Sustainable Energy Technologies and Assessments 52 (2022) 102323

Contents lists available at ScienceDirect

Sustainable Energy Technologies and Assessments

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

Original article

Simplified estimation modeling of land surface solar irradiation: A comparative study in Australia and China

Xuan Liao [a], Rui Zhu [a,b,*], Man Sing Wong [a,b]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[b] Research Institute for Land and Space, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

ARTICLE INFO

Keywords:
Land surface solar irradiation
Machine learning
Cloud optical thickness
Aerosol optical thickness
Himawari-8 satellite images
Meteorological data

ABSTRACT

Solar irradiation maps are fundamental geospatial datasets that have been used in a variety of research fields. However, it is difficult to estimate the continuous distribution of solar irradiation over large areas accurately by using conventional interpolation or extrapolation methods based on only a few observation stations. To tackle this problem, this study proposed a method to estimate spatially continuous land surface solar irradiation based on four machine learning models, namely, Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression (SVR), and Multilayer Perceptron (MLP). Clear-sky solar irradiation data were computed based on time and location, cloud optical thickness (COT) and aerosol optical thickness (AOT) that significantly influence solar irradiation were retrieved from Himawari-8 meteorological satellite images, and land surface solar irradiation data were obtained from observation stations for training and evaluation. To explore the weather effects on land surface solar irradiation, air temperatures, humidity, wind, and atmospheric pressure were also quantified and integrated into the models. As a comparative study, this study collected six-year historical data and estimated solar distribution at a 5-km spatial resolution in Australia and China. Based on the coefficient of determination (R^2), normalized Root Mean Square Error (nRMSE), normalized mean bias error (nMBE), and consumption of time (t), the results show that GBM achieved the highest accuracy with $R^2 > 0.7$ at all stations, followed by RF, SVR, and MLP. It suggests that the proposed method can provide an accurate and reliable estimation of land surface solar irradiation, compared with the theoretical solar irradiation without the obstacle of the atmosphere. The annual solar distribution maps created by the built methods indicate that the proposed method is simple and effective for large geographical regions and can be used worldwide when similar datasets are obtained.

Introduction

During the last few decades, the energy demand has increased by nearly 0.1789 quadrillion kWh with an average growth of 1.2% every year [1], and fossil fuels on the earth tend to be exhausted due to the over-exploitation and utilization of the energy [2]. Simultaneously, the emitted pollutants during the use of traditional energy harm the human living environment, leading to global climate warming and air pollution [3]. To mitigate these problems, the world has recently been focusing on alternative energy sources such as solar energy, wind energy, and tidal energy [4]. Compared to other renewable energy, solar energy is superior in terms of availability, cost-effectiveness, accessibility, capacity, and efficiency [5]. Furthermore, this energy is widely available across the globe, so it can be harvested and utilized in situ without remote transportation. To effectively harvest solar energy, spatio-temporal solar distribution data with accurate quantitative information is increasingly being used in various fields such as agriculture, meteorology, and power systems. For example, solar energy can be used in greenhouses or tunnel farming for the cultivation of crops and vegetables and solar dryers for drying agricultural products [6]. Thus, it is of great importance for the solar industry to estimate the high precision of solar distribution over a large region.

Nowadays, using ground-based stations to observe and record solar irradiation is an effective way to obtain high-precision solar irradiation data. However, the scarcity and uneven distribution of solar irradiation observation stations make it challenging to obtain high precision and continuous solar irradiation data. Moreover, global solar irradiation on the land surface is mainly influenced by astronomical factors and

* Corresponding author at: Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China.
E-mail address: felix.zhu@polyu.edu.hk (R. Zhu).

https://doi.org/10.1016/j.seta.2022.102323
Received 4 October 2021; Received in revised form 7 May 2022; Accepted 22 May 2022
Available online 29 May 2022

Table 1
Climates and ranges of observed solar irradiation of the 22 meteorological stations.

Country	Station Name	Station ID	Climate	Range of observed solar irradiation (kWh/m^2)
Australia	Adelaide	S_1	Mediterranean	0–1.38
	Alice Springs	S_2	Subtropical hot desert	0–1.48
	Broome	S_3	Hot semi-arid	0–1.44
	Cape Grim	S_4	Temperate oceanic	0–1.31
	Cocos Island	S_5	Tropical rainforest	0–1.37
	Darwin	S_6	Tropical savanna	0–1.45
	Geraldton	S_7	Mediterranean	0–1.44
	Kalgoorlie-Boulder	S_8	Semi-arid	0–1.39
	Learmonth	S_9	Hot semi-arid	0–1.36
	Melbourne	S_{10}	Temperate oceanic	0–1.41
	Rockhampton	S_{11}	Humid subtropical	0–1.51
	Townsville	S_{12}	Tropical savanna	0–1.57
	Wagga	S_{13}	Humid subtropical	0–1.43
China	Beijing	S_1	Humid continental	0–9.66
	Guangzhou	S_2	Humid subtropical	0.24–7.81
	Harbin	S_3	Humid continental	0.13–12.13
	Kau Sai Chau	S_4	Humid subtropical	0–1.09
	King’s Park	S_5	Humid subtropical	0–1.08
	Shanghai	S_6	Humid subtropical	0.16–8.65
	Urumqi	S_7	Continental cold semi-arid	0–11.75
	Wenjiang	S_8	Humid subtropical	0.21–8.39
	Wuhan	S_9	Humid subtropical	0.14–8.40

atmospheric factors such as clouds and aerosols [7]. However, it is difficult to obtain large-scale and continuous atmospheric data via atmospheric observation stations, which poses significant challenges for the estimation of high-precision and continuous solar irradiation over large regions. To tackle this problem, researchers have conducted a large number of studies on solar estimation and have made great achievements by developing solar irradiation estimation models over the past two decades [8]. Traditional solar irradiation estimation methods can be organized into three categories: empirical [9–11], physical [12–15], and machine learning models [16–18].

Several researchers employed the empirical model to estimate solar irradiation based on the data from meteorological stations, such as cloudiness-based models, sunshine-based models, temperature-based models, and meteorological parameters-based models [19–23]. However, these empirical models are also limited to a small region although using the empirical models for estimating solar irradiation is convenient. In other words, it is difficult to transform the same empirical model to other regions.

Moreover, plenty of studies focused on physical models to estimate solar irradiation. Physical models commonly used in solar irradiation research include radiation transmission models and parameterized models, such as the METSTA model [24], Bird model [25], Yang model [26], and Page model [27]. Some researchers utilized data acquired from both meteorological stations and satellites [28–30]. Satellite images used in these models can provide large-scale and continuous spatial distribution information, while these models generally estimate low temporal-resolution solar radiation that cannot achieve near real-time monitoring.

Since machine learning can be used in a variety of applications to achieve accurate prediction, various machine learning methods have been developed for estimating solar irradiation in recent years [16]. Generalized machine-learning models have three categories, namely, ANN-based [31–33], Kernel-based [34–36], and Tree-based [37–39]. Compared to physical and empirical models, machine learning models can produce moderate accuracy and wider application for solar irradiation prediction, so it has become one of the most widely used methods for solar estimation.

Ramedani et al. [40] compared the performance of support vector regression (SVR) and fuzzy linear regression for global solar radiation prediction in Iran, in which SVR used the polynomial model (SVR_poly) and radial basis model (SVR_rbf) as the kernel function. The results show that the SVR_rbf model has a better performance than fuzzy linear regression. Srivastava et al. [41] compared the forecasting performance of the 1-day-ahead to 6-day-ahead hourly solar radiation using the Multivariate Adaptive Regression Spline (MARS), Classification and Regression Tree (CART), Piecewise Linear Functions of Regression Trees (M5), and Random Forest (RF) model in India. The result illustrates that the RF model outperformed the MARS, CART, and M5 models. Rabehe et al. [42] assessed the prediction performance of multi-layer perceptron (MLP), boosted decision tree, and a new combination of these models with linear regression for the daily global solar irradiation using a real dataset in the south of Algeria. The results show that the MLP model performs better than the other models. Urraca et al. [43] used the Gradient Boosting Machines (GBMs) to predict daily global horizontal irradiation using the data from 38 ground stations in Castilla-La Mancha with an average mean absolute error (MAE) of 1.63 MJ/m^2 from 2001 to 2013. The results suggest that this model had a good generalization capacity. However, all these studies were limited in available data sources and regions that cannot be solved with empirical models and physical models, so it becomes important for our study to estimate spatially continuous and quantitatively accurate land surface solar irradiation using four machine-learning models with limited datasets.

In summary, although empirical methods for estimating solar irradiation have certain merits, they still have a weak capability to deal with a large geographical extent, such as an estimation covering the whole of China. Physical methods generally combine with satellite images to estimate large-scale solar irradiation, while these images have a relatively low temporal resolution. In this regard, this method is hard to meet the high accuracy requirement on solar irradiation estimation. In comparison, machine learning methods applied in the estimation of solar energy have merits on high prediction accuracy and fast computation. Therefore, combining with the aforementioned reviews, we utilized four machine learning methods, i.e., Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression (SVR), and Multilayer Perceptron (MLP), to establish a robust relation between meteorological data, cloud optical thickness (COT), aerosol optical thickness (AOT), clear-sky radiation and land surface solar irradiation. This study used these data as the main input parameters based on the selected optimal model for estimating solar irradiation with high temporal-spatial resolution over a large geographical extent.

