Earth and Space Science

RESEARCH ARTICLE
10.1029/2019EA001058

Daily Global Solar Radiation in China Estimated From High-Density Meteorological Observations: A Random Forest Model Framework

Zhaoliang Zeng[1], Zemin Wang[1], Ke Gui[2], Xiaoyu Yan[3], Meng Gao[4], Ming Luo[5,6], Hong Geng[7], Tingting Liao[8], Xiao Li[9], Jiachun An[1], Haizhi Liu[10], Chao He[7], Guicai Ning[6], and Yuanjian Yang[11,6,12]

[1]Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan, China, [2]Institute of Atmospheric Composition, Chinese Academy of Meteorological Sciences, CMA, Beijing, China, [3]Environment and Sustainability Institute, University of Exeter, Penryn, UK, [4]Department of Geography, Hong Kong Baptist University, Hong Kong, [5]School of Geography and Planning, and Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, China, [6]Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong, [7]School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, [8]Plateau Atmospheric and Environment Laboratory of Sichuan Province, College of Atmospheric Science, Chengdu University of Information Technology, Chengdu, China, [9]CPI Power Engineering Co., LTD, Shanghai, China, [10]National Meteorological Center, CMA, Beijing, China, [11]School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing, China, [12]State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China

Abstract Accurate estimation of the spatiotemporal variations of solar radiation is crucial for assessing and utilizing solar energy, one of the fastest-growing and most important clean and renewable resources. Based on observations from 2,379 meteorological stations along with scare solar radiation observations, the random forest (RF) model is employed to construct a high-density network of daily global solar radiation (DGSR) and its spatiotemporal variations in China. The RF-estimated DGSR is in good agreement with site observations across China, with an overall correlation coefficient (R) of 0.95, root-mean-square error of 2.34 MJ/m^2, and mean bias of −0.04 MJ/m^2. The geographical distributions of R values, root-mean-square error, and mean bias values indicate that the RF model has high predictive performance in estimating DGSR under different climatic and geographic conditions across China. The RF model further reveals that daily sunshine duration, daily maximum land surface temperature, and day of year play dominant roles in determining DGSR across China. In addition, compared with other models, the RF model exhibits a more accurate estimation performance for DGSR. Using the RF model framework at the national scale allows the establishment of a high-resolution DGSR network, which can not only be used to effectively evaluate the long-term change in solar radiation but also serve as a potential resource to rationally and continually utilize solar energy.

Key Points:
- The RF model exhibits high accuracy and considerable potential in projecting DGSR
- Importance of possible variables for estimating DGSR is identified by the RF model
- Spatiotemporal variations of DGSR across China are constructed from high-density meteorological observations

Correspondence to:
Z. Wang and Y. Yang,
zmwang@whu.edu.cn;
yyj1985@nuist.edu.cn

Citation:
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 and Space Science*, *7*, e2019EA001058. https://doi.org/10.1029/2019EA001058

Received 13 DEC 2019
Accepted 16 JAN 2020
Accepted article online 26 JAN 2020

© 2020 The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Introduction

Daily global solar radiation (DGSR), as the major energy source of the Earth, plays a key role in the terrestrial radiation balance, energy exchange, hydrological cycle, photosynthesis, and the formation of weather and climate (Cooter & Dhakhwa, 1996; Cline et al., 1998; Hoogenboom, 2000; Power, 2001; Pohlert, 2004; Wild, 2009). DGSR is also crucial for the utilization of solar energy through transformation of technologies (Wu et al., 2012, 2016; Tang et al., 2018; Prăvălie et al., 2019; Zou et al., 2019). In recent years, there has been an aggravation of air pollution in China induced by massive fossil fuel consumption and emissions coinciding with unfavorable weather conditions (Guo et al., 2016, 2019; Lou et al., 2019; Yang, et al., 2018; Yang, Ye, et al., 2019; Zheng et al., 2019). On the one hand, this has significantly modulated the change in surface solar radiation (Che et al., 2005; Guo et al., 2018; Wang et al., 2012, 2013; Wang & Wild, 2016; Zheng et al., 2018; He & Wang, 2020; Yang et al., 2020), while on the other hand, it has led to solar radiation becoming one of the fastest-growing and important sources of clean and renewable energy. Therefore, solar radiation is a topic that has attracted broad and increasing attention in China (Che et al., 2005; Sun et al., 2016; Li et al., 2017; Wang et al., 2016; Song et al., 2019; Tang et al., 2016, 2018; Liu et al., 2019; He & Wang et al.,

2020). Meanwhile, DGSR data are urgently needed for designing and evaluating solar energy technologies, not only in China but also in many other regions of the world (Yang et al., 2001, 2006; Wang et al., 2015; Zou et al., 2019; Prăvălie et al., 2019).

However, to date, field observations have been insufficient and solar radiation monitoring stations are distributed unevenly and sparsely across the globe, meaning high-spatial-resolution and high-density DGSR information still remains limited. Besides, high uncertainty also exists in the estimation of solar radiation received at the Earth's surface, mainly due to the influences of astronomical, meteorological, and regional factors (Che et al., 2005; Sun et al., 2015; Wang et al., 2013, 2014; Wang & Wild, 2016; Liu et al., 2019; He & Wang, 2020; Yang et al., 2020). Therefore, a reliable and high-density DGSR observational network and information on the spatiotemporal distribution of surface solar radiation are an ongoing core issue in energy studies (Jiang, 2009b; Wang et al., 2016; Zhang et al., 2017; Assouline et al., 2018; Halabi et al., 2018, Qin et al., 2019). In China, only 97 solar radiation monitoring stations have been established on the mainland, and they are unevenly and sparsely distributed across the region (Tang et al., 2011). In addition, advances in solar radiation instrumentation and the relocation of stations have resulted in a lack of long-term and continuous solar radiation data sets in most parts of China. This limitation seriously hinders the evaluation of solar energy resources in China, leading to gaps in scientific research and engineering applications (Yang et al., 2001; Tang et al., 2018).

To obtain a longer and more detailed data set of DGSR in China, previous studies have made great efforts to estimate DGSR from surface meteorological observations using various methods, such as the traditional empirical formula and artificial neural network methods (Jiang et al., 2009a; Huang et al., 2011; Qin et al., 2011; Chen et al., 2013; Li et al., 2013; Tang et al., 2010, 2013, 2016; Sun et al., 2016; Li et al., 2017). Using these methods, however, the importance of input variables for estimating DGSR could not be identified objectively and automatically. In addition, although these empirical models might be simply established, their input variables are scare, and their regression coefficients usually vary with time and space (Sun et al., 2016). These methods also depend in particular on the time span of the radiation observation data with the climate background, making them unsuitable for DGSR estimation at the national scale. To address the issues with the empirical formula model, by using an artificial neural network method with a hybrid model (Yang et al., 2001), Tang et al. (2013) attempted to produce an estimated DGSR data set across China. However, since data from only 716 meteorological stations were used in their study, their estimated DGSR exhibited large spatial discontinuity. Thus, an in-depth understanding of the distribution and intensity levels of DGSR remains limited and warrants further investigation. As there are 2,474 meteorological stations across China, a more accurate and more robust DGSR network derived from this high-density meteorological data set should be constructed, and the spatial features of DGSR over China should be examined in more detail.

Based on a high-density ground-based meteorological observation data set, we use the random forest (RF) model, a popular and highly flexible machine learning algorithm that can identify variable importance (Wang et al., 2019; Yang, Ye, et al., 2019; Ye et al., 2019), to construct a high-density DGSR network and estimate the spatiotemporal distributions of DGSR across China. Following this introduction, the solar radiation and meteorological observation data sets are introduced in section 2. Section 3 explains in detail the flow of the RF modeling for DGSR estimation and variable selection. The results with respect to model performance, importance of variables, and spatial distributions of DGSR are presented and discussed in section 4. Section 5 discusses and concludes the study.

2. Data

2.1. Solar Radiation Data

The DGSR measurements used in this study are from the National Meteorological Information Center, China Meteorological Administration. The data set contains daily mean global solar radiation, diffuse solar radiation, and horizontal direct solar radiation from 130 stations (including replacement stations) in mainland China from May 1957. However, the radiation instruments were updated around 1993, and some stations subsequently stopped measuring DGSR. Currently, only 97 stations scattered throughout mainland China continue to observe DGSR (red squares in Figure 1). Quality-controlled DGSR data were collected from these 97 radiation stations during 2012–2015 (Shi et al., 2008; Tang et al., 2011). Therefore, we use

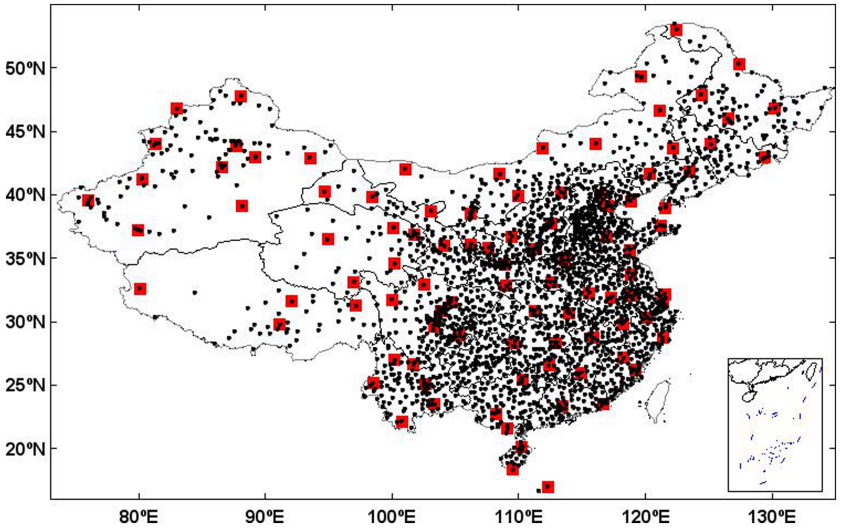

Figure 1. Spatial distribution of the 97 solar radiation observation stations (red squares) and 2379 China Meteorological Administration stations (black dots).

the DGSR data during 2013–2014 to train and test the RF model. We also use the DGSR data in 2012 and 2015 to evaluate the model's hindcast and prediction performances, respectively.

