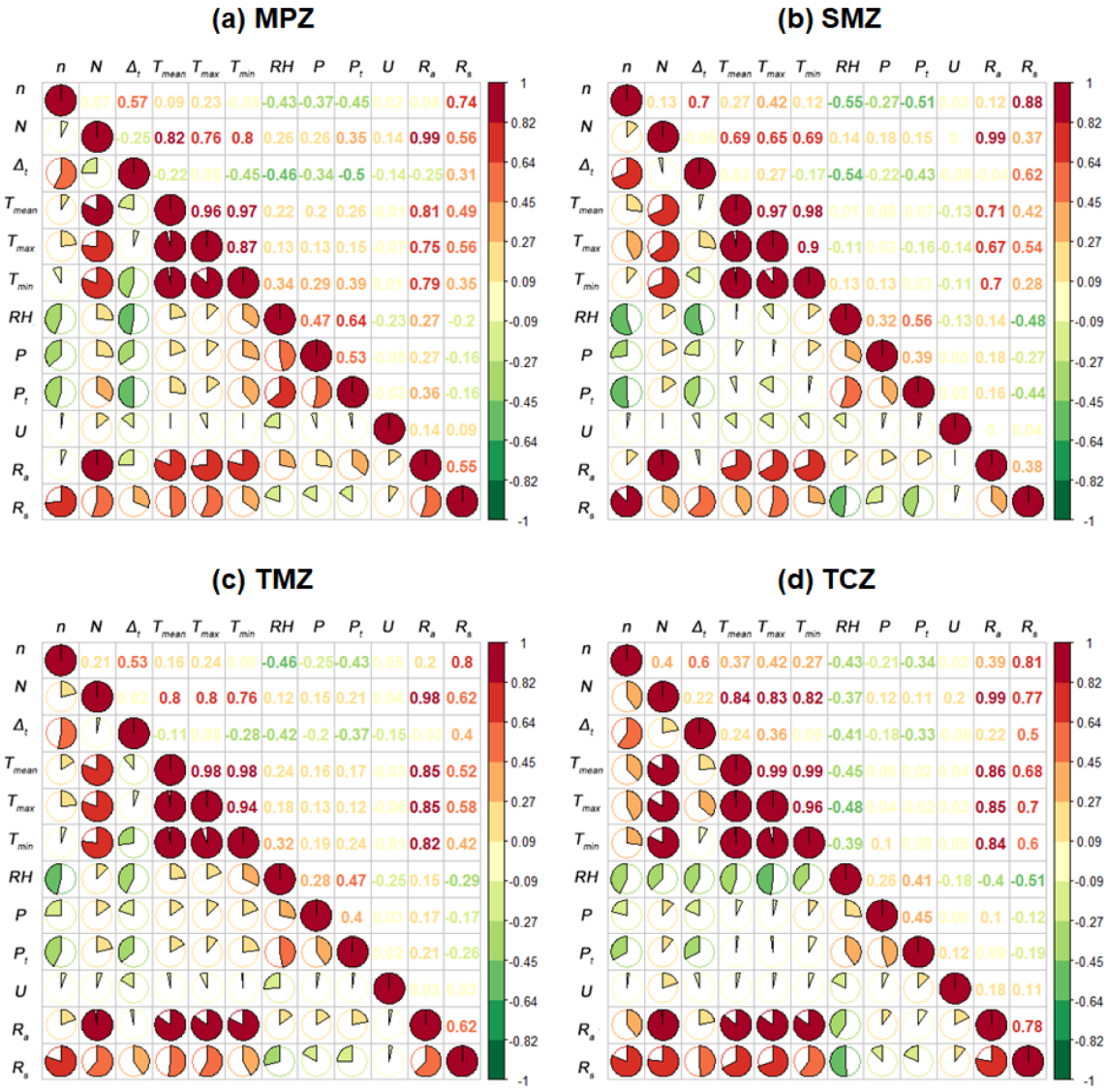

Fig. 2. Correlations among meteorological elements in the four climatic regions of China.

Table 1
Root mean square error (RMSE, MJ m^{-2} d^{-1}) of global solar radiation (R_s) estimation with the support vector machine (SVM) method under different combinations of input meteorological factors of possible maximum sunshine duration (N), sunshine duration hours (n) and horizontal extra-terrestrial radiation (R_a) in the four different climatic regions of China mainland.

Input combination	MPZ	SMZ	TMZ	TCZ
N	5.5021	6.7855	5.8047	4.9406
R_a	5.5005	6.8034	5.8033	4.9407
n + N	2.6488	2.6295	2.3520	2.1036
n + R_a	2.6480	2.6286	2.3495	2.0983
n + N + R_a	2.6455	2.6216	2.3467	2.0929

meteorological factors were comprehensively evaluated, including sunshine duration hours n, h; maximum possible sunshine duration N, h; diurnal temperature range Δt, ˚C; daily average temperature T_{mean}, ˚C; daily maximum temperature T_{max}, ˚C; daily minimum temperature T_{min}, ˚C; relative humidity RH, %; precipitation P, mm; rainfall event P_t; average wind speed U, m s^{-1}; and daily extraterrestrial solar radiation R_a, MJ m^{-2} d^{-1}. The lower triangle of the matrix indicated the correlation between meteorological elements and the upper triangle indicated the corresponding correlation coefficients. Positive correlation was shown in red and negative correlation in green. The color intensity and the size of the sector area were proportional to the correlation coefficient. The results showed that the correlations among different meteorological factors in the four different climatic regions were also different. The rankings of correlations between meteorological factors and R_s was: $n > T_{max} > N > R_a > T_{mean} > T_{min} > \Delta t > RH > P > P_t > U$ in the MPZ (or the mountain plateau zone); $n > \Delta t > T_{max} > RH > Pt > T_{mean} > R_a > N > T_{min} > P > U$ in the SMZ (or the subtropical monsoon zone); $n > R_a > N > T_{max} > T_{mean} > T_{min} > \Delta t > RH > P_t > P > U$ in the TMZ (or the temperate monsoon zone); and $n > R_a > N > T_{max} > T_{mean} > T_{min} > RH > \Delta t > P_t > P > U$ in the TCZ (or the temperate continental zone). The meteorological factors of n, N, R_a, T_{mean}, T_{max}, T_{min} were positively correlated with R_s, while the factors of RH, P and P_t were negatively correlated.

Generally, the correlation between n and R_s was the strongest in each climatic zone, while the correlation between U and R_s was the weakest. The correlation ranking of other meteorological factors was slightly different in different zones. The correlation coefficient between n and R_s differed obviously since the highest was 0.88 in SMZ and the lowest was 0.74 in MPZ, namely the SMZ zone was most affected by sunshine hour n. Since the variable n can reflect the amount of cloud to some extent, the R_s in the SMZ was remarkably affected by the cloud amount. At the same time, the MPZ is located on the mountain and plateau and the weather is relatively sunny. Thus, the R_s in MPZ was less affected by clouds.

In TMZ and TCZ, R_a and N were the most related meteorological factors besides n. The correlation coefficients between the two variables and R_s were 0.62 and 0.62 in TMZ and 0.77 and 0.78 in TCZ, respectively. In MPZ, the correlation coefficients between N, T_{max}, R_a and R_s were 0.56, 0.56 and 0.55, respectively. In SMZ, the correlation coefficients of R_a, N and R_s were 0.38 and 0.37, respectively, which were weaker than the correlation between n, Δt, T_{max}, RH, P_t, T_{mean} and R_s. This indicated that the main meteorological factors affecting R_s were sunshine hour n and geographical factors in TMZ, TCZ and MPZ, but R_s was also strongly affected by other meteorological factors (e.g. temperature, air humidity, and precipitation) in SMZ. This is because SMZ is located on the southeast coast of China, with abundant precipitation and humid air. These climatic conditions had a greater impact on solar radiation. In addition, Δt in SMZ was the smallest among the four regions, but the correlation between Δt and R_s was higher than the other temperature factors. This was probably because Δt could better reflect the changes of R_s in this region.

In general, except for n and N, the correlations of temperature (T_{mean}, T_{max}, T_{min}), RH with R_s were stronger than the correlations of precipitation and U with R_s in all of the four climatic regions in China mainland. The correlation between temperature and R_s was higher in TCZ than that in other regions, which may be due to the dryer climate in Northwest China and temperature could more directly reflect the change of global solar radiation. In SMZ and TCZ, the correlations of RH and R_s were strong since the correlation coefficients were -0.48 and -0.51, while the correlation coefficients were -0.2 and -0.29 in MPZ and TMZ, respectively. The influence of RH on global solar radiation was weak in MPZ and TMZ regions. However, the correlation between rainfall events P_t and R_s was -0.44 in SMZ, which was stronger than the other three regions. This was probably because there was more precipitation in SMZ region. The correlation between U and R_s was the lowest in all of the four climatic regions, which indicated wind speed had limited influence on global solar radiation.