This paper is organized as follows. Section 2 presents a series of datasets used in this study. Section 3 introduces the machine learning-based framework for an accurate estimation of land surface solar irradiation. Section 4 presents estimated results in two countries and analyses influential factors in the results. Finally, Section 5 makes discussion and conclusion.

Datasets

This section introduces study areas and the corresponding datasets used as input and output parameters of the designated machine learning models for the estimation of surface solar irradiation. Since satellite images have several competitive advantages, such as continuity, large-scale coverage, and publicly available, this study used a geostationary satellite called Himawari-8 to collect AOT and COT data with an hourly updated temporal resolution. As meteorological data have a strong correlation with solar irradiation [44], meteorological data, i.e., the maximum temperature, minimum temperature, average humidity, average wind speed, and average atmosphere pressure, were used as the input parameters. To obtain high precision of estimation, solar irradiation under the clear-sky condition was also calculated as an input

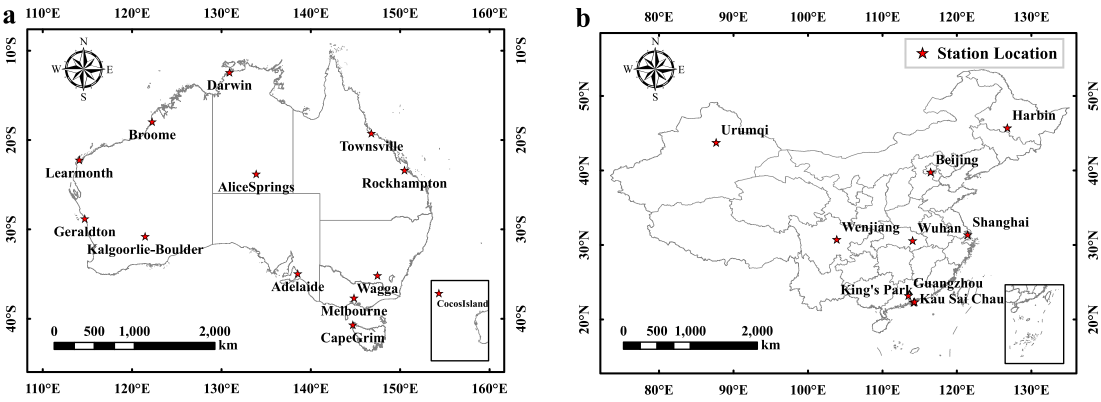

Fig. 1. Distribution of the 22 stations represented in red stars. (a) Stations in Australia. (b) Stations in China.

Table 2
The meteorological data used for the estimation of solar irradiation. AT: average temperature; P: average atmosphere pressure; WS: wind speed; SD: sunshine duration; H: humidity; MaxT: maximum temperature; MinT: minimum temperature; CC: cloud cover; WVP: water vapour pressure; ER: extraterrestrial radiation.

Reference	Parameters	Model
Dahmani et al. [57]	AT, P, WS, SD, H	MLP
Biazar et al. [58]	MaxT, WS, P, CC, H, SD	SVM
Zang et al. [59]	MinT, MaxT, WS, H, SD	BDN
Deo et al. [60]	MinT, SD, WVP, WS, P	SVM
Rabehi et al. [44]	ER, AT, MaxT, MinT, SD	MLP, BDT

parameter for training the machine learning models.

Study areas

To make a comprehensive evaluation of machine learning-based solar estimation, this study focused on two countries, i.e., Australia and China, that cover large geographical extents in the southern and northern hemispheres, respectively. Since the two countries cover a wide range of latitudes with various local climates (Table 1), it is helpful to validate the robustness and generalization of the method proposed in this study. There were 22 stations that contain the required datasets, covering six continuous years from 2015 to 2020. These stations consisted of 13 stations in Australia (Fig. 1 and 9 stations in China (Fig. 1b), in which two are in the Hong Kong Special Administrative Region (SAR), namely, King's Park station and Kau Sai Chau station. Table 1 shows the range of hourly observed solar irradiation in Australia and Hong Kong SAR and the range of daily observed solar irradiation in China.

Himawari-8 satellite products

Himawari-8 is a geostationary weather satellite operated by the Japan Meteorological Agency [45], which covers a large geographical extent in a range between 60°S – 60°N and 80°E – 160°W, including Oceania, Southeast Asia, and Western Pacific. Advanced Himawari Imager (AHI) aboard Himawari-8 provides AOT and COT data. The satellite images in the NetCDF format are freely available from the JAXA Himawari Monitor P-Tree System [46]. This study chose Himawari-8 level-2 AOT and COT data with a temporal resolution of 10 min and a spatial resolution of 5 km from 2015 to 2020. Huang et al. evaluated the Himawari-8 cloud products and suggested that the data quality has high consistency, benefiting from the active Radar-LiDAR observations [47]. In addition, Gao et al. suggested that the Himawari-8 satellite can provide reliably aerosol products for environmental research [48].

Calculated hourly clear-sky solar irradiation

Hourly clear-sky solar irradiation (CSI) in the 22 stations was calculated by using a Python online library called Pysolar [49], which was developed based on the Masters' algorithm [50] for solar irradiation calculation and an algorithm proposed by Reda and Andreas [51] for solar position calculation in its performance. The algorithm utilizes longitude, latitude, and an instant of time on a specific day to calculate the corresponding Sun's location in the sky, the solar irradiation in a clear-sky condition, and the irradiation reaching a horizontal or inclined surface on the ground [52,53]. The computed hourly clear-sky solar irradiation data set contained the same set of attributes for Australia and China, including the station name, time, and hourly clear-sky solar irradiation from 2015 to 2020. Since solar irradiation observed at Chinese stations has a daily-based temporal resolution, the estimated hourly solar irradiation at each Chinese station was further accumulated daily for keeping consistency.

Observed land surface solar irradiation

Surface solar irradiation observed by these stations was used as the ground truth to evaluate machine learning models-based estimation. Solar irradiation in Australia was measured by 13 meteorological stations (Fig. 1a and Table 1), which were operated by the Australian Government Bureau of Meteorology [54]. Notably, the original data was updated every minute, and this study rescaled the temporal resolution to hourly-based updates for the constancy of other datasets. Solar irradiation datasets in China had two independent categories, i.e., daily updated solar irradiation (Fig. 1b and Table 1), which is the highest temporal resolution that can be obtained from China National Meteorological Information Center [55] and hourly updated solar irradiation obtained from the Hong Kong Observatory [56].

Meteorological data

Table 2 suggests that meteorological data are commonly used as the input parameters to estimate solar irradiation. Therefore, we employed meteorological data as the input parameters, including the maximum temperature (MaxT), minimum temperature (MinT), average humidity (H), average wind speed (WS), and average atmosphere pressure (P). The hourly meteorological data in China and Australia are purchased from the OpenWeather website [61].

Construction of the datasets

The dataset in each station consists of meteorological data, AOT, COT, CSI, and the observed land surface solar irradiation from 2015 to 2020. The original AOT and COT data have a temporal resolution of 10

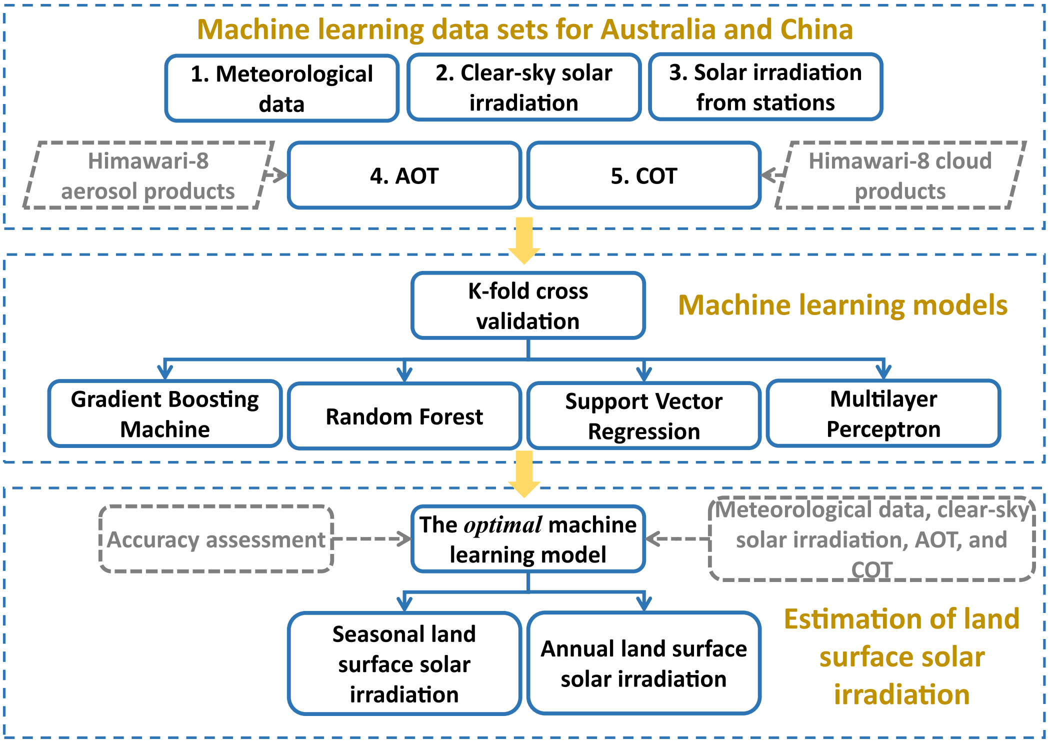
Machine learning data sets for Australia and China
1. Meteorological data
2. Clear-sky solar irradiation
3. Solar irradiation from stations
Himawari-8 aerosol products
4. AOT
5. COT
Himawari-8 cloud products
K-fold cross validation
Machine learning models
Gradient Boosting Machine
Random Forest
Support Vector Regression
Multilayer Perceptron
Accuracy assessment
The optimal machine learning model
Meteorological data, clear-sky solar irradiation, AOT, and COT
Seasonal land surface solar irradiation
Annual land surface solar irradiation
Estimation of land surface solar irradiation