2.2. Meteorological Data

The daily data set of basic meteorological elements of China's national surface meteorological stations (V3.0) contains daily observations of basic meteorological elements measured at 2,474 major stations since January 1951. The main routine meteorological variables, including daily average barometric pressure, daily average relative humidity, daily sunshine duration, daily mean air temperature, daily average wind speed, daily maximum/minimum surface air temperature, daily evaporation capacity, and precipitation from 08:00 to 20:00 (local standard time [LST]), were collected from 2,379 meteorological stations. The spatial distribution of the selected stations can be seen in Figure 1 (black dots), and information on the meteorological variables is provided in Table 1. The quality control of the data set can be referred to in the procedures outlined by Tang et al. (2011). In addition, the variable “evaporation” is directly excluded in this study, due to the lack of observations at many stations.

3. RF Framework for Modeling DGSR

3.1. RF Model Design

The RF model is a popular and highly flexible machine learning algorithm and is capable of analyzing the characteristic of the complex interaction of classification with good robustness for data with noise or missing values (Chen et al., 2018; Wang et al., 2019; Yang, Ye, et al., 2019; Ye et al., 2019). In particular, the RF model has been widely used as a feature selection tool for high-dimensional data for identifying variable importance (Xiao et al., 2018; Wang et al., 2019). Several previous studies using RF to predict DGSR mainly focused on solar radiation at a single site or in a case study within a certain area of China (Sun et al., 2016). On the contrary, for the whole of China, high-density DGSR estimations from surface meteorological observations are scarce. The flow chart for estimating DGSR with the RF algorithm is shown in Figure 2, including three steps as follows:

1. Data matching and variable selection. The model training and testing data pairs are screened by data quality control and spatiotemporal matching. Since each solar radiation station has overlapped with its corresponding meteorological station, spatial matching need not be considered. According to the variable importance characteristics of the RF, the input variables to the RF model are selected and determined by the backward selection method (see section 3.3 for details).

Table 1
List of Predictor Variables for Estimating DGSR

Variable		Unit	Selected[a]	Description
Geographical factors	Longitude	°	Y	—
	Latitude	°	Y	—
	Altitude	m	Y	—
Time factor	DOY	day	Y	Day of year
Estimated factor	DGSR	MJ/m^2	Y	Daily global solar radiation
Meteorological factors	PRS-mean	hPa	Y	Daily average atmospheric pressure
	PRS-min	hPa	Y	Daily minimum atmospheric pressure
	RHU	%	Y	Daily average relative humidity
	SSD	h	Y	Daily sunshine duration
	SAT-mean	°C	Y	Daily mean surface air temperature
	DTR	°C	Y	Diurnal temperature range (SAT-max minus SAT-min)
	WIN	m/s	Y	Daily average wind speed
	LST-mean	°C	Y	Daily mean land surface temperature
	LST-max	°C	Y	Daily maximum land surface temperature
	LST-min	°C	Y	Daily minimum land surface temperature
	EVP	mm	N	Daily evaporation capacity
	PRS-max	hPa	N	Daily maximum atmospheric pressure
	SAT-min	°C	N	Daily minimum surface air temperature
	SAT-max	°C	N	Daily maximum surface air temperature
	PRE-2008	mm	N	Precipitation from 20:00 to 8:00
	PRE-0820	mm	N	Precipitation from 8:00 to 20:00
	PRE-2020	mm	N	Precipitation from 20:00 to 20:00

[a]Y: Included in the reduced model after variable selection.

2. RF model building, training, and testing. RF has interpolation and extrapolation and spatial interpolation functions (Hengl et al., 2018) when inputting the geoinformation parameters. Here, we use the selected predicters and other variables (latitude, longitude, altitude, and day of year [DOY]) as the final variables then train and test the model using the tenfold site-based cross-validation (CV) method. Data pairs of solar radiation and other meteorological observations from 97 sites in 2013 and 2014 are used to build, train, and test the final RF model in our study.
3. Model hindcast performance validation and high-density DGSR network construction. The data for 2012 (2015) are used to assess the performance of the model hindcast (prediction). Meanwhile, the DGSR at those meteorological sites with no solar radiation observations is estimated. Thus, the distribution of site solar radiation (DGSR network) across China can be acquired. Furthermore, the ordinary kriging interpolation model is used to obtain the spatial distribution of grid-scale DGSR across China.

To evaluate the performance of the models and for consistency with previous validation studies (Zou et al., 2019; Yang, Ye, et al., 2019; Wang et al., 2019), the tenfold CV method is used in our study and four statistical metrics—the coefficient of determination (R^2), root-mean-square error (RMSE), mean fractional bias (MB), and correlation coefficient (R)—are used to measure the prediction performance. The four statistical metrics are calculated at each radiation site using the following equations (1)–(4), respectively:

$$R^2=\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y}_i)(x_i-\bar{x}_i)\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}, \tag{1}$$

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(x_i-\bar{x}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}, \tag{2}$$

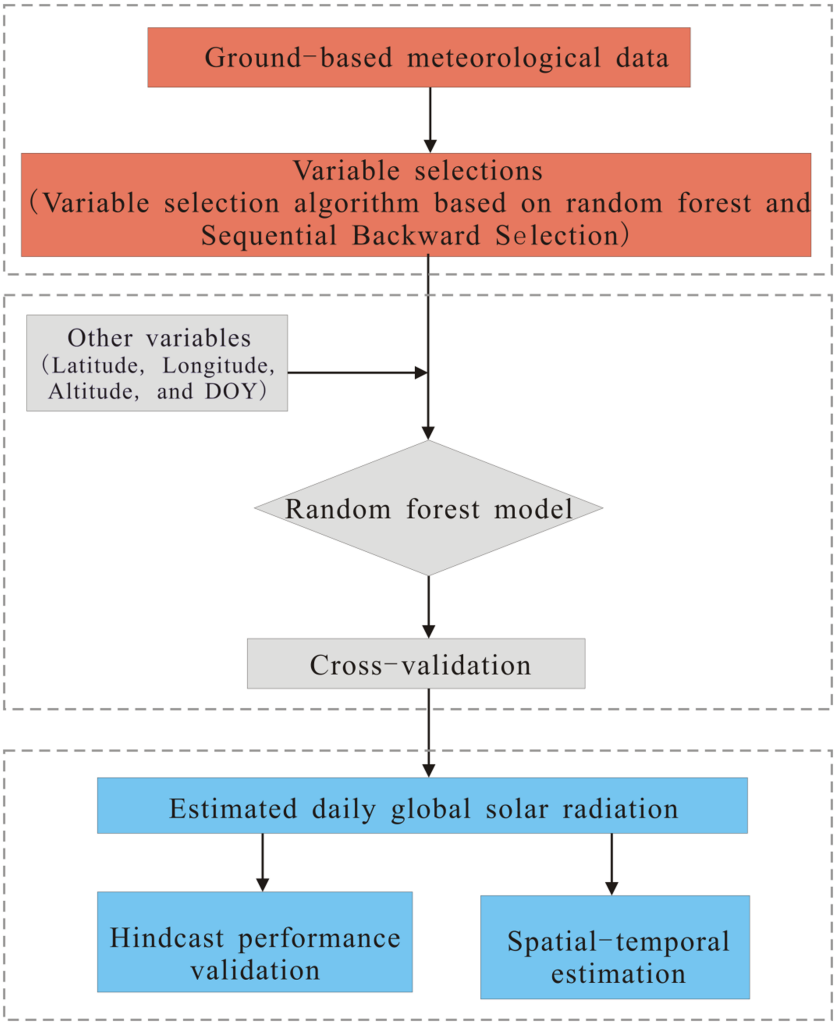

Figure 2. Flowchart of the RF model for estimating DGSR.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}, \tag{3}$$

$$\mathrm{MB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i), \tag{4}$$

where n denotes the number of sample points, x_i represents the observed DGSR value of x for sample i, y_i represents the estimated DGSR value of y for sample i, and $\bar{x}_i$ and $\bar{y}_i$ represent the average of x_i and y_i, respectively.