3.2. Replaceability of meteorological factors in solar radiation estimation

The correlation between N and R_a was very strong (Fig. 2) since the correlation coefficients of N and R_a were close to 1.0 in all of the four regions. Factors of N and R_a can be obtained from the date and the geographic information of the weather station. Therefore, they can replace each other in the process of R_s estimation. Therefore, the combinations of input meteorological factors including N and R_a were used to estimate the global solar radiation using the SVM method. The average $RMSE$ values of model estimation were calculated in the four different climatic regions (Table 1). Generally, the estimation accuracy in TCZ region was the highest; TMZ ranked the second; and the

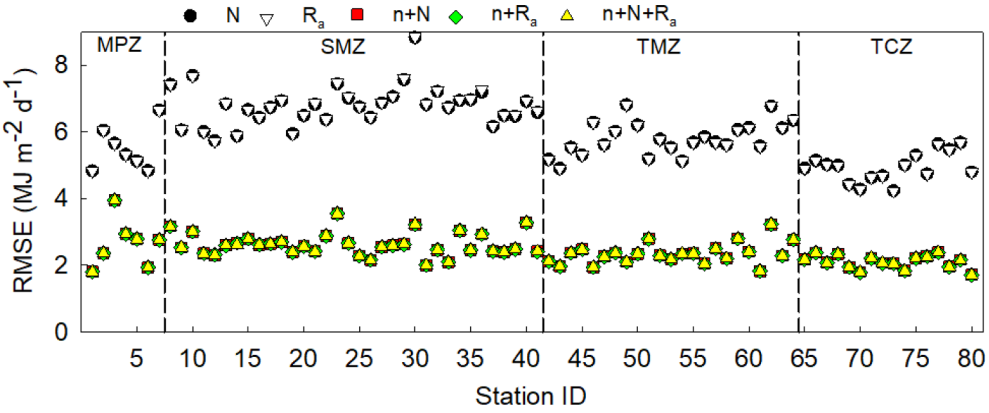

Fig. 3. Root mean square error (RMSE, MJ m^{-2} d^{-1}) of global solar radiation (R_s) estimation using the support vector machine (SVM) method under different combinations of input meteorological factors at the 80 weather stations in China mainland.

estimation accuracy in MPZ and SMZ were relatively low. In all of the four climatic regions, the combination of $n + N + R_a$ was better than the other four combinations (Table 1).

The RMSE values of the 80 meteorological stations in China mainland were further analyzed under different combinations of input meteorological factors (Fig. 3). When only N or R_a was used as the single input factor, the RMSE values were not ideal, and there were large gaps between sites. Then, the estimated error significantly reduced after adding the variable of n. However, the errors were similar for the three combinations of $n + N$, $n + R_a$ and $n + N + R_a$, all between 1.8 and 4.0 MJ m^{-2} d^{-1}. When n was included, adding N or R_a could result in a higher estimation accuracy. The combination of $n + R_a$ had a slightly smaller error than $n + N$ (Table 1). So in the following study, only R_a was chosen as the main input meteorological factor while N was neglected to reduce the complexity in global solar radiation estimation.

For temperature, four different temperature-related meteorological factors T_{max}, T_{mean}, T_{min} and Δt were investigated in this study. Generally, the correlations of the three temperature factors T_{max}, T_{mean} and T_{min} with N and R_a were strong, up to 0.84, while the correlations with other meteorological factors was relatively weak. The correlations among Δt and T_{max}, T_{mean} and T_{min} were very different in the same climatic region, and there were also great differences among different climatic regions. The correlations among T_{mean}, T_{max} and T_{min} were close to 1.0. Thus, there was a redundancy in temperature-related meteorological factors in solar radiation estimation. Based on the replaceability of the four temperature-related variables (Table 2), it can be seen that the estimation error with only R_a as the input factor can be effectively reduced by adding any of the four factors of Δt, T_{max}, T_{mean}, and T_{min}. For the input combinations with T_{max}, T_{mean}, and T_{min} at the same time, adding Δt had some negative impact on the estimation accuracy. Therefore, Δt was not recommended when T_{max}, T_{mean} and T_{min} were already available. The estimation accuracy of the three-factor input combination of $R_a + T_{max} + T_{min}$ was higher than that of the two-factor input combination of $R_a + \Delta t$. This was because Δt only reflected daily temperature changes, while T_{max} and T_{min} could better reflect the amount of radiation received by the atmosphere during this period.

When estimating R_s with R_a and Δt, the estimation accuracy could be slightly improved when T_{max} and T_{mean} were added, namely, the four-factor combination of $R_a + \Delta t + T_{max} + T_{mean}$ could obtain better estimation accuracy. However, the estimation accuracy was not obviously improved after further adding T_{min}. In addition, the estimation accuracy of the four-factor combination of $R_a + \Delta t + T_{max} + T_{min}$ was not significantly improved compared with the three-factor combination of $R_a + \Delta t + T_{max}$, namely, there was no need to add T_{min}. Therefore, the combinations of $T_{max} + T_{mean} + T_{min}$ or $\Delta t + T_{max} + T_{mean}$ should be selected among the temperature-related factors to obtain higher estimation accuracy. Generally, when n was known, the five-factor combination of $n + R_a + T_{max} + T_{mean} + T_{min}$ or $n + R_a + \Delta t + T_{max} + T_{mean}$ was the optimal combination of input meteorological factors for the R_s estimation with SVM method (Table 2).

In the subtropical monsoon zone or SMZ, the correlation coefficient between P_t and R_s was -0.44, but this kind of correlation was not strong in the other climatic regions (Fig. 1). This was because SMZ had more precipitation than other regions and the impact of precipitation on solar radiation was more serious. Since the factor of P_t was derived from precipitation P, the effect of these two meteorological factors on the accuracy of radiation estimation was further analyzed. Based on the estimation errors of different combinations of input meteorological factors including precipitation, it was found that the combination of $R_a + P$ was more accurate than the combination of $R_a + P_t$ if only R_a and precipitation were used as input factors (Table 3). Thus, when temperature or n were available, P_t can be selected as an additional input factor to improve the estimation accuracy. However, the simultaneous use of P and P_t could not obviously improve the estimation accuracy, but even increase the estimation error. According to the estimation results of 80 meteorological stations in China mainland under different input combinations including precipitation related factors (Fig. 4), there were large difference in estimation accuracy among different stations. It can be seen that the addition of Δt and n reduced the error of $R_a + P$ (or P_t). The combination of $n + R_a + P_t$ had the highest estimation accuracy. Therefore, it was recommended the precipitation event P_t as the precipitation-related input factor to estimate R_s.

3.3. Determination of the optimal combinations of input meteorological factors

Based on the different optimal combinations of input meteorological factors, global solar radiation was estimated with the support vector machine method in different climatic regions of China mainland. Then the estimation results were further compared with the estimations by the multiple linear regression method and Angstrom-Prescott formula (Table 4). Since R_a was necessary in the estimation of global solar radiation by conventional methods and it also had a great influence on the estimation accuracy with the SVM method, R_a was added as a meteorological input factor except for n when estimating R_s with the SVM method. The results showed that the estimation accuracy of the input combination of $n + R_a$ was better than that of the Angstrom-Prescott formula in whole China mainland. The estimation accuracy of the SVM was generally higher than the multiple linear regression method under various conditions, which indicated that the SVM method could estimate global solar radiation more accurately in different climatic regions of China mainland. Overall, the simulation accuracy in the TCZ region was the highest, followed by the TMZ region, while the accuracies in the SMZ and MPZ regions were relatively poor. This was probably due to the higher correlation between meteorological factors and R_s in the TCZ region. In SMZ, high estimation accuracy could be obtained only using n, which was consistent with the high correlation between n and R_s in the SMZ region.

In the same climatic zone, the accuracies of R_s estimation with different combinations of input meteorological factors were different. The addition of the same meteorological factor had different effects on the accuracy of R_s estimation in different regions (Table 4). Generally,

Table 2
Root mean square errors (RMSE, MJ m^{-2} d^{-1}) of global solar radiation (R_s) estimations with the support vector machine (SVM) method under different combinations of input meteorological factors including various temperature related factors in the four different climatic regions of China mainland.

Input combination	MPZ	SMZ	TMZ	TCZ
R_a	5.5005	6.8034	5.8033	4.9407
$R_a + T_{max}$	4.9846	5.2781	5.2191	4.6901
$R_a + T_{max} + T_{mean}$	4.3140	4.2254	4.3200	4.2734
$R_a + T_{max} + T_{min}$	3.9112	3.9227	3.9235	3.7427
$R_a + T_{max} + T_{mean} + T_{min}$	3.7448	3.7771	3.8103	3.4506
$R_a + \Delta t$	3.9639	4.2869	4.0471	3.8040
$R_a + \Delta t + T_{max}$	3.9154	3.9337	3.9288	3.7530
$R_a + \Delta t + T_{max} + T_{min}$	3.9153	3.9308	3.9271	3.7519
$R_a + \Delta t + T_{max} + T_{mean}$	3.7663	3.8090	3.8360	3.5137
$R_a + \Delta t + T_{max} + T_{mean} + T_{min}$	3.7622	3.8025	3.8322	3.5025
$n + R_a$	2.6480	2.6286	2.3495	2.0983
$n + R_a + T_{max}$	2.6416	2.5631	2.2954	2.0789
$n + R_a + T_{max} + T_{mean}$	2.6193	2.5148	2.2668	2.0607
$n + R_a + T_{max} + T_{min}$	2.6126	2.4885	2.2527	2.0519
$n + R_a + T_{max} + T_{mean} + T_{min}$	***2.5946***	***2.4593***	***2.2321***	***2.0294***
$n + R_a + \Delta t$	2.6253	2.5194	2.2927	2.0853
$n + R_a + \Delta t + T_{max}$	2.6120	2.4866	2.2571	2.0573
$n + R_a + \Delta t + T_{max} + T_{min}$	2.6122	2.4889	2.2580	2.0579
$n + R_a + \Delta t + T_{max} + T_{mean}$	***2.5955***	***2.4590***	***2.2389***	***2.0400***
$n + R_a + T_{max} + T_{mean} + T_{min} + \Delta t$	2.5948	2.4601	2.2398	2.0397

Note: The bold italics values in the table are the errors in global solar radiation estimations based on the recommended optimal combinations of input meteorological factors.