Fig. 2. The framework of the proposed methodology.

min, whereas solar irradiation data is updated daily in China and hourly in Australia. To obtain the same resolution for building the machine learning models, all data in each country are aggregated to the same temporal resolution, with the lowest resolution serving as the benchmark, i.e., daily in China and hourly in Australia.

Machine learning-based estimation of solar irradiation

In this section, a machine learning framework was proposed to estimate hourly updated solar irradiation at the 22 stations (Fig. 2). There are three modules in the complete framework, i.e., machine learning datasets, machine learning models, and the estimation of the land surface solar irradiation. First of all, this study created a machine learning data set using meteorological data, AOT, COT, CSI, and the solar irradiation measured from the stations in different regions. After that, four machine learning models were used for training and prediction, namely, MLP, RF, SVR, and GBM. Finally, the paper compared the training results to determine the optimal model based on four evaluation indicators (i.e., R^2, nRMSE, nMBE, and t). When the optimal model was determined, land surface solar irradiation in Australia and China was estimated using interpolated meteorological maps, Himawari-8 cloud and aerosol products.

Data pre-processing

Pre-processing operations have been conducted to train machine learning models. First, missing values and default values of all datasets have been checked and removed. In addition, due to the inconsistency of data sources between the two countries, solar irradiation was firstly transformed to the same unit (kWh/m^2). Note that the temporal resolution of solar irradiation in mainland China was daily updated while the data in Australia was hourly updated. Finally, in this study, the datasets were divided into training datasets and validation datasets by using K-fold cross validation [62]. Specifically, the original data set was randomly divided into K equal-sized sub-datasets. Of the K sub-datasets, a single sub-dataset was employed as the validation data to test the performance of machine learning, and the remaining K-1 sub-datasets were used as the training data. In this study, we set K equalling ten.

Constructing machine-learning based estimation models

Machine learning models are expected to be able to estimate solar irradiation accurately with high computational efficiency, while estimation accuracy may be inconsistent when applying different methods in various regions. Therefore, this study adopted four different machine learning models to compare results comprehensively so that an optimal one could be identified to construct a reliable solar irradiation estimation model. The Python IDE, PyCharm [63], was employed to perform all calculations. In particular, the sklearn package [64] was used to train the four machine learning models, and the scipy package [65] was used to perform the calculation of the estimation accuracy. The GridSearchCV [66] function in the sklearn package was used to search for the optimal parameters values for four models.

Construction of the Support Vector Regression

The SVR [67] is used to perform the regression to estimate the land surface solar irradiation. In our study, the process of SVR had the following steps. Meteorological data, AOT, COT, and CSI data were selected as dependent variables for inputting the model, and the solar irradiation data measured from the observation stations were used as label variables for outputting the model. Then training function was employed to train the regression model. After that, the expected result was obtained via adjusting different kernel functions, gamma values, and the parameter of C. The dataset was organized as $\{(X_i, Y_i), i = 1, \dots, n\}$, where X_i is the vector of meteorological data, AOT, COT, and CSI data, Y_i is corresponding solar irradiation of stations, and n denotes the number of the dataset. With an SVR, a linear function is defined as:

$$f(x) = \omega \cdot x + b \quad (1)$$

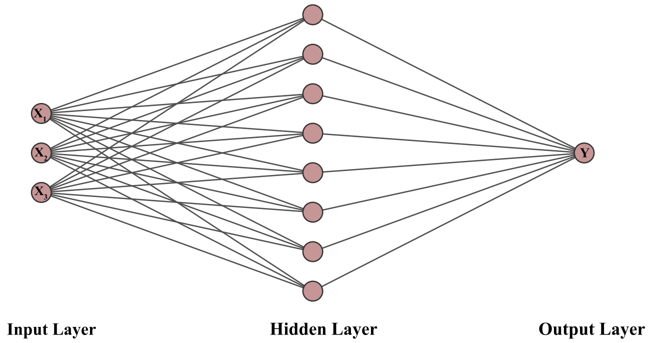

Fig. 3. The architecture of MLP.

where ω is the weight vector and b is the constant. The coefficients ω and b are estimated by the minimization process:

$$y = min\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

s.t.

$$\begin{cases} y_i - \omega\cdot x_i - b\leqslant\omega + \xi_i, & i = 1,2,\ldots,n \\ \omega\cdot x_i + b - y_i\leqslant\omega + \xi_i^*, & i = 1,2,\ldots,n \\ \xi_i, \xi_i^*\geqslant 0, & i = 1,2,\ldots,n \end{cases} \tag{3}$$

where ξ and C are the prescribed parameters, and ξ_i and ξ_i^* are positive slack variables. of the support vector regression-Kuhn-Tucher (KKT) optimizing conditions are applied in the linear regression function as presented below:

$$f(x) = \sum_{i\in SV_s}(a_i - a_i^*)(x_i, x) + b \tag{4}$$

where a_i and a_i^* are Lagrangian multipliers.

Construction of the Random Forest

Random forest [68] is a flexible and easy ensemble learning method, which can usually obtain robust results for classification and regression tasks. Therefore, RF was employed to estimate the land surface solar irradiation. In this study, the input dataset was $\{X_i, i = 1, \ldots, m\}$ and the output dataset was $\{Y_i, i = 1, \ldots, m\}$, where X_i denotes the vector of meteorological data, AOT, COT and CSI data, Y_i is the solar irradiation of stations, and m denotes the number of datasets. On this basis, this study performed the RF regression model with the following three steps.

1. Bootstrap sample method was employed to generate a training dataset by randomly drawing with replacement m samples, where m is the size of the original training dataset.
2. A multitude of decision trees was constructed at training time and outputting the class that is the mode of mean prediction of the individual trees.
3. After repeating step (2) for n times, we can obtain a number of n regression trees to generate the random forest. For any regression tree, the mean error of all the regression trees can be calculated for obtaining an unbiased estimation of the random forest. The calculation formula is as follows:

$$Y(x_i) = \frac{1}{n}\sum_{i=1}^{n}T_n(X_i), n = 1,2,\ldots,n \tag{5}$$

where T_n denotes a regression tree, and n is the number of regression trees.

Construction of the multilayer perceptron

Artificial Neural Networks (ANNs) are computing systems inspired by biological neural networks, which can learn from data relationships and generalize the laws of data to predict data development trends. As one of the most popular structures of ANNs, MLP [69] consists of three layers, i.e., an input layer containing the structured meteorological data, CSI, AOT, and COT, a hidden layer achieved by a Sigmoid function as the activation function, and an output layer providing estimated surface solar irradiation (Fig. 3). This study used MLP with a Back Propagation (BP) algorithm for training, which contains forward data flow calculation and backward error propagation. Particularly, the input layer received solar irradiation datasets, the hidden layer transmits and adjusts network weights for the regression model, and finally, the output layer stored the estimated irradiation data. If the result obtained from the output layer was not consistent with the ground truth, then weight adjustment would be conducted based on an error function achieved by a backward propagation algorithm. The neural network would be optimized continuously by repeating the adjustment of the weight parameters until the error was lower than the established standard.