The steps of the tenfold CV method involve first dividing the data set into 10 equal parts. Nine of them are used as training data sets to construct the RF model, and the remaining one is used as the validation data set for verification. The process is repeated 10 times. Each set of samples is verified once, and the values of the tenfold CV are averaged to obtain the final result.

3.2. Variable Selection

The optimal variable subset is extracted from the original variable set, so that the classification or regression model constructed by the optimal feature subset can achieve a prediction accuracy that is similar to, or even better than, the previous variable selection (Genuer et al., 2010). One of the advantages in the RF model is that it can measure the importance of each variable, so that the most important variable can be selected for model construction. Using the RF algorithm, we first calculate the importance of all variables

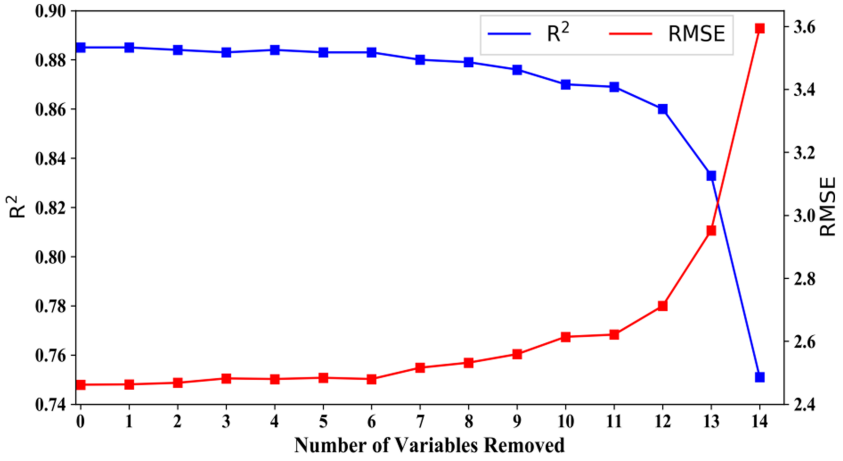

Figure 3. Predictive performance (CV R^2 and RMSE) of the RF model during the variable selection process. Note that Steps 15 and 16, where RMSE increases dramatically, are not shown in the figure. The predictor variables are removed one at a time in the following order: (1) PRE-2020, (2) PRE-0820, (3) PRE-2008, (4) SAT-max, (5) SAT-min, (6) PRS-max, (7) SAT-mean, (8) RHU, (9) DTR, (10) WIN, (11) PRS-min, (12) LST-min, (13) LST-max, (14) PRS-mean, (15) LST-mean, and (16) SSD.

and sort them based on sequential backward selection. We then remove the least important variable and repeat the above procedure until only one variable remains in the model. Finally, according to the evaluation index of the RF model (here, R^2 and RMSE) in each cycle procedure, the variable subset with the least number of variables and the highest prediction accuracy is obtained as the result of variable selection.

To obtain optimum parameters, we perform a sensitivity test and employ the tenfold CV method to test the model estimation performance. In this process, the mean R^2, RMSE, and the corresponding order in relative variable importance after tenfold CV should be recorded. The process and results of variable selection are shown in Figure 3 (note that Steps 15 and 16, where RMSE increases dramatically, are not shown in the figure). The best performance (i.e., highest R^2 [0.884] and lowest RMSE [1.743 MJ/m^2]) for the RF model appears when the sixth variable is removed during variable selection, with 10 meteorological variables remaining as predictors (SAT-mean, RHU, DTR, WIN, PRS-min, LST-min, LST-max, PRS-mean, LST-mean, and SSD).

Finally, the input variables for the solar radiation model are the above 10 remaining meteorological variables, the latitudes, longitudes and altitudes of sites, and dummy variables (DOY). All these variables are key predictors in the RF model to estimate spatiotemporal variations in solar radiation. In addition, we also need to determine the two most important parameters of the model itself, that is, the number of trees to grow (*N*tree) and the number of variables randomly sampled as candidates at each split (*M*try). The default value of *M*try is the square root of p in classification and $p/3$ in regression, where p is the number of all characteristic variables (Liaw et al. 2002). Figure 4 shows the relationship between *N*tree and R^2/RMSE when *M*try is equal to 5 (here, $p/3 = 5$). For instance, when *N*tree is 500, R^2 increases by 0.01, while RMSE is the smallest. With increases in *N*tree, R^2 and RMSE tend to be stable. Therefore, *M*try = 5 and *N*tree = 500 are set for the final RF model.

4. Results

4.1. Statistical Characteristics

Table 2 lists the annual and seasonal statistical properties of the final modeling variables in the sample data set. The DGSR ranges from 0.01–40.28 MJ/m^2, with an annual mean and a standard deviation of 14.82 and 7.59 MJ/m^2, respectively. The annual mean (standard deviation) of LST-mean, LST-max, LST-min, PRS-mean, PRS-min, RHU, SSD, SAT-mean, WIN, and DTR are 15.18 (13.05), 32.74 (17.56), 6.25 (12.54), 913.98 (113.93), 911.19 (113.91), 61.13 (20.65), 6.53 (4.02), 12.09 (12.29), 10.99 (4.93), and 2.19 (1.25),

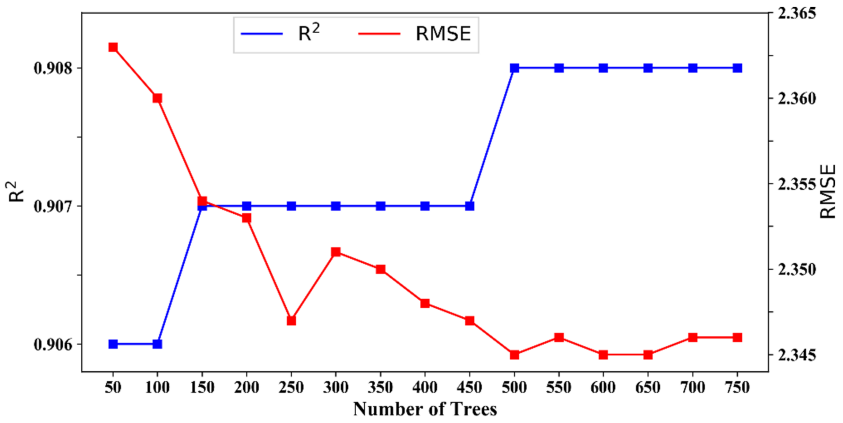

Figure 4. Relationship between *N*tree and R^2/RMSE when *M*try is equal to 5 for the RF model predicting the DGSR in China.

respectively. The mean values for DGSR and most selected meteorological variables (except PRS-mean, PRS-min, WIN, and DTR) are the highest for summer, followed by spring and fall, and the lowest for winter. Due to the large spatiotemporal variability of samples, the ranges of model input variables are also wide.

4.2. Variable Importance Contribution

Figure 5 illustrates the importance of input variables as predictors in the final RF model. SSD accounts for 47.37% of the overall variable importance and therefore is the most important to estimate solar radiation. This is followed by another four dominant variables—namely, LST-max, DOY, latitude, and LST-mean, with importance values of 17.75%, 8.16%, 5.04%, and 5.02%, respectively. The high importance of SSD to daily solar radiation has also been identified in previous modeling studies (Chen et al., 2013; Sun et al., 2015, 2016). Note that SSD can also imply and contain the impacts of air pollution on DGSR (Wang et al., 2012, 2013). Moreover, the correlation coefficient (*R*) is also calculated, to investigate the relationships between DGSR and SSD/LTS-max/LTS-mean (see Figure 6). Their *R* values are 0.8, 0.71, and 0.53, respectively, which are consistent with the variable importance values. DTR, PRS-mean, LST-min, RHU, and altitude show relatively low importance for estimating DGSR; their corresponding importance values are 2.63%, 2.50%, 2.21%, 2.07%, 1.84%, 1.82%, and 1.80%, respectively. In contrast, WIN and SAT-mean have the lowest variable importance, with importance values of 1.17% and 1.00%, respectively. In addition to meteorological variables, the DOY (seasonal effects) and latitude (geographical factor) are critical to estimate DGSR, consistent with previous results (Li et al., 2010; Li et al., 2013; Sun et al., 2016).

Table 2
Descriptive Statistics of the Modeling Variables in the Training Data Set

	Annual		MAM		JJA		SON		DJF	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
LST-mean	15.18	13.05	16.86	9.20	27.74	6.42	14.91	10.02	0.97	9.70
LST-max	32.74	17.56	37.05	14.13	46.70	11.81	31.44	13.49	15.40	14.04
LST-min	6.25	12.54	6.19	9.59	17.86	6.31	6.79	10.26	-5.98	10.37
PRS-mean	913.98	113.93	913.07	113.17	907.70	109.71	917.11	113.18	920.58	117.23
PRS-min	911.19	113.91	909.78	113.07	905.33	109.85	914.49	113.19	917.63	117.19
RHU	61.13	20.65	53.86	23.27	66.46	18.54	64.70	18.29	59.45	19.55
SSD	6.53	4.02	7.08	4.18	7.19	4.43	6.21	3.80	5.57	3.38
SAT-mean	12.09	12.29	13.32	8.84	23.16	5.57	12.68	9.29	−0.91	10.76
WIN	10.99	4.93	11.86	5.08	10.28	4.25	10.69	4.92	11.04	5.25
DTR	2.19	1.25	2.49	1.31	2.16	1.05	2.04	1.26	2.06	1.29
DGSR	14.82	7.59	17.56	7.48	19.28	7.52	12.62	6.00	9.58	4.72

Note. MAM: March, April, and May; JJA: June, July, and August; SON: September, October, and November; DJF: December, January, and February; Std: Standard deviation.