Table 3
Root mean square errors (RMSE, MJ m^{-2} d^{-1}) of global solar radiation (R_s) estimation by the support vector machine (SVM) method under different combinations of input meteorological factors including precipitation related factors of P and P_t in the four climatic regions of China mainland.

Input combination	MPZ	SMZ	TMZ	TCZ
$R_a + P$	4.8423	5.4500	4.8426	4.3529
$R_a + P_t$	4.8557	5.5276	4.8465	4.3309
$R_a + \Delta t + P$	3.8938	3.8997	3.8327	3.6533
$R_a + \Delta t + P_t$	3.8861	3.9363	3.8272	3.6388
$R_a + \Delta t + P + P_t$	3.8848	3.8354	3.8084	3.6426
$n + R_a + P$	2.6917	2.5630	2.3440	2.1830
$n + R_a + P_t$	***2.6345***	***2.5532***	***2.2754***	***2.0715***
$n + R_a + P + P_t$	2.6828	2.5468	2.3332	2.1814

Note: The bold italic values in the table are the errors in global solar radiation estimations based on the recommended optimal combinations of input meteorological factors.

the optimal combination of input meteorological factors was $n + R_a + T_{max} + T_{mean} + T_{min} + RH$ for the SVM method in MPZ and TCZ. The estimation accuracy even reduced after adding U in MPZ. However, the optimal combination of input meteorological factors was $n + R_a + \Delta t + T_{max} + RH + P_t + T_{mean} + U$ in SMZ, which was the combination of all available meteorological elements. In TMZ, the optimal combination of input meteorological factors was $n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t$. Adding U increased the complexity of the model but did not improve the accuracy substantially. After the existing input combination reached a certain accuracy, the effect of additional variables was not obvious on the improvement of estimation accuracy. For the multiple linear regression method, it was usually to combine all meteorological elements to achieve the best accuracy. However, the SVM method could achieve higher accuracy by using fewer meteorological factors. The SVM method could reduce the number of meteorological factors required in the estimation of global solar radiation. Therefore, the machine learning algorithms such as SVM were more suitable for the situations with fewer types of or missing meteorological data.

The accuracy of solar radiation estimation with the SVM method under the optimal combination of input meteorological factors was higher than that of the traditional Angstrom-Prescott formula (Table 4). The accuracies of two methods were then further analyzed (Fig. 5). The size of green dots in the Fig. 5 represents the $RMSE_{reduction}$ of SVM method under the optimal combination of input meteorological factors relative to the Angstrom-Prescott formula in the four climatic regions in China mainland. $RMSE_{reduction}$ (filled dots) was defined as the ratio (%) of the RMSE difference between the Angstrom-Prescott formula and the SVM method to the RMSE of the Angstrom-Prescott formula (Eq. 12). The larger dot indicated greater improvement of estimation accuracy by the SVM method compared to the Angstrom-Prescott formula. It can be seen that the accuracy of *Rs* estimation was improved by the SVM method in different climatic regions in whole China mainland. The improvement of estimation accuracy was more obvious in the SMZ, followed by TCZ, and the $RMSE_{reduction}$ of many stations in these two regions were greater than 20%. Since the MPZ was located in the Qinghai-Tibet Plateau, the special geographical location and climate conditions made the improvement smaller. The $RMSE_{reduction}$ in MPZ were all less than 15%. In conclusion, the machine learning method could effectively improve the estimation accuracy of R_s in relatively humid regions.

3.4. Comparisons among different machine learning methods

The optimal combinations of input meteorological factors were applied to the stations in the four different climatic regions of China mainland. At the same time, different methods (Models 1–6) were used to estimate the R_s of each station (Fig. 6). Compared with the Angstrom-Prescott formula and multiple linear regression method, the four machine learning methods could all effectively improve the estimation accuracy. Under the same combination of input meteorological factors, the estimation accuracies of different machine learning methods were also slightly different. Among the four different machine learning methods, the RMSE of ELM and SVM in each region were lower than other methods. Thus, it was recommended to use the ELM or SVM method to estimate global solar radiation in different climate regions of China mainland. In addition, among the four different climatic regions, the estimation accuracy had been improved most obviously through the machine learning methods in the SMZ region, where the mean of RMSE was reduced from 2.7615 (Angstrom-Prescott formula) to 2.2850 MJ m^{-2} d^{-1} (ELM). This confirmed that the machine learning method could perform better in the SMZ region. Among the four climate zones, the estimation error in TCZ was the smallest, and the error distribution was the most concentrated. The range of box plot of MPZ was the largest, which was probably because there were only seven stations in MPZ and the complexity of the Qinghai-Tibet Plateau led to large differences among the stations. The range of estimation error in SMZ was relatively large, which was probably because there were many observation stations in SMZ and the internal climate was complicated.

ELM was the model with the smallest error among the several methods investigated method. Under the optimal combination of meteorological factors, the median of RMSE were 2.50, 2.20, 1.95, and 2.11 MJ m^{-2} d^{-1} in MPZ, SMZ, TCZ and TMZ, respectively (Fig. 6). Considering the geographical location and climatic conditions comprehensively, eight representative stations in the four climatic regions were selected. The estimated R_s based on ELM and the optimal input

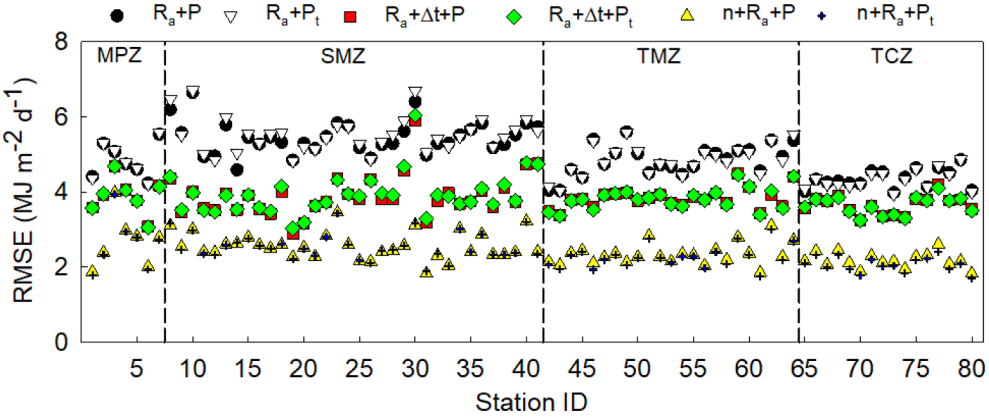

Fig. 4. Root mean square error (RMSE, MJ m^{-2} d^{-1}) of global solar radiation (R_s) estimation using the support vector machine (SVM) method under different combinations of input meteorological factors at 80 stations in China mainland.

Table 4
The determination coefficient (R^2), root mean square error (RMSE, MJ m^{-2} d^{-1}), akaike's information corrected criterion (AICc) of daily global solar radiation (R_s) estimations through support vector machine (SVM), multiple linear regression (MLR), and Angstrom-Prescott formula (A-P) under different combinations of input meteorological factors in different climatic regions of China mainland.