Construction of the gradient boosting machine

In this study, GBM [70] was used to estimate the land surface solar irradiation, which is one class of the Boosting algorithm for producing regression models. It is achieved by establishing an additive model that adds a new decision tree in each iterative step, leading to the minimized deviation in the loss function. The GBM model was performed as follows:

1. Given a training dataset $\{(x_i, y_i), i = 1, \ldots, n\}$ and the loss function $L(y, F(x))$, where x_i was the vector of meteorological data, AOT, COT, and CSI data, y_i is corresponding solar irradiation of stations, and n denotes the number of datasets. The model was initialized using the fixed value γ:

$$F_0(x) = \underset{\gamma}{\text{argmin}}\sum_{i=1}^{n}L(y_i, \gamma) \tag{6}$$

2. Calculation pseudo-residuals r_{im}, the formula is as follows:

$$r_{im} = [\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}, (i = 1, 2, 3, \ldots., n) \tag{7}$$

3. Calculation γ_m to solve the optimization problem:

$$\gamma_m = \underset{\gamma}{\text{argmin}}\sum_{i=1}^{n}L(y_i, F_{m-1}(x_i + \gamma h_m(x_i))) \tag{8}$$

where $h_m(x)$ denotes pseudo-residuals for the decision tree, the formula is as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{9}$$

Estimation surface solar irradiation based on the optimal model

This study employed four evaluation indicators to evaluate the estimation accuracy of each model, namely a coefficient of determination (R^2), normalized Root Mean Square Error (nRMSE), normalized mean bias error (nMBE), and consumption of time (t). Specifically, nRMSE and nMBE were calculated as follows:

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}}{\frac{1}{n}\sum_{i=1}^{n}y_i} \tag{10}$$

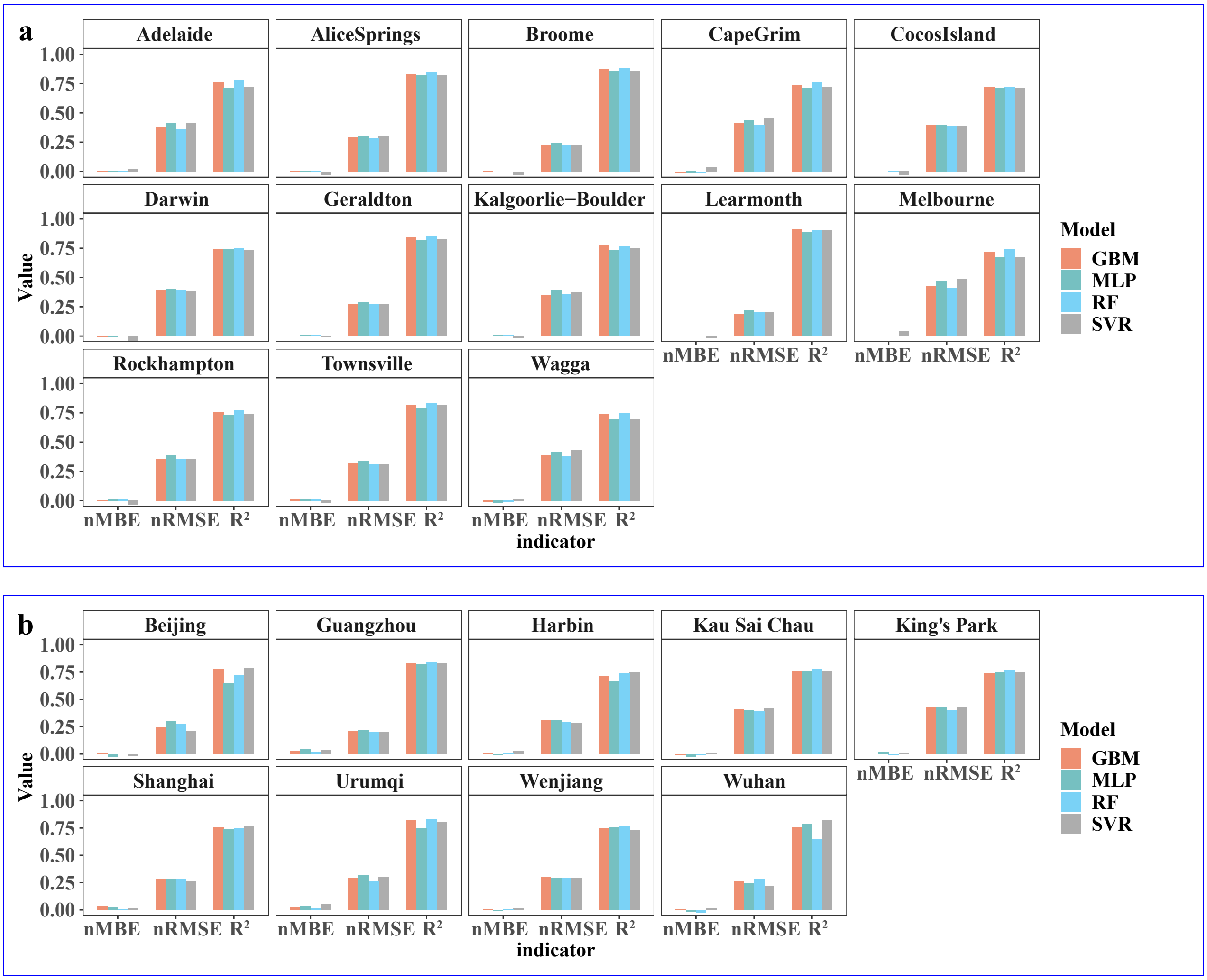

Fig. 4. Estimation accuracy of the four machine learning models using R^2, nRMSE, and nMBE in all stations. (a) Results in Australia. (b) Results in China.

$$nMBE = \frac{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)}{\frac{1}{n}\sum_{i=1}^{n}y_i} \tag{11}$$

where n is the number of data, $\widehat{y}_i$ denotes estimation value, and y_i is actual value.

Results

The four machine learning models were used to evaluate the estimation accuracy at each station independently based on the four evaluation indicators. Through comprehensive comparison, the optimal machine learning model was selected for estimating surface solar irradiation in Australia and China.

Accuracy assessment of the models

Fig. 4 systematically compares the estimated accuracy based on R^2, nRMSE, and nMBE in all the 22 stations. Overall, it is found that the four models have similar estimation performance. Specifically, all stations have $R^2 \geqslant 0.7$ in both countries using the GBM model, and the proportions of the stations are about 38% for Australia and about 22% for China when $R^2 \geqslant 0.8$. Besides, the nMBE values are significantly low in all stations, and the nRMSE values are between 0.2 and 0.4 only. The results suggest that the estimation models are reliable with high estimation accuracy, which indicates that the proposed method can effectively estimate land surface solar irradiation over large regions. From the other perspective, Fig. 5 summarizes the computation time of the four machine learning models in each station, which presents that the GBM model achieves the shortest time consumption. This suggests that GBM is outperformed for the estimation accuracy and computational efficiency, especially for extensive computation when there are a large number of stations confined in a small area.

Feature importance analysis for the input parameters

Furthermore, the feature importance analysis is conducted to evaluate the impacts of each parameter on the estimation models (Fig. 6). It shows that CSI is significantly larger than the second most important feature of H for estimating the solar irradiation in Australia, leaving the rest features almost ignorable. This indicates that Australia has stable and solar favourable meteorological conditions, which thus have weak impacts on the solar estimation. In contrast, the top three impact features are H, CSI, and MaxT in China, suggesting that the land surface solar irradiation is comprehensively affected by the meteorological features.