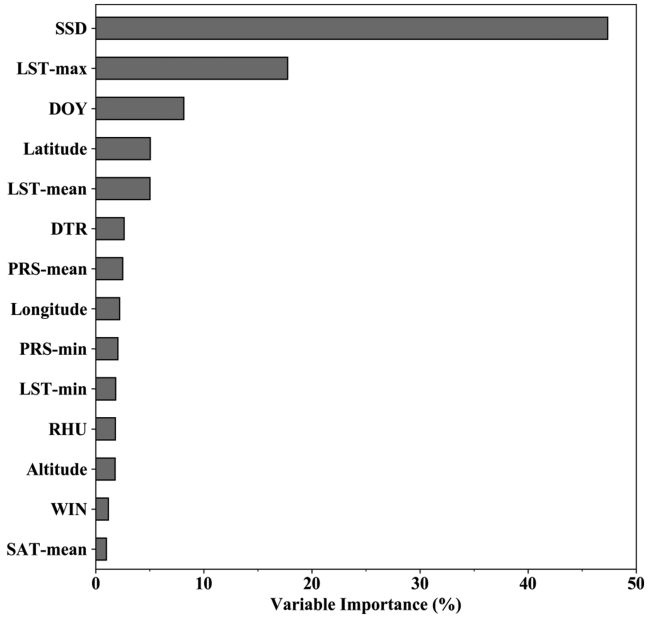

Figure 5. Variable importance plot for the RF model estimating the DGSR in China.

4.3. Predictive Performance

Figure 7a shows scatterplots of observed and estimated DGSR samples from the results of tenfold CV from the RF model on the prediction of DGSR at 97 ground sites during 2013–2014. The *R*, RMSE, and MB values of the model CV results are 0.95, 2.34 MJ/m^2, and −0.04 MJ/m^2, respectively. This indicates that the DGSR estimated from the RF model is in good agreement with the actual observations over China during 2013 and 2014. Moreover, the *R*, RMSE, and MB between the observed and estimated DGSR at each site from 2013 to 2014 are shown in Figures 8a and 8b, respectively. In particular, 88 sites present $R > 0.90$ (accounting for ~91% of the total ground sites), while only 7 sites have $0.8 < R < 0.90$. The smallest *R* is 0.76, also significant at the 95% confidence level. The estimated DGSR shows high correlation with observed DGSR at each ground site, indicating that the RF model has high predictive performance over China. For RMSE and MB, except for Tacheng, Yan'an, and Hangzhou stations, which have large RMSE (absolute values of MB) of 4.56 MJ/m^2 (0.8 MJ/m^2), 5.61 MJ/m^2 (1.7 MJ/m^2), and 4.97 MJ/m^2 (1.8 MJ/m^2), respectively, values are small at all sites. In general, ~52 (73) sites present RMSE values of <2 (2.5) MJ/m^2, and spatial differences of RMSE are not evident over China. Higher values are mainly located in central and northern China, followed by southern China, northwestern China, and the Tibetan Plateau. Lower values are in northwestern China, except for the three sites mentioned above. Correspondingly, 87 sites, which account for 91% of all sites, have MB values between −1 and 1. This result suggests that the deviation between estimated and observed solar radiation data sets in China is small. Meanwhile, the geographical distributions in terms of positive and negative MB present regional differences as follows: negative in northern China, the Tibetan Plateau, and northwestern China, close to 0 in southern China, and positive in central China.

To further explore the effects of seasonal and regional differences on the predictions of RF models, the seasonal and regional CV performances of the RF model are also summarized, in Table 3. The *R* values are 0.95, 0.95, 0.96, and 0.95 in spring, summer, fall, and winter, respectively, suggesting that the RF model performs well in estimating DGSR for all seasons. Both the RMSE and absolute values of MB are also small. Nevertheless, the smallest (largest) RMSE values are found in winter (summer), which may be related to the lowest (highest) values of DGSR. Negatively small MB values indicate predictive values are slightly underestimated in all seasons. In general, the RF model has good predictive performance for all seasons, demonstrating that the seasonal means of estimated DGSR are highly consistent with observed DGSR.

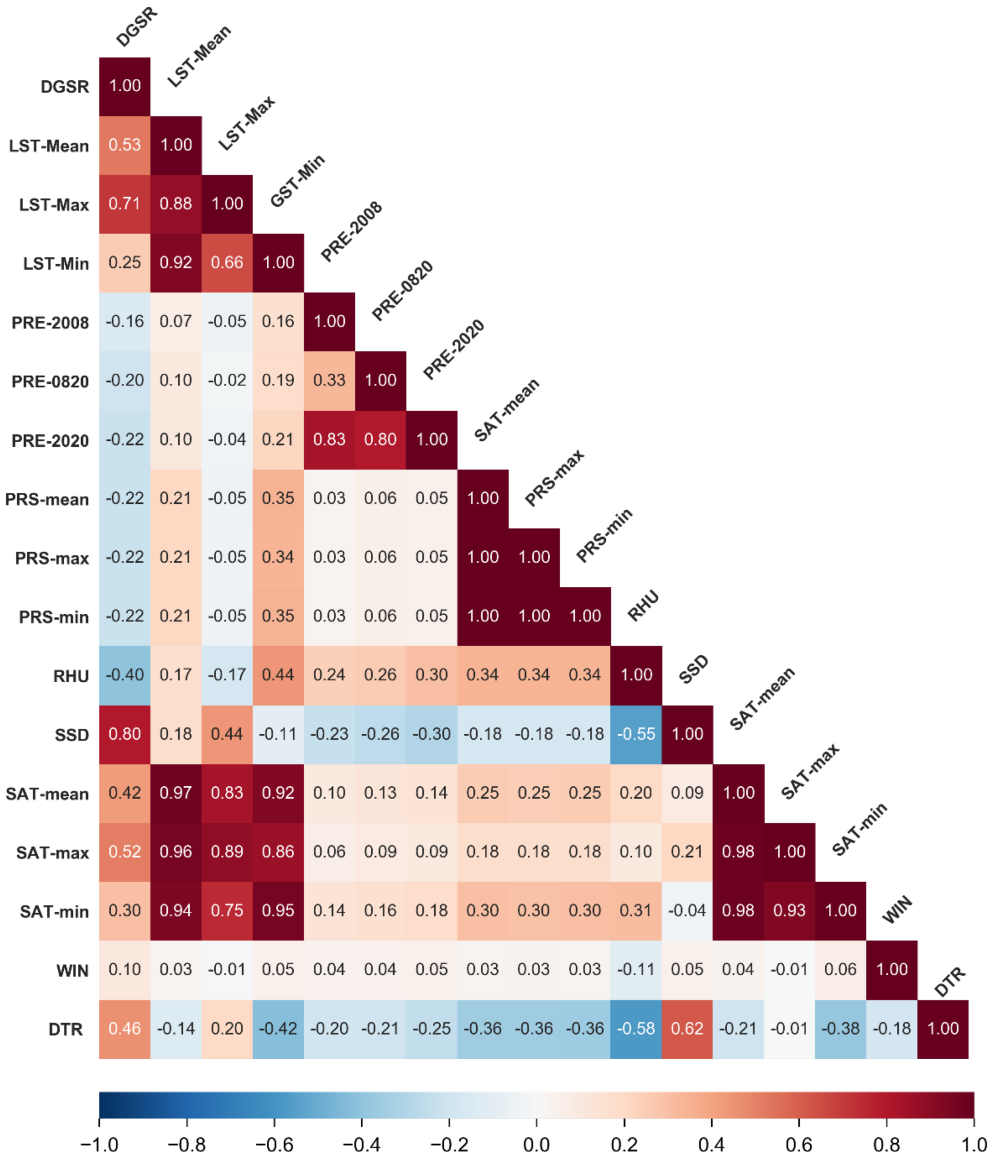

Figure 6. Correlations among the DGSR and the meteorological variables. Colors indicate the Spearman's rank correlation coefficient values.

Based on the Chinese power transmission system (Li et al., 2017), seven subregions, representing the northeastern grid (a), the northwestern grid (b), the northern grid (c), the Tibetan grid (d), the central grid (e), the eastern grid (f), and the southern grid (g), are used to evaluate the predictive performance of the RF model across China. Table 3 also shows the statistical information for the predictive performance of the RF model for these seven subregions, where high R values (0.94–0.96) and small RMSE (absolute values of MB) [2.02–2.69 MJ/m^2 (0.01–0.04 MJ/m^2)] are observed. Slight positive MB values can be found in the northern, Tibetan, and central subregions, while slight negative MB appears in the other subregions, where DGSR is slightly underestimated. Generally, the predictive performance of the RF model is high and reasonably consistent across the seven subregions of China, implying that DGSR can be estimated by the final RF model for any region where DGSR observations are unavailable.