	Input combination	R^2			RMSE (MJ m^{-2} d^{-1})			AICc		
		SVM	MLR	A-P	SVM	MLR	A-P	SVM	MLR	A-P
MPZ	n	0.613	0.542		4.057	4.400		3.799	3.962	
	$n + R_a$	0.824	0.811	0.820	2.632	2.737	2.668	2.875	2.965	2.901
	$n + R_a + T_{max}$	0.826	0.811		2.615	2.735		2.862	2.964	
	$n + R_a + T_{max} + T_{mean}$	0.829	0.811		2.587	2.733		2.839	2.963	
	$n + R_a + T_{max} + T_{mean} + T_{min}$	0.833	0.813		2.558	2.716		2.816	2.950	
	$\boldsymbol{n + R_a + T_{max} + T_{mean} + T_{min} + RH}$	0.836	0.814		2.526	2.709		2.790	2.945	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t$	0.836	0.814		2.524	2.707		2.790	2.944	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t + U$	0.837	0.815		2.523	2.702		2.792	2.941	
SMZ	n	0.802	0.773		3.260	3.468		3.357	3.481	
	$n + R_a$	0.872	0.850	0.856	2.610	2.826	2.761	2.905	3.067	3.018
	$n + R_a + \Delta t$	0.883	0.854		2.485	2.783		2.806	3.036	
	$n + R_a + \Delta t + T_{max}$	0.887	0.858		2.445	2.751		2.773	3.014	
	$n + R_a + \Delta t + T_{max} + RH$	0.895	0.865		2.356	2.686		2.697	2.966	
	$n + R_a + \Delta t + T_{max} + RH + P_t$	0.897	0.867		2.332	2.661		2.675	2.947	
	$n + R_a + \Delta t + T_{max} + RH + P_t + T_{mean}$	0.898	0.868		2.314	2.649		2.660	2.939	
	$\boldsymbol{n + R_a + \Delta t + T_{max} + RH + P_t + T_{mean} + U}$	0.900	0.869		2.293	2.643		2.643	2.935	
TMZ	n	0.729	0.651		3.864	4.300		3.694	3.909	
	$n + R_a$	0.897	0.871	0.891	2.332	2.615	2.402	2.678	2.914	2.739
	$n + R_a + T_{max}$	0.903	0.872		2.266	2.608		2.622	2.910	
	$n + R_a + T_{max} + T_{mean}$	0.906	0.873		2.229	2.603		2.590	2.907	
	$n + R_a + T_{max} + T_{mean} + T_{min}$	0.909	0.877		2.187	2.553		2.553	2.870	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH$	0.913	0.879		2.148	2.534		2.519	2.856	
	$\boldsymbol{n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t}$	0.915	0.882		2.126	2.509		2.498	2.838	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t + U$	0.915	0.882		2.120	2.504		2.493	2.834	
TCZ	n	0.763	0.657		3.771	4.430		3.652	3.977	
	$n + R_a$	0.925	0.906	0.922	2.075	2.326	2.133	2.453	2.685	2.507
	$n + R_a + T_{max}$	0.927	0.907		2.048	2.318		2.427	2.679	
	$n + R_a + T_{max} + T_{mean}$	0.929	0.907		2.024	2.312		2.404	2.675	
	$n + R_a + T_{max} + T_{mean} + T_{min}$	0.931	0.913		1.985	2.242		2.366	2.613	
	$\boldsymbol{n + R_a + T_{max} + T_{mean} + T_{min} + RH}$	0.933	0.914		1.964	2.231		2.346	2.605	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t$	0.933	0.914		1.957	2.226		2.340	2.601	
	$n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t + U$	0.933	0.914		1.959	2.224		2.342	2.600	

Note: The bold italic symbols in the table were the optimal combinations of input meteorological factors in each climatic region of China mainland.

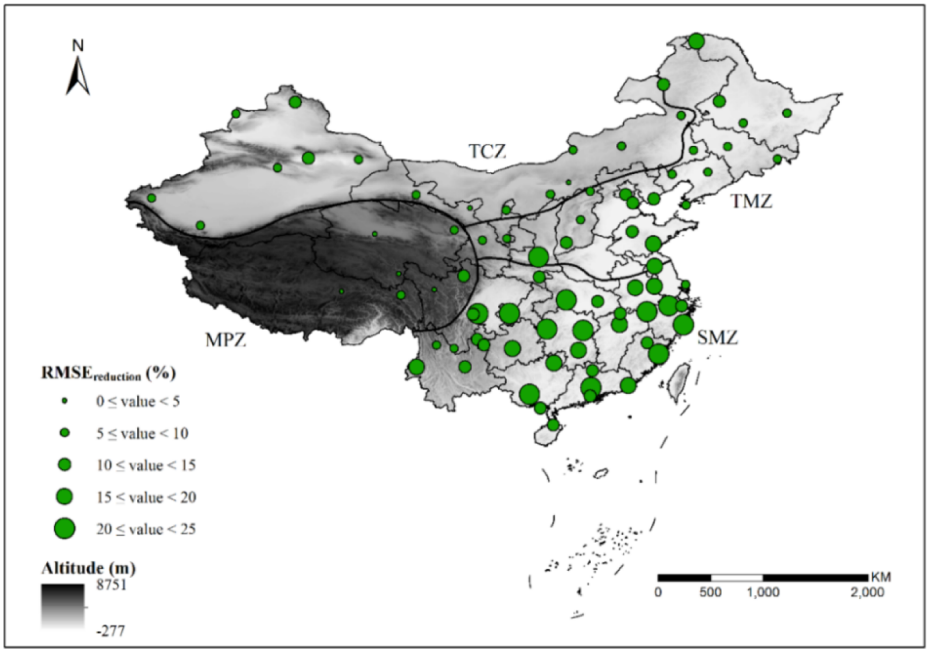

Fig. 5. $RMSE_{reduction}$ of the SVM method under the optimal combination of input meteorological factors relative to the Angstrom-Prescott formula in the four climatic regions in China mainland.

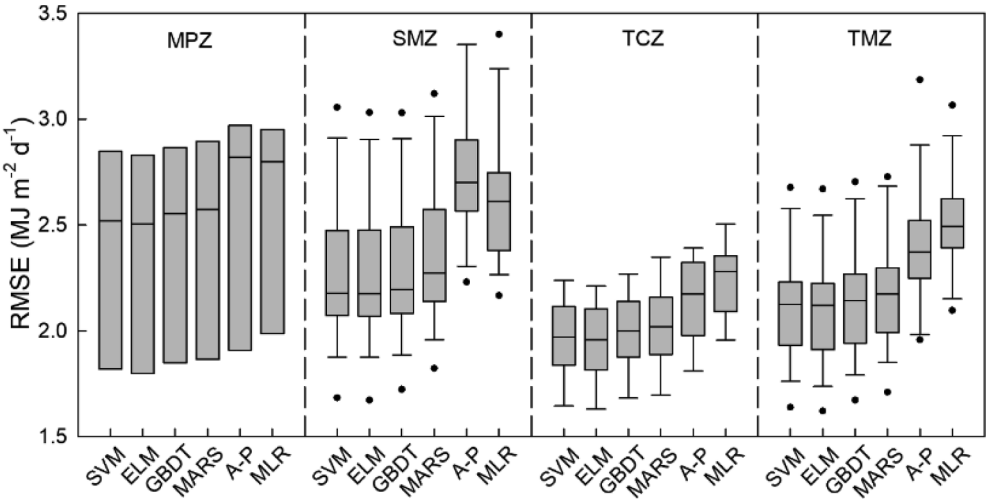

Fig. 6. RMSE (MJ m^{-2} d^{-1}) of estimation of daily global solar radiation through different machine learning methods under the same optimal combinations of input meteorological factors in the four different climatic regions of China mainland. The optimal combination of input meteorological factors were $n + R_a + T_{max} + T_{mean} + T_{min} + RH$, $n + R_a + \Delta t + T_{max} + RH + P_t + T_{mean} + U$, $n + R_a + T_{max} + T_{mean} + T_{min} + RH$ and $n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t$ in MPZ, SMZ, TCZ and TMZ, respectively.

combination were shown in Fig. 7. The eight typical stations had obtained good estimation results, especially the Turpan and Xilinhot stations in the TCZ, with R^2 of 0.955 and 0.942, respectively. R^2 of Golmud station in MPZ was 0.944, while R^2 of Changdu station in the same area was only 0.797, indicating that the internal situation of MPZ area was complicated and there was a big difference between the stations.

4. Discussion

It is very complicated to estimate the actual solar radiation energy on the earth's surface due to the atmosphere around the Earth, the undulating earth surface, and the unevenly distributed regions of land and sea. When solar radiation penetrates the atmosphere, it will be weakened by the atmosphere; when the solar radiation reaches the earth's surface, different reflections will occur due to the different ground properties. In addition, the earth's atmosphere is also constantly changing. It reflects, absorbs, and scatters the short-wave radiation of the sun, and absorbs the long-wave radiation of the ground. At the same time, the atmosphere itself also emits long-wave radiation. Therefore, global solar radiation varies not only with latitude and season, but also with geographical characteristics and atmospheric conditions. To estimate the solar radiation on the earth's surface, the first step was to investigate the solar radiation energy that reached the upper bound of the earth's atmosphere, i.e. the extraterrestrial radiation of R_a. The solar radiation energy at different times and locations is determined by the astronomical position of the sun to the earth. A day is divided into two parts, day and night. The duration of radiation available on the earth's surface is expressed by possible maximum sunshine duration, or N. Due to the existence of the atmosphere around the Earth, surface radiation and extra-terrestrial radiation are very different. Atmospheric transparency and sunny weather conditions can affect global solar radiation, which is mainly reflected by the meteorological factor of sunshine hours, or n. The solar radiation reaching the surface is absorbed by the atmosphere in the form of long-wave radiation or emitted into the air to cause temperature changes. The heat exchange causes weather changes. Precipitation, humidity, and wind speed will affect the particulate matter and cloud volume in the atmosphere. Therefore, the interaction of various meteorological factors affects solar radiation.

In this study, based on the analysis of correlations among meteorological factors in China mainland, it was found that R_s was strongly correlated with the meteorological factors of n, N, R_a, temperature and RH. A study by Benghanem and Mellit [56] in Saudi Arabia found that the correlation coefficient between n and R_s was 0.94, and the correlation coefficients between n with temperature and RH was 0.68 and −0.72. Their another study also showed that n played a very important role in estimating R_s [57]. These studies were consistent with the correlations between R_s and various meteorological factors found in this study.