Generation of the land surface solar irradiation

To create seasonal and annual land surface solar irradiation maps at

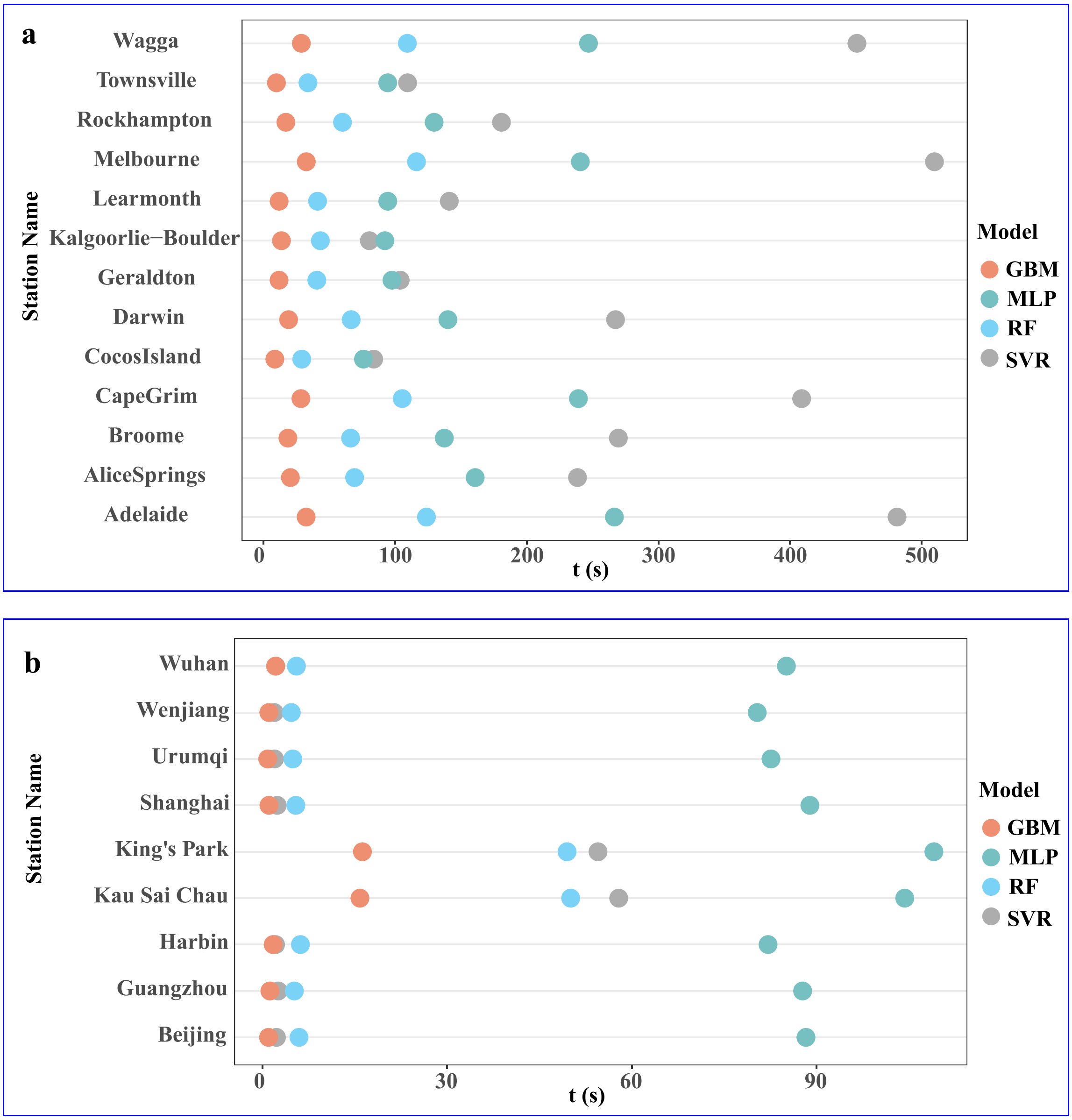

Fig. 5. Computation time of the four machine learning in all stations. (a) Results in Australia. (b) Results in China.

a 5-km spatial resolution in the two countries in 2020, the GBM model is used because it has achieved the highest estimation accuracy in both countries. The meteorological, COT, AOT, and CSI images are well prepared and used as the input parameters of the trained model. In addition, a set of meteorological images are obtained by using the Kriging interpolation method. Since the trained GBM model based on each observation station has relatively high accuracy as presented in Fig. 4, this study used all the trained models to create the solar irradiation maps over the whole territory of Australia and China.

To systematically evaluate the accuracy of each created solar irradiation map, this study investigated the relative errors between the estimated values and correspondingly measured values located at all the stations in each solar irradiation map. Fig. 7 shows that the relative errors in all stations are between 0.1 and 0.2, which suggests that the estimation results in all stations are accurate. Therefore, the mean values of all estimation maps were calculated and used as the final estimated solar irradiation map in the two countries. To avoid extremely big data computation, the solar irradiation on the middle day of each month is considered as the daily mean irradiation of that month, so that the monthly, seasonal, and annual solar irradiation can be accumulated over the corresponding time interval in each country.

Maximum and minimum monthly land surface solar irradiation

Fig. 8 and Fig. 9 show the maximum and minimum horizontal surface global solar irradiation in Australia and China, respectively. Overall, the solar distribution in January is significantly higher than that in August in Australia, whereas the maximum solar distribution is in August and the minimum solar distribution is in January in China. In Australia, solar irradiation gradually increases from the northwest region to the southeast region in August (Fig. 8a), with monthly values ranging from 171.78 to 76.08 kWh/m^2, while the irradiation in the central region is lower than in the other regions in January, (Fig. 8b), with monthly values ranging from 200.18 to 95.12 kWh/m^2. In China, solar irradiation in the southeast and central regions is lower than in other regions in January (from 69.88–147.98 kWh/m^2). In contrast, the irradiation is overall high in the whole country in August, with only part of the central region relatively low (from 97.56–223.89 kWh/m^2).

Seasonal land surface solar irradiation

Furthermore, seasonal land surface solar irradiation maps were

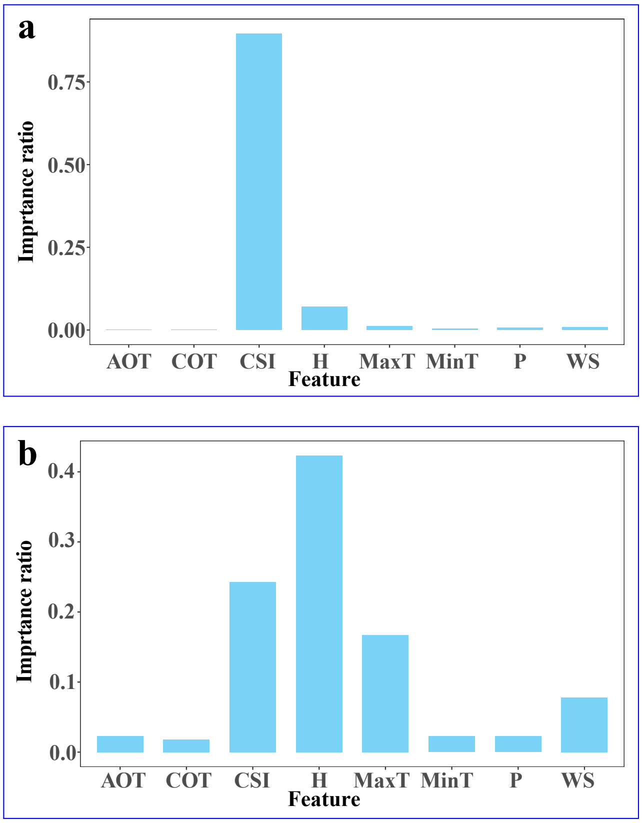

Fig. 6. Importance ratios for the input features. (a) Australia. (b) China.

created for Australia (Fig. 10) and China (Fig. 11). Overall, the highest solar irradiation values are in summer in the two countries, followed by those in spring, autumn, and winter.The solar irradiation values in all seasons in Australia exhibit the narrow distribution, whereas those in China give the wide distribution. Fig. 10 shows that Australia has an insignificant change in solar distribution during the four seasons, and most areas in Australia have a large amount of solar energy near 632 kWh/m^2. In China, solar irradiation in western and northeastern regions maintains a high level near 535 kWh/m^2 all year round, whereas, for southeastern regions in spring and summer, it is higher than that in autumn and winter.

Annual land surface solar irradiation

Lastly, the annual land surface solar irradiation was estimated by accumulating four seasonal solar energy. Overall, the total irradiation in Australia (Fig. 12a) is higher than that in China (Fig. 12b). In detail, the vast majority of areas in Australia have abundant solar resources, suggesting that Australia is feasible to promote solar energy in most areas. In comparison, the distribution of the annual irradiation in China presents a gradual decrease from the northeast to the southwest. This indicates that southwest China has a relatively thick cloud cover that hinders the receiving of solar energy, meaning that latitude may not be a conclusive factor for using solar energy in large regions. In addition, heterogeneous distribution of solar energy is apparent in central China, which indicates that our model is also sensitive to depicting regional differences in solar distribution. It is found that our results are consistent with the published maps created by Solargis [71,72] when comparing the quantitative ranges and the distribution patterns of the solar irradiation maps.

Discussion and conclusion

This study developed a method by integrating the machine learning models and remote sensing technologies to estimate land surface solar irradiation at fine temporal resolutions (i.e., hourly to daily) over large geographical areas. Even though the study areas of Australia and China are two big countries that contain a variety of climate zones, the trained models based on only a few stations still achieved high prediction accuracy with $R^2 > 0.7$ for all the stations. By comparing the generated maps with the published maps in terms of the spatio-temporal distributions and the quantitative ranges, it is found that our results are broadly in line with the published maps. This suggests that the established models are accurate and reliable, and the proposed method can be used to estimate land surface solar irradiation in large-scale regions. In addition, the high availability of Himawari-8 satellite products with free licensed characteristics makes it possible to be widely used for an accurate estimation of solar irradiation over large regions, which is especially important for nations that aim to promote using solar energy.

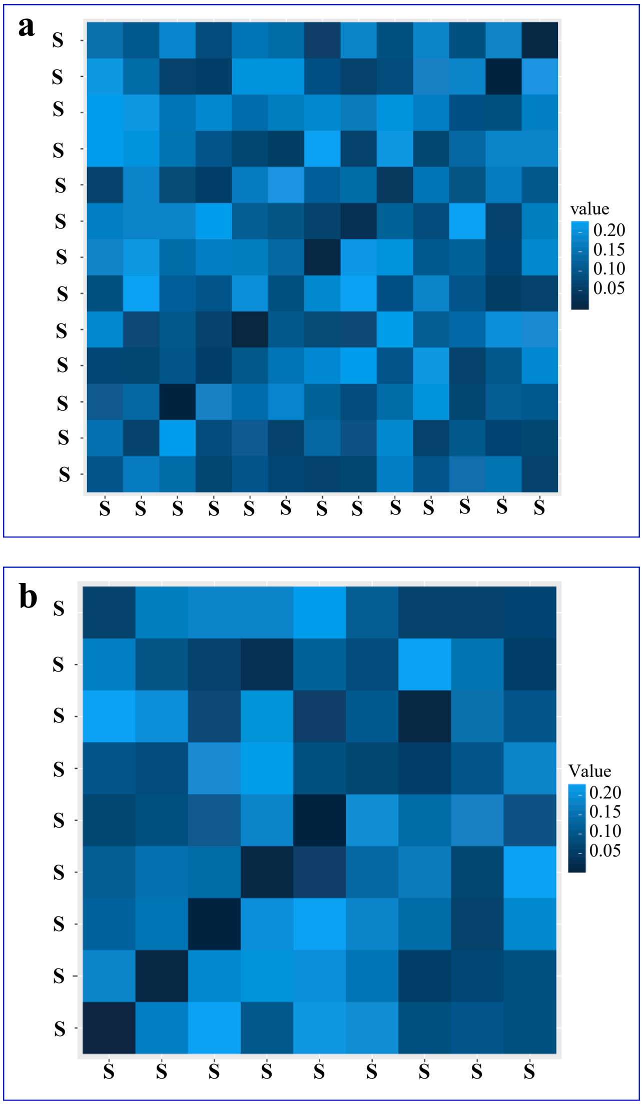

Fig. 7. The relative errors between the estimated values and measured values in each station for each solar irradiation map. (a) There are 13 stations in Australia that correspond to a 13 × 13 matrix. (b) There are 9 stations in China that correspond to a 9 × 9 matrix.