To comprehensively evaluate the performance of the RF model, the same training samples are used to establish other models for estimating DGSR, including a decision tree (DT) (Quinlan, 1986), BP neural network (BP) (Wen et al., 2002), support vector machine (SVM) (Cortes & Vapnik, 1995), and multiple linear

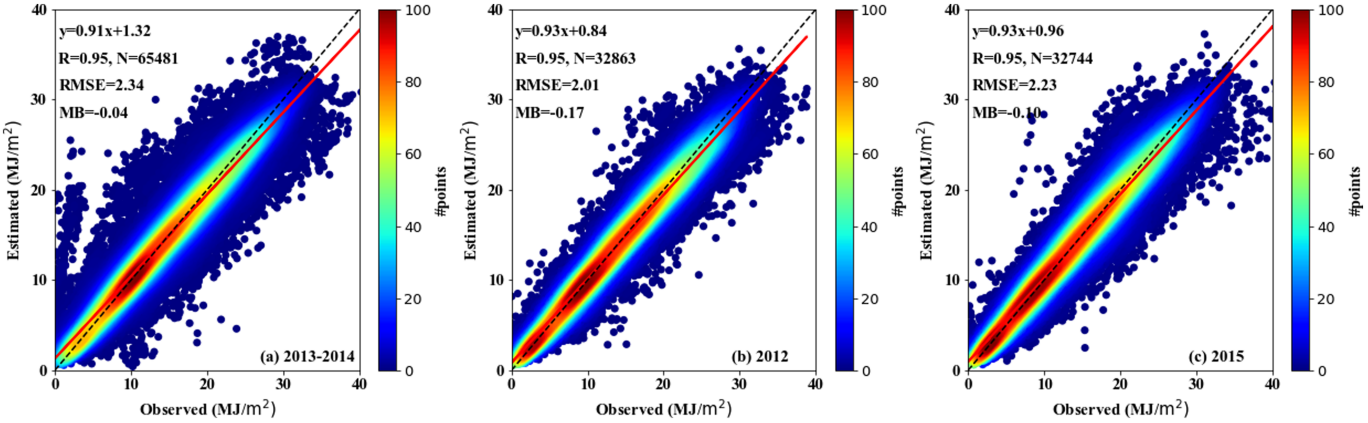

Figure 7. Scatterplots of CV results for the predictive performance of the RF model in predicting DGSR in (a) 2013–2014, (b) 2012, and (c) 2015 for 97 sites over China.

regression (MLR) (Zelterman, 2015). The samples from 2013 to 2014 are trained and predicted, and the prediction performances of the five models are compared (see Table 4). The RF model produces the best performance in estimating DGSR, with the highest R (0.95) and the smallest RMSE (2.34 MJ/m^2), followed by MLR, SVM, and BP, while DT has the lowest CV performance.

4.4. Model Hindcast Performance

Figures 7b and 7c show the RF-estimated and the observed DGSR for all sites across China in 2012 and 2015. The estimated DGSR presents good consistency with the observed DGSR, with a high correlation coefficient of 0.95. The slope of the regression equation is 0.93, showing that the estimated DGSR is slightly smaller than the observed DGSR. Therefore, we can expect that it is feasible to construct (predict) the historical (future) DGSR through the RF model.

In addition, for evaluating the spatial hindcast performance of the RF model, Figure 9 shows the spatial distribution of the annual mean DGSR estimated by the RF model across China at the high-density station scale and 10-km grid scale after kriging interpolation during 2013–2014, 2012, and 2015, respectively. Note that grid-scale DGSR is important to the inputs of climate reanalysis, numerical weather forecasting, and geographical and hydrological models. Figures 9a and 9b show that the annual average DGSR ranges from 6 to 26 MJ/m^2, with a national-average value of 15.55 MJ/m^2 during 2013–2014. Spatial differences are evident across China, indicating that the solar radiation intensity in northern China (western China) is higher than that in southern China (eastern China). The spatial distributions of annual mean DGSR both in 2012 and

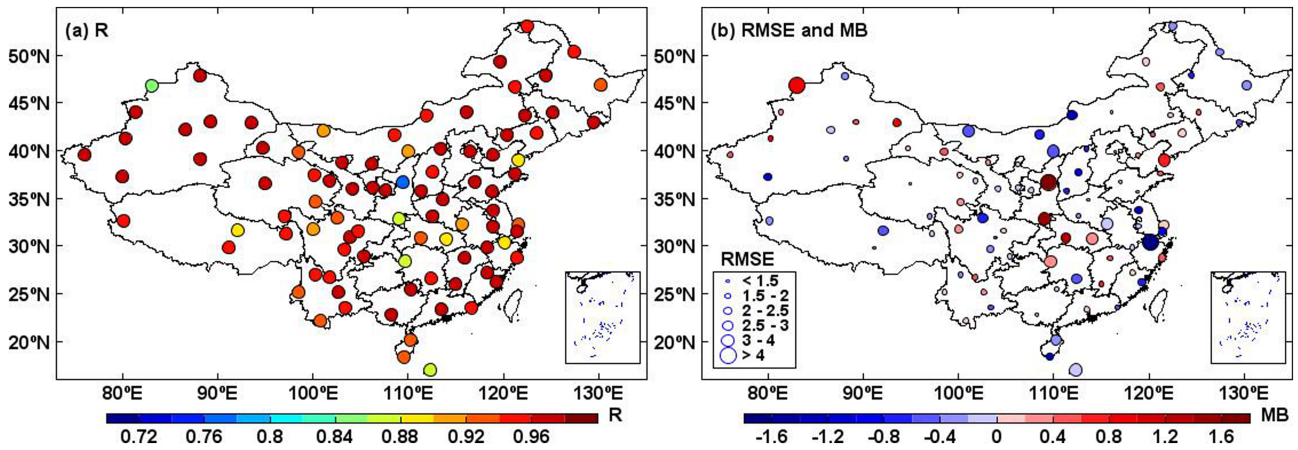

Figure 8. (a) Spatial distributions of the CV correlation coefficient (R) between observed and estimated DGSR. (b) Spatial distributions of the RMSE and MB of DGSR. The circle size (color shading) represents the RMSE (MB) values of the DGSR.

Table 3
Tenfold Cross-Validation Results of the Random Forest Model in Different Seasons and in Different Subregions

	N	*R*	RMSE (MJ/m^2)	MB (MJ/m^2)	Slope
MAM	16,500	0.95	2.40	−0.03	0.89
JJA	16,573	0.95	2.36	−0.02	0.89
SON	16,336	0.96	1.84	−0.06	0.89
DJF	16,142	0.95	1.53	−0.04	0.88
Northeastern grid	8,727	0.96	2.25	−0.03	0.91
Northwestern grid	17,480	0.95	2.52	−0.04	0.91
Northern grid	9,487	0.96	2.30	0.01	0.93
Tibetan grid	2,907	0.94	2.02	0.01	0.87
Central grid	10,853	0.95	2.50	0.03	0.89
Eastern grid	7,293	0.94	2.69	−0.04	0.88
Southern grid	8,734	0.95	2.05	−0.01	0.91
Whole China	65,481	0.95	2.34	−0.04	0.91

Note. MAM: March, April, and May; JJA: June, July, and August; SON: September, October, and November; DJF: December, January, and February; *N*: number of samples; *R*: correlation coefficient; RMSE: root-mean-square error; MB: mean prediction bias.

2015 are highly consistent with that during 2013–2014 (see Figure 9), suggesting that the spatial pattern of DGSR changes little year on year. Comparing the interpolation results with observations at stations (Figures 9b, 9d, and 9f), the spatial distribution and magnitude of the annual observed and estimated DGSR are highly consistent, suggesting that both the prediction and hindcast performance of the RF model are relatively high. Therefore, it will be possible in future work to reconstruct high-accuracy, high-density, and long-term DGSR data sets by using the combined RF model and kriging model.

5. Discussion and Conclusion

5.1. Discussion

Many approaches have been developed to estimate surface solar radiation accurately (Khatib et al., 2012; Liu et al., 2019; Olatomiwa et al., 2015; Sun et al., 2015; Wang et al., 2016; Zhang et al., 2017), which can be generally categorized as follows:

1. Estimation by semiempirical and semiphysical formulae with conventional meteorological parameters (e.g., sunshine hour, cloud, temperature, and humidity) as inputs (Angstrom, 1924; Davies et al., 1975; Thornton & Running, 1999; Bakirci, 2009; Besharat et al., 2013; Hassan et al., 2016; Liu et al., 2019). A simple empirical model can be established with its coefficients varying with time and space, which need to be calibrated using long-term radiation observation data in certain areas.
2. Estimation by the artificial neural network method with meteorological observations (Jiang, 2009a; Linares-Rodriguez et al., 2013; Tang et al., 2013; Ramedani et al., 2014; Kashyap et al., 2015; Wang et al., 2016). The artificial neural network method requires a large number of samples to train the model in a local area, and the trained model may not be applicable in other areas.
3. Retrieval by satellite-based radiation (Ceballos, 2004; Liang et al., 2006; Mueller et al., 2009; Lu et al., 2010; Huang et al., 2011; Qin et al., 2011; Ma & Pinker, 2012; Jia et al., 2013; Zhang et al., 2014). The satellite inversion method is relatively new. However, it is limited by a low sampling frequency, and it is difficult to produce time series data sets over long periods.