In traditional empirical models of global solar radiation (e.g. the Angstrom-Prescott formula), it is necessary to input both N and R_a. However, this study showed that when using machine learning to estimate global solar radiation, R_a and N are interchangeable. In other words, using only one of these two factors could meet the estimation accuracy requirements. In addition, the study also found that in the process of global solar radiation estimation with machine learning method, the estimation accuracy could be further improved if the meteorological factor T_{mean} was added. Moreover, different forms of the same meteorological factor (e.g. Temperature and Precipitation) may also influence the final estimation accuracy. In previous studies, many scholars used T_{mean} as an input factor and achieved good estimation results [58]. Belaid and Mallet [59] used the SVM method to estimate R_s in Algeria based on different combinations of temperature-related meteorological factors (i.e. T_{max}, T_{min}, T_{mean}, Δt), R_a and N. It was found that the global solar radiation could be effectively estimated based on the temperature, and the introduction of R_a or N could greatly improve the estimation accuracy. This result was basically consistent with this study.

Studies have shown that rainfall could have a certain impact on the estimation of global solar radiation. Based on 39 stations in Australia, Liu and Scott [26] used different combinations of rainfall and temperature to estimate global solar radiation. They found that the accuracy of estimation could be improved by using P_t instead of P. Similar result was also obtained in China mainland in this study. Quej et al. [60] added rainfall factors to the temperature-based estimation model to improve the accuracy of global solar radiation estimation. Fan et al. [61] also found that rainfall information had a positive impact on estimating daily solar radiation in humid areas in China. However, precipitation had different effects on different climatic regions of China mainland in this study. For example, when the combination of $n + R_a + T_{max} + T_{mean} + T_{min} + RH$ was used to estimate global solar radiation in the MPZ and TCZ regions, accuracy improvement by additional precipitation information was very small, but the complexity of the model was increased. This may be due to the different climatic characteristics in different regions. The MPZ and TCZ belong to arid regions with less precipitation, where precipitation cannot sufficiently reflect the changes of global solar radiation. Therefore, local climatic conditions need to be considered carefully if rainfall was used as the input meteorological factor for the global solar radiation estimation.

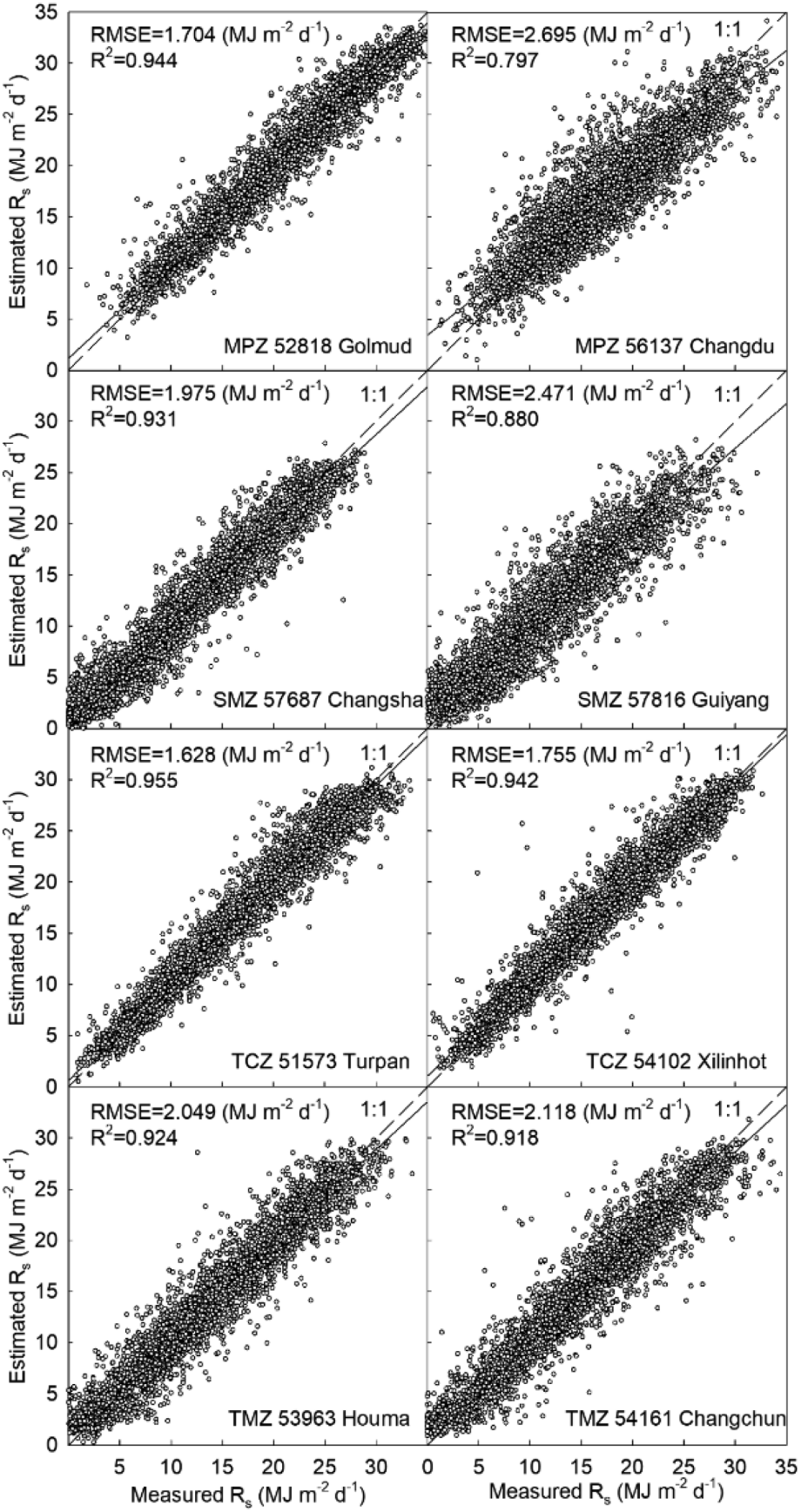

Fig. 7. RMSE (MJ m^{-2} d^{-1}) of estimation of daily global solar radiation through ELM method under the optimal combinations of input meteorological factors in eight typical stations (Golmud, Changdu, Changsha, Guiyang, Tuipan, Xilinhot, Houma and Changchun).

In previous studies, many scholars chose different combinations of input meteorological factors to estimate global solar radiation. For example, Wu et al. [62] estimated global solar radiation using n, N, R_a, T_{max}, T_{min}, RH, and P in humid regions of China. Torabi et al. [63] estimated the global solar radiation in Iran using n, T_{max}, T_{min}, R_a and N as input factors. Bhardwaj et al. [64] used Julian day, n, temperature, RH, wind speed, and atmospheric pressure to estimate global solar radiation in India. In addition, similar studies were also conducted in Korea [65], Malaysia [66], and Turkey [46]. However, the combinations of input meteorological factors were very different in the above studies, which indicated that it was necessary to select the most suitable combinations of input meteorological factors according to local climate characteristics in different regions, especially for a country with a large area and various climates. In this study, China mainland was divided into four different climate zones. The optimal combinations of input meteorological factors and the machine-learning estimation methods were explored for each of the four different climatic zones. However, it is still unclear whether there exits significant difference in the estimation accuracy within the same climatic zone. Thus, further study is needed to make sure that the division of the climatic zones matches the variation of global solar radiation in China mainland.

The Support Vector Machine (SVM) method is one of the most widely used machine learning methods. In previous studies, the SVM method has been recommended for global solar radiation estimation [59]. Therefore, to reduce the computational complexity, this study only used the SVM as the representative machine learning method to determine the optimal combination of input meteorological factors in the four different climatic regions of China mainland. It was assumed that the optimal combination of input meteorological factors selected for the SVM method was also valid for other machine learning methods.

This study found that machine learning methods could more accurately estimate global solar radiation in China mainland compared to traditional empirical models. Previous studies had similar results [67], mainly because machine learning methods were more capable of dealing with nonlinear and noisy data, such as extreme weather events (heavy rain and sand etc.) [12]. Many scholars also conducted comparative studies on different machine learning methods. For example, in the study by Shamshirband et al. [68], the performance of different machine learning methods ranked from good to bad as ELM > SVM > GP (genetic programming) > ANN (artificial neural network). Wang et al. [69] showed the ranking was: ANFIS (adaptive neuro-fuzzy inference systems) > M5Tree > Empirical Model. Keshtegar et al. [46] found that MARS was superior to RSM (response surface method), Kriging and M5Tree. Fan et al. [70] used n, N, and R_a as input meteorological factors and compared the performance of 12 different machine learning algorithms in China mainland. They found that ANFIS, ELM, SVM, and MARS generally performed better. However, according to the results of this study, the performance of different machine learning methods was related to regional climate conditions and the combinations of input meteorological factors.