This study used 22 datasets to train the machine learning models independently, which thus created a well-trained model for each of the 22 solar observation stations. As all the trained models obtained high estimation accuracy, all the models were used to create solar irradiation maps to make full use of the currently available datasets. However, as

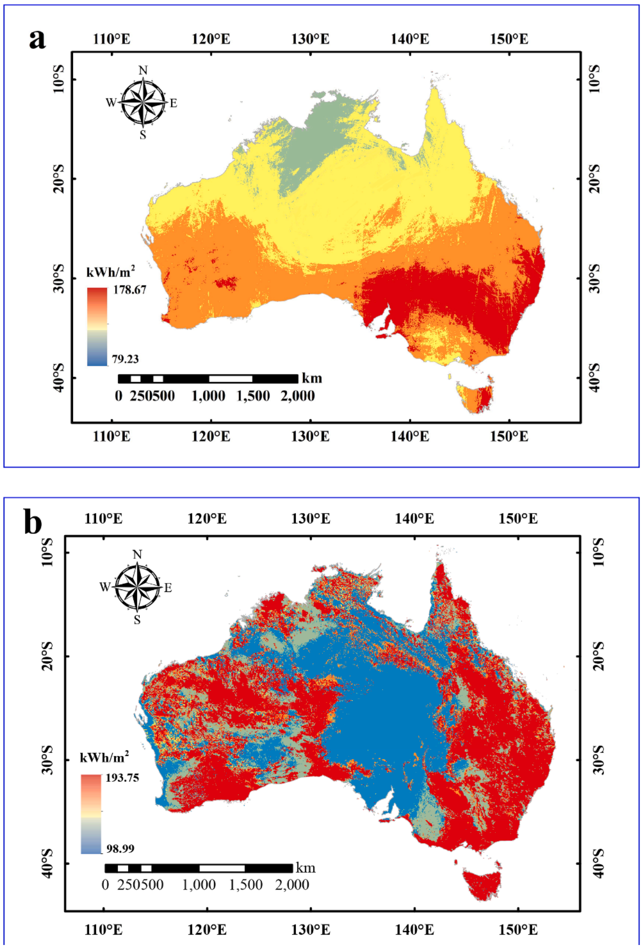

Fig. 8. Distribution of the maximum and minimum horizontal surface global solar irradiation in Australia. (a) Distribution in August. (b) Distribution in January.

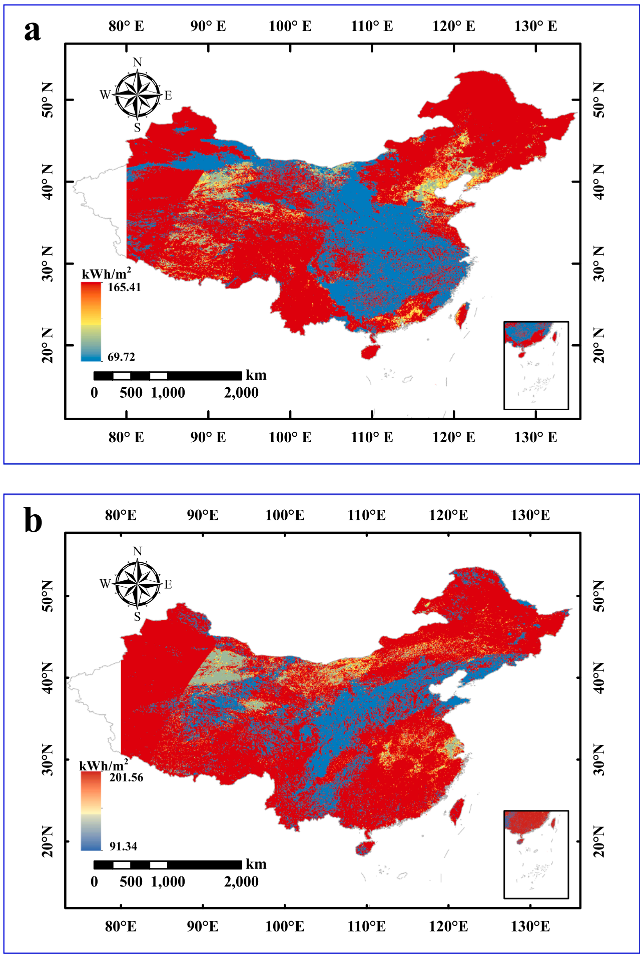

Fig. 9. Distribution of the maximum and minimum horizontal surface global solar irradiation in China. (a) Distribution in January. (b) Distribution in August.

the solar observation stations have sparse distribution in each country, it is difficult to validate the prediction accuracy of each pixel value in the finally created solar irradiation maps. Alternatively, the observed solar irradiation data with determined geo-locations can be used as real samples to systematically investigate the final prediction accuracy.

The Kriging interpolation method was used to generate the spatially continuous meteorological images, which were used as the input parameters for estimating solar irradiation. Although the analysis shows that the overall interpolation accuracy is significantly high, it is hard to make sure that the whole areas maintain the same high accuracy. Nevertheless, the comparison of the published maps and the relative error matrices help confirm that this method is feasible and the results are reliable. Meanwhile, this study conducted the importance-analysis for the input parameters and it was found that the impacts of these parameters on solar estimation are different between the two countries. While in the same country, the impacts of the parameters are consistent for different models. This implies the effectiveness of the selected parameters for the solar estimation. It is worth mentioning that meteorological conditions can affect land surface solar irradiation to some extent, in which the humidity makes a great contribution.

The average values of a set of the estimated solar irradiation maps in the same spatial and temporal domains are used to create the final solar irradiation map because of two reasons. First, the estimation accuracies (R^2) of all the models are basically consistent in a small range between 0.7 and 0.9. Second, the relative error matrices (Fig. 7) between the estimated values and measured values are between 0.1 and 0.2 only. This demonstrates that the difference between each estimation solar irradiation map is rather small. Therefore, the estimated solar irradiation maps can make an equal contribution to create the final solar map.

To conclude, this study proposes a simple and effective method for the estimation of land surface solar irradiation based on machine learning models using meteorological data, Himawari-8 satellite cloud and aerosol products, and solar observation data in Australia and China. The estimation of solar irradiation based on four machine learning models, i.e., RF, SVR, MLP, and GBM, is effective and reliable, and GBM has achieved the best performance in terms of accuracy and computational efficiency. The estimation of seasonal and annual solar irradiation at nationwide levels is useful for planning solar-related applications.

CRediT authorship contribution statement

Xuan Liao: Conceptualization, Methodology, Visualization, Validation, Software, Writing - original draft. **Rui Zhu:** Conceptualization, Visualization, Investigation, Writing - review & editing, Supervision. **Man Sing Wong:** Conceptualization, Investigation, Writing - review & editing, Supervision.

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.

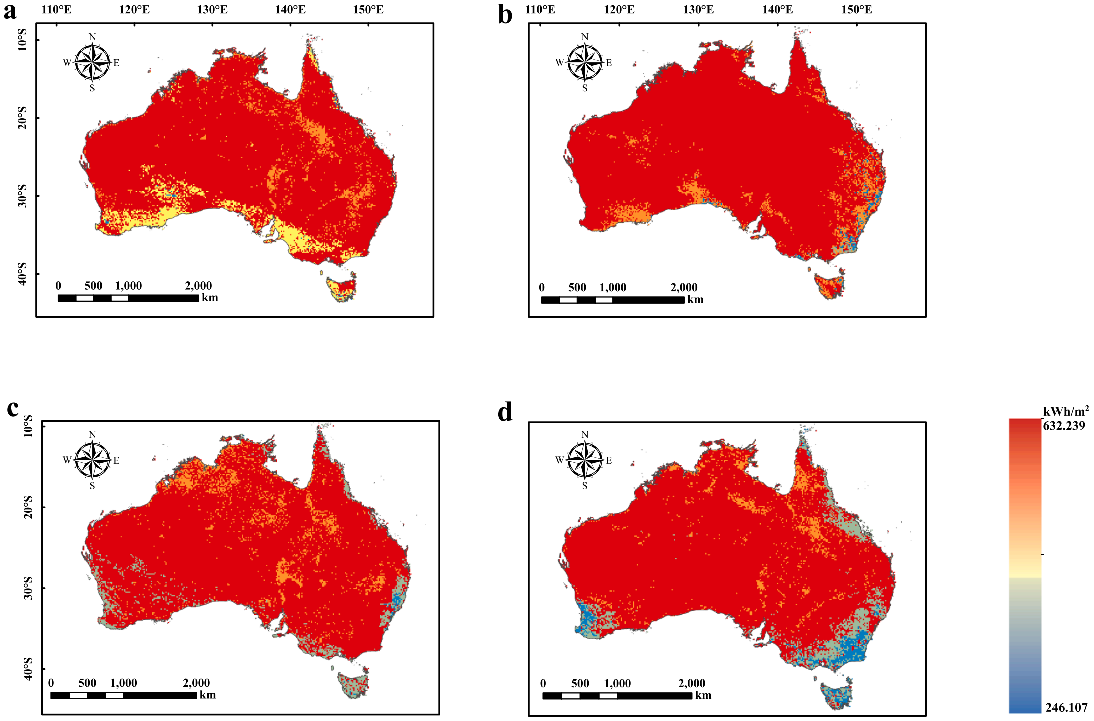

Fig. 10. Seasonal distribution of land horizontal surface global solar irradiation in Australia. (a) The irradiation in spring (September to November). (b) The irradiation in summer (December to February). (c) The irradiation in autumn (March to May). (d) The irradiation in winter (June to August).