Table 4
Comparison of the Accuracies in Estimating DGSR Based on Different Models

Model	*R*	RMSE (MJ/m^2)
DT (decision tree)	0.90	3.22
BP (BP neural network)	0.90	3.17
SVM (support vector machine)	0.92	2.85
MLR (multiple linear regression)	0.93	2.84
RF (random forest)	0.95	2.34

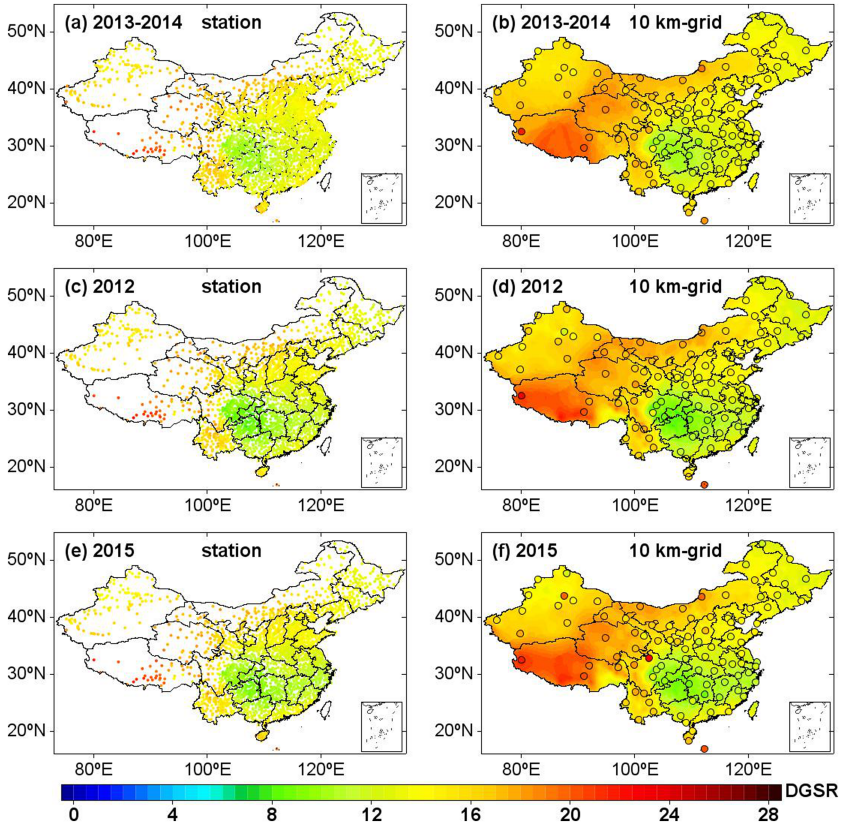

Figure 9. The high-density station distribution (left-hand panel) and 10-km grid distribution (right-hand panel) annual mean DGSR estimated by the RF model over China in (a, b) 2013–2014, (c, d) 2012, and (e, f) 2015. The dots in the right-hand panel represent the corresponding observed annual mean DGSR at 97 ground sites. The units of DGSR are in MJ/m^2.

4. Prediction by an atmospheric radiative transfer model or numerical prediction model (Yang et al., 2001, 2006; Mathiesen & Kleissl, 2011. Perez et al., 2013; Lorenz et al., 2016; Ruiz-Arias et al., 2011). This model simulation method has a solid physical basis, but the structure of the model is complex, and the input parameters—including ozone thickness, aerosol content, atmospheric precipitation, and other variables—are difficult to obtain in real time.

Overall, the aforementioned methods need to input different parameters to estimated DGSR models and then output different solar-related products, while each method has its own advantages and disadvantages.

According to our results from other models (DT, BP, SVM, and MLR), RF is more suitable for DGSR predictions at a large scale with good performance and shows the importance of all input variables for estimating DGSR at the national scale, which is consistent with previous work (Sun et al., 2016). Specifically, 10 meteorological variables, as well as the latitudes, longitudes and altitudes of sites, and dummy variables, are evaluated in our work as predictors to employ in RF models, indicating that daily sunshine duration is the most important contributing factor in estimating DGSR, and daily maximum land surface temperature and DOY also play crucial roles in determining DGSR across China. In summary, the present work has built an RF modeling framework for estimating high-density DGSR at the national level in China, implying that high-resolution DGSR data can be made available to effectively evaluate the long-term change and trend of solar radiation, and their causes, which will be reported separately in a future publication.

6. Conclusion

The present study introduces the RF model, a popular and highly flexible machine learning algorithm, to estimate solar radiation across China at the national scale. The estimated DGSR is in good agreement with

site observations across China, with mean R, RMSE, and MB values of 0.95, 2.34, and −0.04 MJ/m^2, respectively. The geographical distributions of R, RMSE, and MB values at each ground site indicate that the RF model has high predictive performance in estimating DGSR under different climatic and geographic conditions across China. In addition, the RF model also presents good predictive performance for all seasons across different regions where DGSR is not observed. Moreover, the importance of all the input variables for the RF model for the estimation of DGSR indicates that daily sunshine duration, daily maximum land surface temperature, and DOY play crucial roles in determining DGSR across China. Using the RF model, both at the station and grid scale, it is possible to effectively evaluate the long-term trend in solar energy through reconstructing high-spatial-resolution DGSR data sets from high-density meteorological stations in China. It can also help to more strategically utilize and regulate solar energy on an ongoing basis as one of the most important green energy resources in China. In addition, the estimated high-spatial-resolution DGSR has the potential to be applied in other sectors in China, such as agriculture, ecology, hydrology, and meteorology, to explore spatial details more accurately.

Acknowledgments
This work was supported by the National Natural Science Foundation of China under Grants 41776195, 41531069, and 41871029 and the open funding of State Key Laboratory of Loess and Quaternary Geology (SKLLQG1842). http://data.cma.cn/site/index.htmlAll meteorological data input Random Forest Model 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, hourly data can be found at http://data.cma.cn/en/?r=data/detail&dataCode=A.0012.0001, and daily observations are found at http://data.cma.cn/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0.

References

Angstrom, A. (1924). Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. *Quarterly Journal of the Royal Meteorological Society*, *50*(210), 121–126. https://doi.org/10.1002/qj.49705021008

Assouline, D., Mohajeri, N., & Scartezzini, J. L. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using random forests. *Applied Energy*, *217*, 189–211. https://doi.org/10.1016/j.apenergy.2018.02.118

Bakirci, K. (2009). Models of solar radiation with hours of bright sunshine: A review. *Renewable and Sustainable Energy Reviews*, *13*(9), 2580–2588. https://doi.org/10.1016/j.rser.2009.07.011

Besharat, F., Dehghan, A. A., & Faghih, A. R. (2013). Empirical models for estimating global solar radiation: A review and case study. *Renewable and Sustainable Energy Reviews*, *21*, 798–821. https://doi.org/10.1016/j.rser.2012.12.043

Ceballos, J. C. (2004). A simplified physical model for assessing solar radiation over Brazil using GOES 8 visible imagery. *Journal of Geophysical Research*, *109*(D2), D02211. https://doi.org/10.1029/2003JD003531

Che, H. Z., Shi, G. Y., Zhang, X. Y., Arimoto, R., Zhao, J. Q., Xu, L., et al. (2005). Analysis of 40 years of solar radiation data from China, 1961-2000. *Geophysical Research Letters*, *32*(6), L06803. https://doi.org/10.1029/2004GL022322

Chen, J. L., Li, G. S., & Wu, S. J. (2013). Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy Conversion and Management*, *75*, 311–318. https://doi.org/10.1016/j.enconman.2013.06.034

Chen, W., Xie, X., Peng, J., Shahabi, H., Hong, H., Bui, D. T., et al. (2018). GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena*, *164*, 135–149. https://doi.org/10.1016/j.catena.2018.01.012

Cline, D. W., Bales, R. C., & Dozier, J. (1998). Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling. *Water Resources Research*, *34*(5), 1275–1285. https://doi.org/10.1029/97WR03755

Cooter, E. J., & Dhakhwa, G. B. (1996). A solar radiation model for use in biological applications in the South and Southeastern USA. *Agricultural and Forest Meteorology*, *78*(1-2), 31–51. https://doi.org/10.1016/0168-1923(95)02241-4

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Davies, J. A., Schertzer, W., & Nunez, M. (1975). Estimating global solar radiation. *Boundary-Layer Meteorology*, *9*(1), 33–52. https://doi.org/10.1007/BF00232252

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Guo, J., Li, Y., Cohen, J. B., Li, J., Chen, D., Xu, H., et al. (2019). Shift in the temporal trend of boundary layer height in China using long-term (1979-2016) radiosonde data. *Geophysical Research Letters*, *46*(11), 6080–6089. https://doi.org/10.1029/2019GL082666