This study clarified the differences in solar radiation estimation in different climatic regions of China mainland, and improved the estimation accuracy of radiation through the best combination of meteorological factors and machine learning algorithms. The results could help China to better assess and utilize solar energy resources. It also provided a new way for the solar radiation estimation in regions with complex climate. In this study, only six representative models were selected to estimate global solar radiation in China mainland. There are still many other machine learning methods that were not tested. Therefore, further study is needed to see whether there are other better machine learning methods for global solar radiation estimation. In the application of machine learning method, only the key parameters of the model itself were optimized in this study, while the other parameters were set as default values. This procedure simplified the process of model correction to a certain extent, but further study is also needed to see whether the solar radiation estimation accuracy could be further improved through optimizing the remaining model parameters. Furthermore, some uncommon meteorological factors such as air pressure and visibility were not considered in this study. The optimal combination of input meteorological factors was determined when all common meteorological data were available. However, further study may be needed to determine the alternative combination of input meteorological factors when sunshine hours, temperatures, or other influential meteorological factors are in scarcity or missing.

5. Conclusions

In this study, using the meteorological data of 80 stations in China mainland, based on the correlation of meteorological elements, different input combinations of meteorological factor were established to evaluate the performance of six different estimation models in the four climatic regions of China mainland. Some main conclusions were drawn as follows.

In the process of estimating global solar radiation R_s with machine learning method, the addition of n had the greatest impact on the estimation accuracy. R_a could be used to replace N to reduce the number of input meteorological factors. For temperature-related meteorological factors, a combination of T_{max} + T_{min} + T_{mean} was recommended. If relative humidity data were available, it was also suggested as an input meteorological factor. A binary rainfall event P_t (rainfall = 1 and no rainfall = 0) was recommended as an input factor in humid regions, but not in arid regions. Wind speed contributed little to the improvement of estimation accuracy. Geographical location and climatic conditions should be taken into account to select the optimal combinations of input meteorological factors for R_s estimation. In MPZ, the combination of $n + R_a + T_{max} + T_{mean} + T_{min} + RH$ was recommended; in SMZ, the combination of $n + R_a + \Delta t + T_{max} + RH + P_t + T_{mean} + U$; in TMZ, the combination of $n + R_a + T_{max} + T_{mean} + T_{min} + RH + P_t$; and in TCZ, the combination of $n + R_a + T_{max} + T_{mean} + T_{min} + RH$. No matter under what kind of combination of input meteorological factors, the accuracy of machine learning method was higher than the multiple linear regression method and Angstrom-Prescott empirical formula. Especially in the SMZ, the advantages of machine learning methods were particularly obvious. The ELM and SVM methods were recommended for global solar radiation estimation in China mainland. For regions with large areas or complex climatic conditions, there are certain differences in regional radiation estimates. The accuracy of estimation could be effectively improved through dividing China into different regions and studying each region separately. This study provided a reference for the selection of appropriate combination of input meteorological factors and the methods for the estimation of global solar radiation in different climatic regions of China mainland. This method can also be applied to other similar climate regions in the world for further research.

CRediT authorship contribution statement

Chuan He: Conceptualization, Methodology, Software, Writing - original draft, Visualization. **Jiandong Liu:** Resources, Data curation. **Fang Xu:** Validation, Investigation. **Teng Zhang:** Validation, Investigation. **Shang Chen:** Validation, Writing - review & editing. **Zhe Sun:** Validation, Investigation. **Wenhui Zheng:** Validation, Investigation. **Runhong Wang:** Validation, Investigation. **Liang He:** Writing - review & editing. **Hao Feng:** Supervision, Project administration, Funding acquisition. **Qiang Yu:** Supervision, Project administration, Funding acquisition. **Jianqiang He:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

This research was supported by the Natural Science Foundation of China (No. 41961124006, 41730645), the Key Research and Development Program of Shaanxi (No. 2019ZDLNY07-03), the "Open Project Fund" from the State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Chinese Academy of Sciences and Ministry of Water Resources (No. A314021402-1611), the Science Promotion Project of Test and Demonstration Stations in the Norwest A&F University (No. TGZX2018-32), and the "111 Project" (No. B12007) of China.

Appendix A. Basic information about the 80 meteorological stations involved in this study in China mainland. The precipitation in the table is annual mean value, while the other meteorological factors are daily mean values. The acronym of MPZ represents the mountain plateau zone, SMZ the subtropical monsoon zone, TMZ the temperate monsoon zone, and TCZ the temperate continental zone, respectively.

Station No.	Code	Station name	Climatic Region	Latitude (N)	Longitude (E)	Altitude (*m*)	R_s (MJ m^{-2} d^{-1})	*n* (h)	T_{max} (°C)	T_{mean} (°C)	T_{min} (°C)	*RH* (%)	*P* (mm y^{-1})	*U* (m s^{-1})	Record period
1	52,818	Golmud	MPZ	36.4	94.9	2808.4	19.1	8.4	13.0	5.3	−1.4	32.4	43.0	2.6	1957–2017
2	52,866	Xining	MPZ	36.7	101.8	2296.0	15.9	7.3	14.0	6.0	0.0	56.2	385.5	1.5	1959–2017
3	55,299	Naqu	MPZ	31.5	92.1	4508.2	17.6	7.6	7.1	−0.9	−7.6	51.9	444.0	2.6	1961–2017
4	56,029	Yushu	MPZ	33.0	97.0	3717.7	16.6	6.8	12.0	3.5	−3.0	53.7	486.3	1.1	1960–2017
5	56,137	Changdu	MPZ	31.2	97.2	3316.2	16.8	6.5	16.8	7.8	1.0	50.3	477.6	1.1	1957–2017
6	56,146	Ganzi	MPZ	31.6	100.0	3394.4	18.1	6.9	14.8	6.3	0.2	55.8	659.1	1.8	1994–2017
7	56,173	Hongyuan	MPZ	32.8	102.6	3492.8	16.8	6.4	10.8	2.1	−4.6	69.4	729.7	2.3	1994–2017
8	56,385	Emeishan	SMZ	29.5	103.3	3048.3	12.7	3.9	7.7	3.3	0.5	85.7	1763.2	3.0	1959–2017
9	56,386	Leshan	SMZ	29.6	103.8	425.2	9.5	2.9	20.9	17.1	14.3	80.5	1323.1	1.3	1973–1990
10	56,586	Zhaotong	SMZ	27.4	103.7	1950.7	14.3	5.2	18.2	11.6	7.4	74.6	723.5	2.5	1961–1990
11	56,651	Lijiang	SMZ	26.9	100.2	2382.1	17.0	6.7	19.5	12.9	8.0	62.5	964.0	3.1	1961–2017
12	56,666	Panzhihua	SMZ	26.6	101.7	1225.9	16.1	7.4	27.8	20.9	15.7	56.9	816.4	1.4	1992–2017
13	56,691	Weining	SMZ	26.9	104.3	2238.6	13.1	4.9	16.3	10.4	6.8	79.8	927.0	3.2	1961–1990
14	56,739	Tengchong	SMZ	25.0	98.5	1696.9	15.2	5.9	21.6	15.2	10.6	77.4	1481.0	1.6	1957–2017
15	56,778	Kunming	SMZ	25.0	102.7	1889.1	15.1	6.2	21.1	15.2	10.6	71.4	987.1	2.1	1959–2017
16	57,245	Ankang	SMZ	32.7	109.0	291.7	11.4	4.6	21.4	16.0	12.2	73.8	802.5	1.4	1990–2017
17	57,461	Yichang	SMZ	30.7	111.4	257.5	10.9	4.2	21.6	17.0	13.6	75.3	1144.8	1.3	1957–2017
18	57,494	Wuhan	SMZ	30.6	114.1	24.4	12.3	5.3	21.4	16.8	13.2	76.9	1262.4	2.0	1957–2017
19	57,516	Shapingba	SMZ	29.6	106.5	259.6	8.6	2.7	22.4	18.6	15.9	78.4	1103.5	1.4	1987–2017
20	57,649	Jishou	SMZ	28.2	109.7	255.6	9.6	3.4	21.8	17.0	13.8	79.1	1404.6	1.2	1992–2017
21	57,687	Changsha	SMZ	28.1	112.8	120.0	10.7	4.2	21.8	17.6	14.6	77.9	1458.7	2.2	1987–2017
22	57,816	Guiyang	SMZ	26.6	106.7	1224.9	10.3	3.2	19.6	15.1	12.1	77.4	1091.7	2.3	1959–2017
23	57,874	Changning	SMZ	26.4	112.4	117.8	11.0	3.7	22.8	18.5	15.4	77.4	1421.3	1.9	1992–2017
24	57,957	Guilin	SMZ	25.3	110.3	165.6	11.5	4.1	23.4	19.0	16.0	74.9	1870.8	2.4	1957–2017
25	58,238	Nanjing	SMZ	31.9	118.9	36.4	12.6	5.5	20.5	15.7	11.9	75.1	1074.0	2.5	1959–2017
26	58,265	Lvsi	SMZ	32.1	121.6	6.5	13.2	6.0	19.4	15.8	12.9	78.2	1100.7	3.4	1992–2017
27	58,321	Hefei	SMZ	31.8	117.3	28.2	12.3	5.3	20.6	16.1	12.4	75.3	1004.6	2.6	1959–2017
28	58,457	Hangzhou	SMZ	30.2	120.2	42.6	11.8	4.8	21.2	16.8	13.4	76.5	1414.9	2.2	1959–2017
29	58,467	Cixi	SMZ	30.2	121.3	5.7	12.7	5.6	20.4	16.2	12.9	81.0	1259.0	2.8	1961–1990
30	58,506	Lushan	SMZ	29.6	116.0	1165.4	13.2	5.0	15.3	11.6	8.8	78.0	1953.3	5.0	1960–1990
31	58,531	Tunxi	SMZ	29.7	118.3	143.9	12.1	4.6	22.4	17.0	13.1	78.1	1806.5	1.3	1992–2017
32	58,606	Nanchang	SMZ	28.6	115.9	47.9	12.4	5.1	21.8	17.9	14.9	76.0	1585.9	2.5	1959–2017
33	58,665	Hongjia	SMZ	28.6	121.4	5.3	12.7	4.6	22.2	18.0	15.0	77.3	1590.7	2.3	1992–2017
34	58,737	Jian'ou	SMZ	27.1	118.3	155.7	13.5	4.6	25.0	19.2	15.4	80.1	1742.2	1.4	1992–2017
35	58,847	Fuzhou	SMZ	26.1	119.3	84.8	12.2	4.5	24.6	20.0	17.0	75.3	1389.5	2.6	1959–2017
36	59,082	Shaoguan	SMZ	24.7	113.6	122.3	12.3	4.9	25.4	20.5	17.0	75.7	1500.0	1.4	1960–1990
37	59,287	Guangzhou	SMZ	23.2	113.5	71.5	11.9	4.6	26.5	22.1	19.0	77.0	1781.5	1.9	1957–2017
38	59,316	Shantou	SMZ	23.4	116.7	3.9	14.0	5.6	25.5	21.8	19.0	79.6	1568.4	2.5	1957–2017
39	59,431	Nanning	SMZ	22.6	108.2	122.6	12.6	4.4	26.4	21.7	18.5	79.1	1297.3	1.5	1961–2017
40	59,485	Zhongshan	SMZ	22.5	113.4	34.5	12.1	4.9	25.7	21.8	18.9	82.9	1801.8	2.1	1965–1990
41	59,644	Beihai	SMZ	21.5	109.1	14.0	14.2	5.1	26.8	23.1	20.4	79.8	1828.2	3.3	1993–2017
42	50,136	Mohe	TMZ	53.0	122.5	439.7	12.2	6.6	4.8	−4.2	−12.1	68.6	443.3	1.8	1993–2017
43	50,742	Fuyu	TMZ	47.8	124.5	163.8	14.1	7.0	9.2	3.4	−1.9	62.9	433.9	3.1	1993–2017
44	50,873	Jiamusi	TMZ	46.8	130.3	83.1	12.4	6.6	9.4	3.5	−2.1	66.5	534.4	3.1	1961–2017
45	50,953	Harbin	TMZ	45.9	126.6	117.7	12.9	6.7	10.1	4.4	−1.0	65.1	516.3	3.3	1959–2017
46	52,983	Yuzhong	TMZ	35.9	104.2	1875.6	15.3	7.0	14.6	7.4	1.6	62.1	359.0	2.1	2005–2017
47	53,487	Datong	TMZ	40.1	113.4	1053.6	15.4	7.4	14.1	7.0	0.8	52.3	374.6	2.8	1960–2017
48	53,772	Taiyuan	TMZ	37.6	112.6	777.3	14.5	6.9	17.1	10.2	4.2	58.4	442.9	2.1	1959–2017
49	53,817	Guyuan	TMZ	36.0	106.3	1754.2	15.2	7.1	13.7	7.3	1.9	60.5	432.9	2.6	1985–2017
50	53,963	Houma	TMZ	35.7	111.4	435.1	13.5	6.1	19.6	12.9	7.3	64.5	506.0	1.9	1959–2017
51	54,135	Tongliao	TMZ	43.6	122.3	179.7	14.0	8.2	13.1	6.8	1.1	54.3	364.5	3.6	1960–2017
52	54,161	Changchun	TMZ	43.9	125.2	237.5	13.6	7.1	11.3	5.7	0.7	62.9	584.6	3.7	1959–2017
53	54,292	Yanji	TMZ	42.9	129.5	258.5	13.0	6.3	12.1	5.4	−0.2	64.6	523.5	2.6	1960–2017
54	54,324	Chaoyang	TMZ	41.6	120.4	175.3	14.2	7.4	16.1	9.2	2.9	51.7	467.7	2.8	1963–2017
55	54,342	Shenyang	TMZ	41.7	123.5	49.5	13.5	6.8	14.1	8.3	3.1	63.5	688.7	2.9	1957–2017
56	54,511	Beijing	TMZ	39.8	116.5	32.5	14.4	7.2	18.1	12.5	7.4	56.1	569.0	2.4	1957–2017
57	54,527	Tianjin	TMZ	39.1	117.1	4.3	14.0	6.8	18.1	12.7	8.3	61.2	531.4	2.6	1959–2017
58	54,539	Leting	TMZ	39.4	118.9	9.7	14.0	6.7	16.9	11.5	7.2	63.9	539.7	2.3	1992–2017
59	54,662	Dalian	TMZ	38.9	121.6	92.5	13.7	7.4	14.7	11.1	8.1	64.7	615.5	4.4	1963–2017