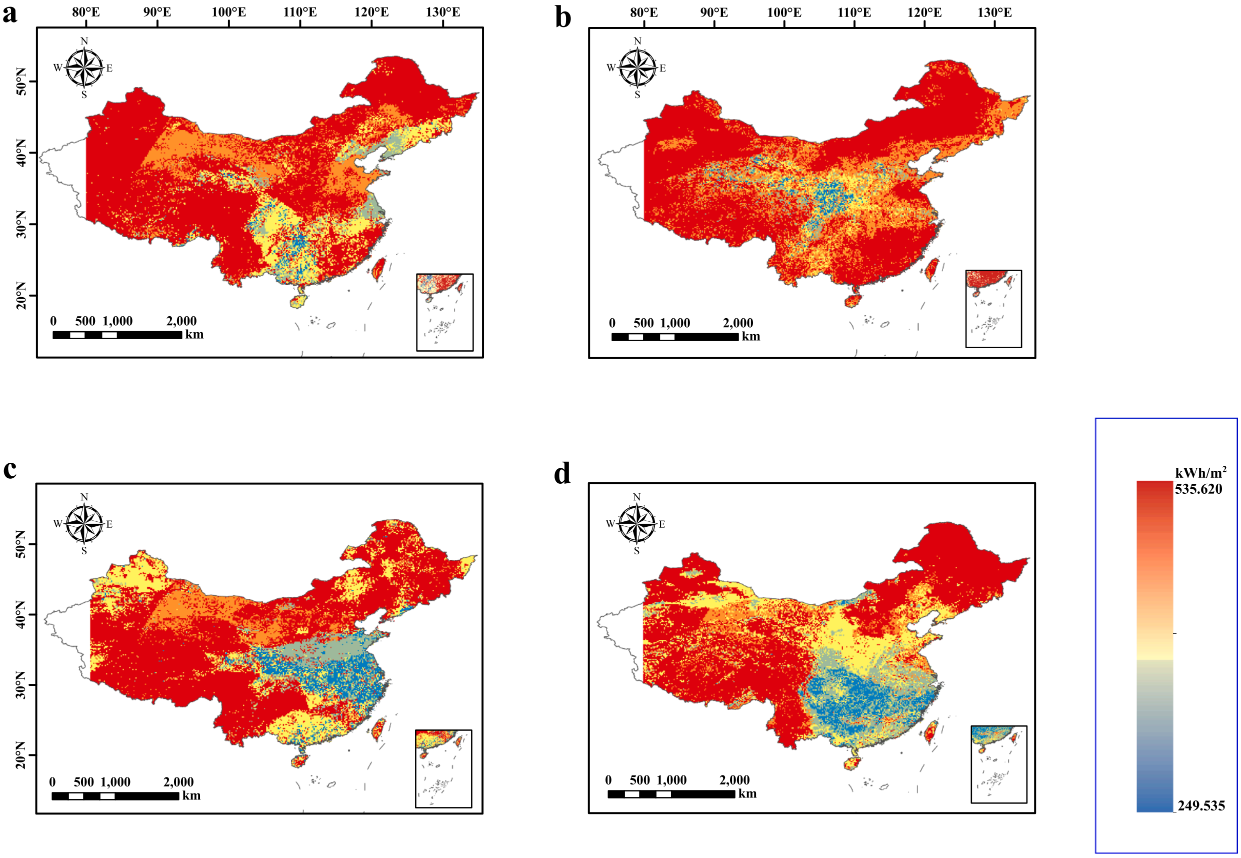

Fig. 11. Seasonal distribution of land horizontal surface global solar irradiation in China. (a) The irradiation in Spring (March to May). (b) The irradiation in Summer (June to August). (c) The irradiation in Autumn (September to November). (d) The irradiation in Winter (December to February).

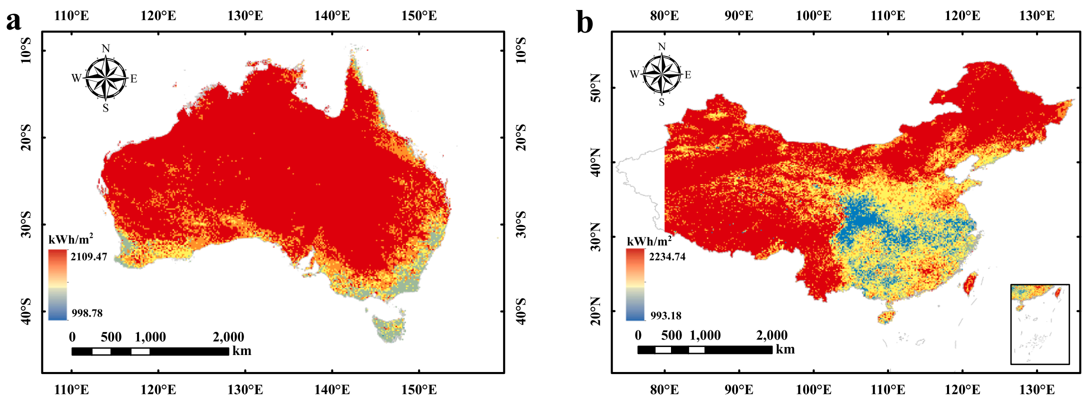

Fig. 12. Annual horizontal surface global solar irradiation in the two countries. (a) Distribution of annual irradiation in Australia. (b) Distribution of annual irradiation in China.

Acknowledgements

Authors thank the funding support from the General Research Fund (Grant No. 15602619), and the Strategic Hiring Scheme (Grant No. P0036221) at the Hong Kong Polytechnic University.

References

[1] Outlook, Energy. International Energy Outlook. Outlook; 2010.

[2] Erickson P, van Asselt H, Koplow D, Lazarus M, Newell P, Oreskes N, Supran G. Why fossil fuel producer subsidies matter. Nature 2020;578:E1–4.

[3] Kammen DM, Sunter DA. City-integrated renewable energy for urban sustainability. Science 2016;352:922–8.

[4] Kumari P, Toshniwal D. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. J Clean Prod 2021: 279:123285.

[5] Kannan N, Vakeesan D. Solar energy for future world:-A review. Renewable Sustain Energy Rev 2016;62:1092–105.

[6] Tariq GH, Ashraf M, Hasnain US. Solar Technology in Agriculture. Technology in Agriculture; 2021.

[7] Wong MS, Zhu R, Liu ZZ, Lu L, Peng JQ, Tang ZQ, Lo CH, Chan WK. Estimation of Hong Kong's solar energy potential using GIS and remote sensing technologies. Renewable Energy 2016;99:325–35.

[8] Zhang J, Zhao L, Deng S, Xu W, Zhang Y. A critical review of the models used to estimate solar radiation. Renewable Sustain Energy Rev 2017;70:314–29.

[9] Benatiallah D, Bouchouicha K, Benatiallah A, Harrouz A, Nasri B. Forecasting of solar radiation using an empirical model. Algerian J Renewable Energy Sustain Devel 2019;1:212–9.

[10] Bailek N, Bouchouicha K, Al-Mostafa Z, El-Shimy M, Aoun N, Slimani A, Al-Shehri S. A new empirical model for forecasting the diffuse solar radiation over Sahara in the Algerian Big South. Renewable Energy 2018;117:530–7.

[11] Makade RG, Chakrabarti S, Jamil B. Prediction of global solar radiation using a single empirical model for diversified locations across India. Urban Climate 2019; 29(100492).

[12] Zhu R, Wong MS, You LL, Santi P, Nichol J, Ho HC, Lu L, Ratti C. The effect of urban morphology on the solar capacity of three-dimensional cities. Renewable Energy 2020;153:1111–26.

[13] Cogliani E, Ricchiazzi P, Maccari A. Physical model SOLARMET for determinating total and direct solar radiation by meteosat satellite images. Sol Energy 2007;81: 791–8.