Guo, J., Liu, H., Li, Z., Rosenfeld, D., Jiang, M., Xu, W., et al. (2018). Aerosol-induced changes in the vertical structure of precipitation: A perspective of TRMM precipitation radar. *Atmospheric Chemistry and Physics*, *18*, 13,329–13,343. https://doi.org/10.5194/acp-18-13329-2018

Guo, J., Miao, Y., Zhang, Y., Liu, H., Li, Z., Zhang, W., et al. (2016). The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data. *Atmospheric Chemistry and Physics*, *16*, 13,309–13,319. https://doi.org/10.5194/acp-16-13309-2016

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

Hassan, G. E., Youssef, M. E., Mohamed, Z. E., Ali, M. A., & Hanafy, A. A. (2016). New temperature-based models for predicting global solar radiation. *Applied Energy*, *179*, 437–450. https://doi.org/10.1016/j.apenergy.2016.07.006

He, Y., & Wang, K. (2020). Variability in direct and diffuse solar radiation across China from 1958 to 2017. *Geophysical Research Letters*, *47*, e2019GL084570. https://doi.org/10.1029/2019GL084570

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518. https://doi.org/10.7717/peerj.5518

Hoogenboom, G. (2000). Contribution of agrometeorology to the simulation of crop production and its applications. *Agricultural and Forest Meteorology*, *103*(1-2), 137–157. https://doi.org/10.1016/S0168-1923(00)00108-8

Huang, G., Ma, M., Liang, S., Liu, S., & Li, X. (2011). A LUT-based approach to estimate surface solar irradiance by combining MODIS and MTSAT data. *Journal of Geophysical Research – Atmospheres*, *116*(D22), n/a. https://doi.org/10.1029/2011JD016120

Jia, B., Xie, Z., Dai, A., Shi, C., & Chen, F. (2013). Evaluation of satellite and reanalysis products of downward surface solar radiation over East Asia: Spatial and seasonal variations. *Journal of Geophysical Research – Atmospheres*, *118*(9), 3431–3446. https://doi.org/10.1002/jgrd.503532013

Jiang, Y. (2009a). Computation of monthly mean daily global solar radiation in China using artificial neural networks and comparison with other empirical models. *Energy*, *34*(9), 1276–1283. https://doi.org/10.1016/j.energy.2009.05.009

Jiang, Y. (2009b). Estimation of monthly mean daily diffuse radiation in China. *Applied Energy*, *86*(9), 1458–1464. https://doi.org/10.1016/j.apenergy.2009.01.002

Kashyap, Y., Bansal, A., & Sao, A. K. (2015). Solar radiation forecasting with multiple parameters neural networks. *Renewable and Sustainable Energy Reviews*, *49*, 825–835. https://doi.org/10.1016/j.rser.2015.04.077

Khatib, T., Mohamed, A., & Sopian, K. (2012). A review of solar energy modeling techniques. *Renewable and Sustainable Energy Reviews*, *16*(5), 2864–2869.

Li, H., Ma, W., Lian, Y., & Wang, X. (2010). Estimating daily global solar radiation by day of year in China. *Applied Energy*, *87*, 3011–3017. https://doi.org/10.1016/j.apenergy.2010.03.028

Li, M. F., Tang, X. P., Wu, W., & Bin, L. H. (2013). General models for estimating daily global solar radiation for different solar radiation zones in mainland China. *Energy Conversion and Management*, *70*, 139–148. https://doi.org/10.1016/j.enconman.2013.03.004

Li, X., Wagner, F., Peng, W., Yang, J., & Mauzerall, D. L. (2017). Reduction of solar photovoltaic resources due to air pollution in China. *Proceedings of the National Academy of Sciences*, *114*(45), 11,867–11,872. https://doi.org/10.1073/pnas.1711462114

Liang, S., Zheng, T., Liu, R., Fang, H., Tsay, S. C., & Running, S. (2006). Estimation of incident photosynthetically active radiation from Moderate Resolution Imaging Spectrometer data. *Journal of Geophysical Research – Atmospheres*, *111*(D15), D15208. https://doi.org/10.1029/2005JD006730

Linares-Rodriguez, A., Ruiz-Arias, J. A., Pozo-Vazquez, D., & Tovar-Pescador, J. (2013). An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images. *Energy*, *61*, 636–645. https://doi.org/10.1016/j.energy.2013.09.008

Liu, Y., Tan, Q., & Pan, T. (2019). Determining the parameters of the Ångström-Prescott model for estimating solar radiation in different regions of China: Calibration and modeling. *Earth and Space Science*, *6*(10), 1976–1986. https://doi.org/10.1029/2019EA000635

Lorenz, E., Kühnert, J., Heinemann, D., Nielsen, K. P., Remund, J., & Müller, S. C. (2016). Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. *Progress in Photovoltaics: Research and Applications*, *24*(12), 1626–1640. https://doi.org/10.1002/pip.2799

Lou, M., Guo, J., Wang, L., Xu, H., Chen, D., Miao, Y., et al. (2019). On the relationship between aerosol and boundary layer height in summer in China under different thermodynamic conditions. *Earth Space Science.*, 2019EA000620. https://doi.org/10.1029/2019EA000620

Lu, N., Liu, R., Liu, J., & Liang, S. (2010). An algorithm for estimating downward shortwave radiation from GMS 5 visible imagery and its evaluation over China. *Journal of Geophysical Research – Atmospheres*, *115*(D18), D18102. https://doi.org/10.1029/2009JD013457

Ma, Y., & Pinker, R. T. (2012). Modeling shortwave radiative fluxes from satellites. *Journal of Geophysical Research – Atmospheres*, *117*(D23), n/a. https://doi.org/10.1029/2012JD018332

Mathiesen, P., & Kleissl, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Solar Energy*, *85*(5), 967–977. https://doi.org/10.1016/j.solener.2011.02.013

Mueller, R. W., Matsoukas, C., Gratzki, A., Behr, H. D., & Hollmann, R. (2009). The CM-SAF operational scheme for the satellite based retrieval of solar surface irradiance—A LUT based eigenvector hybrid approach. *Remote Sensing of Environment*, *113*(5), 1012–1024. https://doi.org/10.1016/j.rse.2009.01.012

Olatomiwa, L., Mekhilef, S., Shamshirband, S., & Petković, D. (2015). Adaptive neuro-fuzzy approach for solar radiation prediction in Nigeria. *Renewable and Sustainable Energy Reviews*, *51*, 1784–1791. https://doi.org/10.1016/j.rser.2015.05.068

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Van Knowe, G., Hemker, K., et al. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, *94*, 305–326. https://doi.org/10.1016/j.solener.2013.05.005

Pohlert, T. (2004). Use of empirical global radiation models for maize growth simulation. *Agricultural and Forest Meteorology*, *126*(1-2), 47–58. https://doi.org/10.1016/j.agrformet.2004.05.003

Power, H. C. (2001). Estimating clear-sky beam radiation from sunshine duration. *Solar Energy*, *71*(4), 217–224. https://doi.org/10.1016/S0038-092X(01)00049-4

Prăvălie, R., Patriche, C., & Bandoc, G. (2019). Spatial assessment of solar energy potential at global scale. A geographical approach. *Journal of Cleaner Production*, *209*, 692–721. https://doi.org/10.1016/j.jclepro.2018.10.239

Qin, J., Chen, Z., Yang, K., Liang, S., & Tang, W. (2011). Estimation of monthly-mean daily global solar radiation based on MODIS and TRMM products. *Applied Energy*, *88*(7), 2480–2489. https://doi.org/10.1016/j.apenergy.2011.01.018

Qin, W., Wang, L., Zhang, M., Niu, Z., Luo, M., Lin, A., & Hu, B. (2019). First effort at constructing a high-density photosynthetically active radiation dataset during 1961–2014 in China. *Journal of Climate*, *32*(10), 2761–2780. https://doi.org/10.1175/jcli-d-18-0590.1

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., & Khoshnevisan, B. (2014). Potential of radial basis function based support vector regression for global solar radiation prediction. *Renewable and Sustainable Energy Reviews*, *39*, 1005–1011. https://doi.org/10.1016/j.rser.2014.07.108

Ruiz-Arias, J. A., Pozo-Vázquez, D., Lara-Fanego, V., Santos-Alamillos, F. J., & Tovar-Pescador, J. (2011). A high-resolution topographic correction method for clear-sky solar irradiance derived with a numerical weather prediction model. *Journal of Applied Meteorology and Climatology*, *50*(12), 2460–2472. https://doi.org/10.1175/2011JAMC2571.1

Shi, G. Y., Hayasaka, T., Ohmura, A., Chen, Z. H., Wang, B., Zhao, J. Q., et al. (2008). Data quality assessment and the long-term trend of ground solar radiation in China. *Journal of Applied Meteorology and Climatology*, *47*(4), 1006–1016. https://doi.org/10.1175/2007JAMC1493.1

Song, Z., Chen, L., Wang, Y., Liu, X., Lin, L., & Luo, M. (2019). Effects of urbanization on the decrease in sunshine duration over eastern China. *Urban Climate*, *28*, 100471. https://doi.org/10.1016/j.uclim.2019.100471