60	54,823	Jinan	TMZ	36.6	117.0	171.2	13.5	6.8	19.6	14.7	10.4	57.0	692.1	3.0	1959–2017
61	54,936	Juxian	TMZ	35.6	118.8	108.4	13.8	5.9	18.7	12.9	8.2	70.8	777.6	2.4	1990–2017
62	57,131	Jinghe	TMZ	34.4	109.0	411.0	12.6	5.2	19.9	14.7	10.6	62.6	521.1	2.5	2006–2017
63	58,141	Huaian	TMZ	33.6	118.9	13.7	13.1	5.3	19.7	14.9	11.0	72.3	980.1	2.4	2001–2017
64	59,758	Haikou	TMZ	20.0	110.3	64.7	14.0	5.7	28.1	24.2	21.5	83.3	1706.5	2.8	1957–2017
65	50,527	Hailar	TCZ	49.3	119.7	650.4	13.9	7.3	5.6	−0.8	−6.6	66.3	353.5	3.2	1960–2017
66	50,834	Suolun	TCZ	46.6	121.2	501.0	14.7	7.7	10.6	3.0	−3.5	56.8	447.3	2.8	1992–2017
67	51,076	Altay	TCZ	47.7	88.1	736.5	15.2	8.2	10.9	4.6	−1.2	58.1	197.8	2.3	1960–2017
68	51,133	Tacheng	TCZ	46.7	83.0	536.0	15.2	7.9	14.8	8.0	2.3	57.4	304.8	2.2	1993–2017
69	51,567	Yandie	TCZ	42.1	86.6	1056.5	15.4	8.1	16.6	9.1	2.4	58.2	79.9	1.6	1993–2017
70	51,573	Turpan	TCZ	42.9	89.2	35.2	15.4	7.9	21.8	14.8	8.6	39.6	15.0	1.2	1960–2017
71	51,709	Kashi	TCZ	39.5	75.8	1386.7	15.7	7.7	18.4	12.1	6.0	50.1	70.4	1.8	1957–2017
72	51,828	Hetian	TCZ	37.1	79.9	1376.0	16.2	7.2	19.3	12.9	7.3	41.2	39.1	1.9	1957–2017
73	52,203	Hami	TCZ	42.8	93.5	738.3	17.1	9.2	18.1	10.1	3.1	43.4	39.8	1.8	1961–2017
74	52,533	Jiuquan	TCZ	39.8	98.5	1478.4	16.6	8.4	15.4	8.1	1.6	46.9	90.7	2.1	1993–2017
75	52,681	Minqin	TCZ	38.6	103.1	1368.7	16.6	8.5	16.3	8.6	1.7	44.4	114.1	2.7	1957–2017
76	53,068	Erenhot	TCZ	43.6	111.9	964.1	17.3	8.7	12.0	4.4	−2.2	47.2	135.2	4.0	1957–2017
77	53,463	Hohhot	TCZ	40.9	111.6	1154.4	16.5	8.3	12.2	5.1	−1.1	54.5	347.1	1.7	1959–1968
78	53,543	Dongsheng	TCZ	39.8	110.0	1463.1	16.1	8.3	12.7	7.2	2.6	47.5	380.3	2.8	1992–2017
79	53,614	Yinchuan	TCZ	38.5	106.2	1111.6	16.4	7.9	16.3	9.5	3.6	55.2	198.7	2.0	1959–2017
80	54,102	Xilinhot	TCZ	44.0	116.1	1003.8	15.4	8.1	10.3	3.2	−3.0	55.7	277.6	3.3	1990–2017