[14] Ceballos JC, Bottino MJ, De Souza JM. A simplified physical model for assessing solar radiation over Brazil using GOES 8 visible imagery. J Geophys Res: Atmos 2004;109(D2).

[15] Yeom JM, Seo YK, Kim DS, Han KS. Solar radiation received by slopes using COMS imagery, a physically based radiation model, and GLOBE. J Sensors 2016, 2016,: 1–15.

[16] Zhou Y, Liu Y, Wang D, Liu X, Wang Y. A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Convers Manage 2021;235(113960).

[17] Voyant C, Notton G, Kalogirou S, Nivet M, Paoli C, Motte F, Fouilloy A. Machine learning methods for solar radiation forecasting: A review. Renewable Energy 2017;105:569–82.

[18] Guermoui M, Melgani F, Gairaa K, Mekhalfi ML. A comprehensive review of hybrid models for solar radiation forecasting. J Clean Prod 2020;258(120357).

[19] Besharat F, Dehghan AA, Faghih AR. Empirical models for estimating global solar radiation: A review and case study. Renewable Sustain Energy Rev 2013;21: 798–821.

[20] Swartman RK, Ogunlade O. Solar radiation estimates from common parameters. Solar Energy 1967;11:170–2.

[21] De SJL, Lyra GB, Dos SCM, Junior RAF, Tiba C, Lyra GB, Lemes MAM. Empirical models of daily and monthly global solar irradiation using sunshine duration for Alagoas State, Northeastern Brazil. Sustain Energy Technol Assess 2016;14:35–45.

[22] Allen RG. Self-calibrating method for estimating solar radiation from air temperature. J Hydrol Eng 1997;2:56–67.

[23] Nikitidou E, Zagouras A, Salamalikis V, Kazantzidis A. Short-term cloudiness forecasting for solar energy purposes in Greece, based on satellite-derived information. Meteorol Atmospheric Phys 2019;131:175–82.

[24] Maxwell EL. METSTAT–The solar radiation model used in the production of the National Solar Radiation Data Base (NSRDB). Sol Energy 1998;62:263–79.

[25] Bird RE. A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. Solar Energy 1984;32:461–71.

[26] Yang K, Huang GW, Tamai N. A hybrid model for estimating global solar radiation. Solar Energy 2001;70:13–22.

[27] Page JK. Proposed quality control procedures for the meteorological office data tapes relating to global solar radiation, diffuse solar radiation, sunshine and cloud in the UK. Report FCIBSE 1997.

[28] Chen JL, Xiao BB, Chen CD, Wen ZF, Jiang Y, Lv MQ, Wu SJ, Li GS. Estimation of monthly-mean global solar radiation using MODIS atmospheric product over China. J Atmospheric Solar-Terrestrial Phys 2014;110:63–80.

[29] Zhang YL, Li X, Bai YL. An integrated approach to estimate shortwave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. Sol Energy 2015;113:347–57.

[30] Feng F, Wang KC. Merging ground-based sunshine duration observations with satellite cloud and aerosol retrievals to produce high-resolution long-term surface solar radiation over China. Earth Syst Sci Data 2021;13:907–922.

[31] Khosravi A, Koury R, Machado L, Pabon J. Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. J Clean Prod 2018;176: 63–75.

[32] Wang J, Jiang H, Wu Y, Dong Y. Forecasting solar radiation using an optimized hybrid model by Cuckoo Search algorithm. Energy Convers Manage 2015;81: 627–44.

[33] Behrang M, Assareh E, Ghanbarzadeh A, Noghrehabadi A. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Sol Energy 2010;84:1468–80.

[34] Sun S, Wang S, Zhang G, Zheng J. A decomposition-clustering-ensemble learning approach for solar radiation forecasting. Sol Energy 2018;163:189–99.

[35] Rohani A, Taki M, Abdollahpour M. A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I). Renewable Energy 2018;115:411–22.

[36] Prada J, Dorronsoro JR. General noise support vector regression with non-constant uncertainty intervals for solar radiation prediction. J Modern Power Syst Clean Energy 2018;6:268–80.

[37] Sharafati A, Khosravi K, Khosravinia P, Ahmed K, Salman SA, Yaseen ZM, Shahid S. The potential of novel data mining models for global solar radiation prediction. Int J Environ Sci Technol 2019;16:7147–64.

[38] 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.

[39] Yagli GM, Yang D, Srinivasan D. Automatic hourly solar forecasting using machine learning models. Renewable Sustain Energy Rev 2019;105:487–98.

[40] Ramedani Z, Omid M, Keyhani A, Khoshnevisan B, Saboohi H. A comparative study between fuzzy linear regression and support vector regression for global solar radiation prediction in Iran. Sol Energy 2014;109:135–43.

[41] Srivastava R, Tiwari A, Giri V. Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. Heliyon 2019;5(e02692).

[42] Rabehi A, Guermoui M, Lalmi D. Hybrid models for global solar radiation prediction: a case study. Int J Ambient Energy 2020;41:31–40.

[43] Urraca R, Antoñanzas J, Antoñanzas-Torres F, Martinez-de-Pison FJ. Estimation of daily global horizontal irradiation using extreme gradient boosting machines. International Joint Conference SOCO'16-CISIS'16-ICEUTE'16 2016:105–13.

[44] Rabehi A, Guermoui M, Lalmi D. Hybrid models for global solar radiation prediction: a case study. Int J Ambient Energy 2020;41:31–40.

[45] Japan Meteorological Agency; 2021. url: https://www.jma.go.jp/jma/indexe.html [Accessed 27 July 2021].

[46] JAXA Himawari Monitor P-Tree System; 2021. url: https://www.eorc.jaxa.jp/ptree/ [Accessed 27 July 2021].

[47] Huang Y, Siems S, Manton M, Protat A, Majewski L, Nguyen H. Evaluating Himawari-8 cloud products using shipborne and CALIPSO observations: Cloud-top height and cloud-top temperature. J Atmospheric Oceanic Technol 2019;36:2327–47.

[48] Gao L, Chen L, Li CC, Li J, Che HZ, Zhang YP. Evaluation and possible uncertainty source analysis of JAXA Himawari-8 aerosol optical depth product over China. Atmos Res 2021;248(105248).

[49] Pysolar; 2021. url: https://pysolar.readthedocs.io/en/latest/ [Accessed 27 July 2021].

[50] Masters GM. Renewable and efficient electric power systems. John Wiley & Sons; 2013.

[51] Reda I, Andreas A. Solar position algorithm for solar radiation applications. Solar Energy 2004;76:577-589.

[52] Gilbert MM. Renewable and efficient electric power systems. John Wiley & Sons; 2004.

[53] Bishop JK, Rossow WB, Dutton EG. Surface solar irradiance from the international satellite cloud climatology project 1983–1991. J Geophys Res: Atmospheres 1997;102:6883–910.

[54] Australian Government Bureau of Meteorology; 2021. url: http://reg.bom.gov.au/index.php/ [Accessed 27 July 2021].

[55] China National Meteorological Information Center; 2021. url: http://data.cma.cn/ [Accessed 27 July 2021].

[56] Hong Kong Observation; 2021. url: https://www.hko.gov.hk/tc/ [Accessed 27 July 2021].

[57] Dahmani K, Notton G, Voyant C, Dizene R, Nivet ML, Paoli C, Tamas W. Multilayer Perceptron approach for estimating 5-min and hourly horizontal global irradiation from exogenous meteorological data in locations without solar measurements. Renewable Energy 2016;90:267–82.

[58] Biazar SM, Rahmani V, Isazadeh M, Kisi O, Dinpashoh Y. New input selection procedure for machine learning methods in estimating daily global solar radiation. Arab J Geosci 2020;13:1–17.

[59] Zang H, Cheng L, Ding T, Cheung KW, Wang M, Wei Z, Sun G. Application of functional deep belief network for estimating daily global solar radiation: A case study in China. Energy Convers Manage 2020;191(116502).

[60] Deo RC, Wen X, Qi F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl Energy 2016;168:568–93.

[61] OpenWeahter; 2022. url: https://openweathermap.org/ [Accessed 14 February 2022].

[62] Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 2009;32:569–575.

[63] PyCharm; 2022. url: https://www.jetbrains.com/pycharm/ [Accessed 14 February 2022].

[64] Sklearn; 2022. url: https://scikit-learn.org/stable/ [Accessed 14 February 2022].

[65] SciPy; 2022. url: https://scipy.org/ [Accessed 14 February 2022].

[66] GridSearchCV; 2022. url: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Accessed 14 February 2022].

[67] Awad M, Khanna R. Support vector regression. Efficient learning machines 2015;67–80.

[68] Segal MR. Machine learning benchmarks and random forest regression. In: UCSF: Center for Bioinformatics and Molecular Biostatistics; 2004. p. 1–14.

[69] Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing 1991;2:183–97.

[70] Friedman JH. Greedy function approximation: a gradient boosting machine. In: Ann Stat; 2001. p. 1189–232.

[71] Global Solar Atlas of China; 2022. url: https://solargis.com/maps-and-gis-data/download/china [Accessed 14 February 2022].

[72] Global Solar Atlas of Australia; 2022. url: https://solargis.com/maps-and-gis-data/download/australia [Accessed 14 February 2022].