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 Conversion and Management*, *119*, 121–129. https://doi.org/10.1016/j.enconman.2016.04.051

Sun, H., Zhao, N., Zeng, X., & Yan, D. (2015). Study of solar radiation prediction and modeling of relationships between solar radiation and meteorological variables. *Energy Conversion and Management*, *105*, 880–890. https://doi.org/10.1016/j.enconman.2015.08.045

Tang, W., Qin, J., Yang, K., Liu, S., Lu, N., & Niu, X. (2016). Retrieving high-resolution surface solar radiation with cloud parameters derived by combining MODIS and MTSAT data. *Atmospheric Chemistry and Physics*, *16*(4), 2543–2557. https://doi.org/10.5194/acp-16-2543-2016

Tang, W., Yang, K., He, J., & Qin, J. (2010). Quality control and estimation of global solar radiation in China. *Solar Energy*, *84*(3), 466–475. https://doi.org/10.1016/j.solener.2010.01.006

Tang, W., Yang, K., Qin, J., Min, M., & Niu, X. (2018). First effort for constructing a direct solar radiation data set in China for solar energy applications. *Journal of Geophysical Research – Atmospheres*, *123*(3), 1724–1734. https://doi.org/10.1002/2017JD028005

Tang, W. J., Yang, K., Qin, J., Cheng, C. C. K., & He, J. (2011). Solar radiation trend across China in recent decades: A revisit with quality-controlled data. *Atmospheric Chemistry and Physics*, *11*(1), 393–406. https://doi.org/10.5194/acp-11-393-2011

Tang, W. J., Yang, K., Qin, J., & Min, M. (2013). Development of a 50-year daily surface solar radiation dataset over China. *Science China Earth Sciences*, *56*(9), 1555–1565. https://doi.org/10.1007/s11430-012-4542-9

Thornton, P. E., & Running, S. W. (1999). An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agricultural and Forest Meteorology*, *93*(4), 211–228. https://doi.org/10.1016/S0168-1923(98)00126-9

Wang, H., Li, J., Gao, Z., Yim, S. H., Shen, H., Ho, H. C., et al. (2019). High-spatial-resolution population exposure to PM2.5 pollution based on multi-satellite retrievals: A case study of seasonal variation in the Yangtze River Delta, China in 2013. *Remote Sensing*, *11*(23), 2724. https://doi.org/10.3390/rs11232724

Wang, K., Ma, Q., Li, Z., & Wang, J. (2015). Decadal variability of surface incident solar radiation over China: Observations, satellite retrievals, and reanalyses. *Journal of Geophysical Research: Atmospheres*, *120*, 6500–6514. https://doi.org/10.1002/2015JD023420

Wang, L., Kisi, O., Zounemat-Kermani, M., Salazar, G. A., Zhu, Z., & Gong, W. (2016). Solar radiation prediction using different techniques: Model evaluation and comparison. *Renewable and Sustainable Energy Reviews*, *61*, 384–397. https://doi.org/10.1016/j.rser.2016.04.024

Wang, Y. W., & Wild, M. (2016). A new look at solar dimming and brightening in China. *Geophysical Research Letters*, *43*(22), 11,777–11,785.

Wang, Y. W., Yang, Y. H., Han, S. M., Wang, Q. X., & Zhang, J. H. (2013). Sunshine dimming and brightening in Chinese cities (1955–2011) was driven by air pollution rather than clouds. *Climate Research*, *56*(1), 11–20.

Wang, Y. W., Yang, Y. H., Zhao, N., Liu, C., & Wang, Q. X. (2012). The magnitude of the effect of air pollution on sunshine hours in China. *Journal of Geophysical Research-Atmospheres*, *117*, D00V14. https://doi.org/10.1029/2011jd016753

Wang, Y. W., Yang, Y. H., Zhou, X. Y., Zhao, N., & Zhang, J. H. (2014). Air pollution is pushing wind speed into a potential regulator of solar dimming in China. *Environmental Research Letters*, *9*, 054004. https://doi.org/10.1088/1748-9326/9/5/054004

Wen, J., Zhao, J., Luo, S., & Han, Z. (2002). The improvements of BP neural network learning algorithm. https://doi.org/10.1109/icosp.2000.893417

Wild, M. (2009). Global dimming and brightening: A review. *Journal of Geophysical Research – Atmospheres*, *114*, D00D16. https://doi.org/10.1029/2008JD011470

Wu, Y., Connelly, K., Liu, Y., Gu, X., Gao, Y., & Chen, G. Z. (2016). Smart solar concentrators for building integrated photovoltaic façades. *Solar Energy*, *133*, 111–118. https://doi.org/10.1016/j.solener.2016.03.046

Wu, Y., Eames, P., Mallick, T., & Sabry, M. (2012). Experimental characterisation of a Fresnel lens photovoltaic concentrating system. *Solar Energy*, *86*(1), 430–440. https://doi.org/10.1016/j.solener.2011.10.032

Xiao, Q., Chang, H. H., Geng, G., & Liu, Y. (2018). An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environmental Science & Technology*, *52*(22), 13,260–13,269. https://doi.org/10.1021/acs.est.8b02917

Yang, K., Huang, G. W., & Tamai, N. (2001). A Hybrid model for estimating global solar radiation. *Solar Energy*, *70*(1), 13–22. https://doi.org/10.1016/S0038-092X(00)00121-3

Yang, K., Koike, T., & Ye, B. (2006). Improving estimation of hourly, daily, and monthly solar radiation by importing global data sets. *Agricultural and Forest Meteorology*, *137*(1-2), 43–55. https://doi.org/10.1016/j.agrformet.2006.02.001

Yang, X., Ye, T., Zhao, N., Chen, Q., Yue, W., Qi, J., et al. (2019). Population mapping with multi sensor remote sensing images and point-of-interest data. *Remote Sensing*, *11*(5), 574. https://doi.org/10.3390/rs11050574

Yang, Y., Yim, S. H. L., Haywood, J., Osborne, M., Chan, J. C. S., Zeng, Z., & Cheng, J. C. H. (2019). Characteristics of heavy particulate matter pollution events over Hong Kong and their relationships with vertical wind profiles using high-time-resolution Doppler lidar measurements. *Journal of Geophysical Research: Atmospheres*, *124*, 9609–9623. https://doi.org/10.1029/2019JD031140

Yang, Y., Zheng, X., Gao, Z., Wang, H., Wang, T., Li, Y., Lau, G. N. C., & Yim, S. H. L. (2018). Long-term trends of persistent synoptic circulation events in planetary boundary layer and their relationships with haze pollution in winter half year over Eastern China. *Journal of Geophysical Research: Atmospheres*, *123*, 10,991–11,007. https://doi.org/10.1029/2018JD028982

Yang, Y., Zheng, Z., Yim, S. Y. L., Roth, M., Ren, G., Gao, Z., et al. (2020). PM2.5 pollution modulates wintertime urban heat island intensity in the Beijing-Tianjin-Hebei Megalopolis, China. *Geophysical Research Letters*, *47*, e2019GL084288. https://doi.org/10.1029/2019GL084288

Ye, T., Zhao, N., Yang, X., Ouyang, Z., Liu, X., Chen, Q., et al. (2019). Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Science of The Total Environment*, *658*, 936–946. https://doi.org/10.1016/j.scitotenv.2018.12.276

Zelterman, D. (2015). *Applied Multivariate Statistics with R*. Cham: Springer. https://doi.org/10.1007/978-3-319-14093-3

Zhang, J., Zhao, L., Deng, S., Xu, W., & Zhang, Y. (2017). A critical review of the models used to estimate solar radiation. *Renewable and Sustainable Energy Reviews*, *70*, 314–329. https://doi.org/10.1016/j.rser.2016.11.124

Zhang, X., Liang, S., Zhou, G., Wu, H., & Zhao, X. (2014). Generating Global Land Surface Satellite incident shortwave radiation and photosynthetically active radiation products from multiple satellite data. *Remote Sensing of Environment*, *152*, 318–332. https://doi.org/10.1016/j.rse.2014.07.003

Zheng, Z., Li, Y., Wang, H., Ding, H., Li, Y., Gao, Z., & Yang, Y. (2019). Re-evaluating the variation in trend of haze days in the urban areas of Beijing during a recent 36-year period. *Atmospheric Science Letters*, *20*(1), e878. https://doi.org/10.1002/asl.878

Zheng, Z. F., Ren, G. Y., Wang, H., Dou, J. X., Gao, Z. Q., Duan, C. F., et al. (2018). Relationship between Fine Particle Pollution and the Urban Heat Island in Beijing, China: Observational Evidence. *Boundary-Layer Meteorology*, *169*(1), 93–113. https://doi.org/10.1007/s10546-018-0362-6

Zou, L., Wang, L., Li, J., Lu, Y., Gong, W., & Niu, Y. (2019). Global surface solar radiation and photovoltaic power from Coupled Model Intercomparison Project Phase 5 climate models. *Journal of Cleaner Production*, *224*, 304–324. https://doi.org/10.1016/j.jclepro.2019.03.268