References

[1] Almorox J, Bocco M, Willington E. Estimation of daily global solar radiation from measured temperatures at Cañada de Luque, Córdoba, Argentina. Renew Energy 2013;60:382–7.
[2] Liu X, Mei X, Li Y, Wang Q, Zhang Y, Porter JR. Variation in reference crop evapotranspiration caused by the Ångström-Prescott coefficient: Locally calibrated versus the FAO recommended. Agric Water Manage 2009;96:1137–45.
[3] Gueymard CA. The sun's total and spectral irradiance for solar energy applications and solar radiation models. Sol Energy 2004;76:423–53.
[4] Irvine PJ, Schäfer S, Lawrence MG. Solar radiation management could be a game changer. Nat. Clim Change 2014;4:842-.
[5] Wang J, Wang E, Yin H, Feng L, Zhao Y. Differences between observed and calculated solar radiations and their impact on simulated crop yields. Field Crop Res 2015;176:1–10.
[6] Thornton PE, Running SW. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agric Forest Meteorol 1999;93:211–28.
[7] Pan T, Wu S, Dai E, Liu Y. Estimating the daily global solar radiation spatial distribution from diurnal temperature ranges over the Tibetan Plateau in China. Appl Energy 2013;107:384–93.
[8] Rashid K, Mohammadi K, Powell K. Dynamic simulation and techno-economic analysis of a concentrated solar power (CSP) plant hybridized with both thermal energy storage and natural gas. J Cleaner Prod 2020;248.
[9] He G, Kammen DM. Where, when and how much solar is available? A provincial-scale solar resource assessment for China. Renew Energy 2016;85:74–82.
[10] Gosens J, Kåberger T, Wang Y. China's next renewable energy revolution: goals and mechanisms in the 13th Five Year Plan for energy. Energy Sci Eng 2017;5:141–55.
[11] Bayrakçı HC, Demircan C, Keçebaş A. The development of empirical models for estimating global solar radiation on horizontal surface: a case study. Renewable Sustainable Energy Rev 2018;81:2771–82.
[12] Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: a review. Renew Energy 2017;105:569–82.
[13] Pillot B, Muselli M, Poggi P, Dias JB. Satellite-based assessment and in situ validation of solar irradiation maps in the Republic of Djibouti. Sol Energy 2015;120:603–19.
[14] Rashid K, Sheha MN, Powell KM. Real-time optimization of a solar-natural gas hybrid power plant to enhance solar power utilization. Annual American Control Conference (ACC) 2018:3002–7.
[15] Besharat F, Dehghan AA, Faghih AR. Empirical models for estimating global solar radiation: a review and case study. Renewable Sustainable Energy Rev 2013;21:798–821.
[16] Zhang J, Zhao L, Deng S, Xu W, Zhang Y. A critical review of the models used to estimate solar radiation. Renewable Sustainable Energy Rev 2017;70:314–29.
[17] Elagib NA, Mansell MG. New approaches for estimating global solar radiation across Sudan. Energy Convers Manage 2000;41:419–34.
[18] Almorox J, Hontoria C. Global solar radiation estimation using sunshine duration in Spain. Energy Convers Manage 2004;45:1529–35.
[19] Kisi O, Parmar KS. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 2016;534:104–12.
[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[21] Tymvios FS, Jacovides CP, Michaelides SC, Scouteli C. Comparative study of Ångström's and artificial neural networks' methodologies in estimating global solar radiation. Sol Energy 2005;78:752–62.
[22] Chen J-L, Liu H-B, Wu W, Xie D-T. Estimation of monthly solar radiation from measured temperatures using support vector machines – a case study. Renew Energy 2011;36:413–20.
[23] Chen JL, Li GS. Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables. Theor Appl Climatol 2014;115:627–38.
[24] Fan J, Wang X, Wu L, Zhang F, Bai H, Lu X, et al. New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: a case study in South China. Energy Convers Manage 2018;156:618–25.
[25] Meenal R, Selvakumar AI. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renew Energy 2018;121:324–43.
[26] Liu DL, Scott BJ. Estimation of solar radiation in Australia from rainfall and temperature observations. Agric Forest Meteorol 2001;106:41–59.
[27] Yadav AK, Chandel SS. Solar radiation prediction using Artificial Neural Network techniques: a review. Renewable Sustainable Energy Rev 2014;33:772–81.
[28] Alizamir M, Kim S, Kisi O, Zounemat-Kermani M. A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. Energy 2020;197:117239.
[29] Ming Z, Shaojie O, Hui S, Yujian G. Is the "Sun" still hot in China? The study of the present situation, problems and trends of the photovoltaic industry in China. Renewable Sustainable Energy Rev 2015;43:1224–37.
[30] Liu J, Liu J, Linderholm HW, Chen D, Yu Q, Wu D, et al. Observation and calculation of the solar radiation on the Tibetan Plateau. Energy Convers Manage 2012;57:23–32.
[31] Tang W, Yang K, He J, Qin J. Quality control and estimation of global solar radiation in China. Sol Energy 2010;84:466–75.
[32] Teke A, Yıldırım HB, Çelik Ö. Evaluation and performance comparison of different models for the estimation of solar radiation. Renewable Sustainable Energy Rev 2015;50:1097–107.
[33] Song Y, Achberger C, Linderholm HW. Rain-season trends in precipitation and their effect in different climate regions of China during 1961–2008. Environ Res Lett 2011;6.
[34] Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome, 1998. p. D05109.
[35] Angstrom A. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Q J Roy Meteor Soc 1924;50:121–6.
[36] Prescott JA, Prescott J, Prescott JS. Evaporation from a water surface in relation to solar radiation. T Roy Soc South Aust 1940;46:114–8.
[37] Vapnik V. The nature of statistical learning theory. New York: Springer science & business media; 2013.
[38] Chen JL, Li GS, Wu SJ. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Convers Manage 2013;75:311–8.
[39] Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. Journal of statistical software2004. pp. 1-20.
[40] Friedman JH. Stochastic gradient boosting. Comput Stat Data An 2002;38:367–78.
[41] Ma X, Ding C, Luan S, Wang Y, Wang Y. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans Intell Transp Sys 2017;18:2303–10.
[42] Persson C, Bacher P, Shiga T, Madsen H. Multi-site solar power forecasting using gradient boosted regression trees. Sol Energy 2017;150:423–36.
[43] Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. R Package Version 2.1 2018.
[44] Friedman JHJT. Multivariate adaptive regression splines. Ann Statist 1991;19:1–67.
[45] Kisi O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 2015;528:312–20.
[46] Keshtegar B, Mert C, Kisi O. Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree. Renewable Sustainable Energy Rev 2018;81:330–41.
[47] Milborrow S. Derived from mda: mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth:

Multivariate Adaptive Regression Splines. R package version 4.4.4 2019.
[48] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70:489–501.
[49] Salcedo-Sanz S, Casanova-Mateo C, Pastor-Sánchez A, Sánchez-Girón M. Daily global solar radiation prediction based on a hybrid coral reefs optimization – extreme learning machine approach. Sol Energy 2014;105:91–8.
[50] Mouselimis L, Gosso A. elmNNRcpp: The Extreme Learning Machine Algorithm. R package version 1.0.1 2018.
[51] Coulibaly O, Ouedraogo A. Correlation of global solar radiation of eight synoptic stations in burkina faso based on linear and multiple linear regression methods. J Sol Energy 2016;2016:1–9.
[52] R Core Team. R: A language and environment for statistical computing. 2013.
[53] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1975;19:716–23.
[54] Wagenmakers E-J, Farrell S. AIC model selection using Akaike weights. Psychon B Rev 2004;11:192–6.
[55] McQuarrie ADR, Tsai C-L. Regression and time series model selection. World Scientific 1998.
[56] Benghanem M, Mellit A. Radial Basis Function Network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy 2010;35:3751–62.
[57] Benghanem M, Mellit A, Alamri SN. ANN-based modelling and estimation of daily global solar radiation data: a case study. Energy Convers Manage 2009;50:1644–55.
[58] Rehman S, Mohandes M. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy 2008;36:571–6.
[59] Belaid S, Mellit A. Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Convers Manage 2016;118:105–18.
[60] Quej VH, Almorox J, Arnaldo JA, Saito L. ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. J Atmos Sol-terr Phys 2017;155:62–70.
[61] Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, et al. 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 Manage 2018;164:102–11.
[62] Wu L, Huang G, Fan J, Zhang F, Wang X, Zeng W. Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions. Energy Convers Manage 2019;183:280–95.
[63] Torabi M, Mosavi A, Ozturk P, Varkonyi-Koczy A, Istvan V. A hybrid machine learning approach for daily prediction of solar radiation. In: Laukaitis G, editor. Recent Advances in Technology Research and Education. Cham: Springer International Publishing; 2019. p. 266–74.
[64] Bhardwaj S, Sharma V, Srivastava S, Sastry OS, Bandyopadhyay B, Chandel SS, et al. Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model. Sol Energy 2013;93:43–54.
[65] Lee M, Koo C, Hong T, Park HS. Framework for the mapping of the monthly average daily solar radiation using an advanced case-based reasoning and a geostatistical technique. Environ Sci Technol 2014;48:4604–12.
[66] Ibrahim IA, Khatib T. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers Manage 2017;138:413–25.
[67] Hassan MA, Khalil A, Kaseb S, Kassem MA. Potential of four different machine-learning algorithms in modeling daily global solar radiation. Renew Energy 2017;111:52–62.
[68] Shamshirband S, Mohammadi K, Yee PL, Petković D, Mostafaeipour A. A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation. Renewable Sustainable Energy Rev 2015;52:1031–42.
[69] Wang L, Kisi O, Zounemat-Kermani M, Zhu Z, Gong W, Niu Z, et al. Prediction of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree. Int J Climatol 2017;37:1141–55.
[70] Fan J, Wu L, Zhang F, Cai H, Zeng W, Wang X, et al. Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. Renewable Sustainable Energy Rev 2019;100:186–212.