Applied Energy 339 (2023) 121005

Contents lists available at ScienceDirect

Applied Energy

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

Assessment of solar radiation resource and photovoltaic power potential across China based on optimized interpretable machine learning model and GIS-based approaches

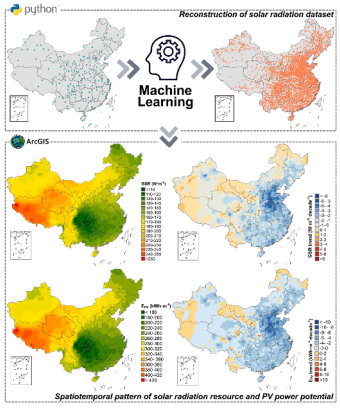
Zhe Song [a,*], Sunliang Cao [a,b,c], Hongxing Yang [a,*]

[a] Renewable Energy Research Group (RERG), Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Research Institute for Sustainable Urban Development (RISUD), The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[c] Research Institute for Smart Energy (RISE), The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

HIGHLIGHTS

- Long-term solar radiation datasets were reconstructed across China.
- Global solar radiation in summer decreased by up to 1.83 $W \cdot m^{-2} \cdot decade^{-1.}$
- China's PV power potential decreased by 1.69 $kWh \cdot m^{-2} \cdot decade^{-1}$ from 1961 to 2016.
- 30 provinces saw a 0.25–10.27% reduction in PV potential in the 2010s versus the 1960s.
- China's PV sector showed a regional mismatch between PV potential and installed capacity.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Solar radiation resource
Solar photovoltaic power potential
Machine learning
Geographic information system
Spatiotemporal pattern

ABSTRACT

In light of the rapidly expanding solar photovoltaic (PV) sector, it is important to provide a deeper understanding of solar energy resources to successfully implement solar energy projects. In this study, an interpretable machine learning model based on extreme gradient boosting (XGBoost) optimized by particle swarm optimization (PSO) algorithms was developed to estimate global solar radiation. The results show that the proposed PSO-XGBoost model possesses the most superior accuracy and stability, with the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) of 0.953, 1.597 $MJ \cdot m^{-2} \cdot day^{-1}$, 1.138 $MJ \cdot m^{-2} \cdot day^{-1}$, and 10.500%, respectively. With the geographic information system (GIS) -based approaches, a 50 km by 50 km spatial resolution map of long-term national average solar radiation resources was generated based on the reconstructed solar radiation dataset, as well as the PV power potential map. The findings reveal that the nationwide annual mean solar radiation resources were decreasing at an estimated attenuation of −0.83 $W \cdot m^{-2} \cdot decade^{-1}$, with a downward trend of the greatest magnitude of −1.83

* Corresponding authors.
E-mail addresses: zhe9501.song@connect.polyu.hk (Z. Song), hong-xing.yang@polyu.edu.hk (H. Yang).

https://doi.org/10.1016/j.apenergy.2023.121005
Received 9 January 2023; Received in revised form 20 February 2023; Accepted 19 March 2023
Available online 28 March 2023

W·m^{-2}·decade^{-1} for summer. China's long-term average yearly PV power potential reached 285.00 kWh·m^{-2}, indicating a spatial pattern of higher potentials in the northwestern and northern provinces, while lower values in the southeastern provinces. Moreover, the PV power potential in China decreased by 1.69 kWh·m^{-2}·decade^{-1} from 1961 to 2016, with an attenuation of above 5 kWh·m^{-2}·decade^{-1} in heavily polluted regions. During the 2010s, 30 out of the 31 provinces experienced a reduction in the PV power potential between 0.25% and 10.27%, with an average national reduction of 2.88%, compared to the 1960s scenario. Also, policy recommendations for long-term PV project deployment were given regarding the regional mismatch between PV power potential and installed capacity in China.

1. Introduction

Since the 2015 Paris Agreement on climate change and the IPCC special report on global warming of 1.5 °C, there has been a global goal to drive the transition in energy markets from fossil fuel dominance to clean energy dominance [1,2]. Indeed, the use of renewable energy has increased globally over the past decade and is expected to play a critical role in achieving greenhouse gas emission reduction targets to combat climate change. In 2021, global renewable power capacity increased by more than 314 GW, with renewable energy accounting for approximately 28.3% of the global electricity mix [3]. Renewable energy led the investment in the global power sector, comprising about 75% of 2021's total of more than $600 billion invested in global new power generation capacity [4]. Moreover, projections show that more than 52% of global power generation will be met by renewable electricity by 2030 [5]. Despite the greatest success of renewable energy in the global power sector, a surging global total final energy consumption still relies heavily on traditional fossil fuels [3]. It is evident that there are significant challenges to overcome to complete the low-carbon transition of global energy markets to achieve net-zero commitments for a livable climate.

Reducing carbon emissions has spurred the global proliferation of renewable energy solutions, such as hybrid renewable energy systems [6,7], thermal energy grid storage [8–10], pumped hydro storage [11,12], and fuel cells [13,14], for the decarbonization of the electricity grid. In the past decade, solar photovoltaic (PV) has become the fastest-growing energy sector among all renewables, driven by rapid technology improvements, supportive governmental policies, and significant cost reductions [15–17]. Fig. 1 summarizes the development of the major PV markets in 2021. According to statistics, the global PV market maintained its record-breaking momentum in 2021 with an additional 25% (175 GW) of capacity growth, increasing the cumulative global PV capacity to 942 GW, approximately 13 times greater than a decade ago [3,18]. China, the United States, Japan, India and Germany remain the leading countries in cumulative PV capacity, with China representing 308.5 GW accounting for about a third of global installed capacity. Moreover, PV contributes nearly 5% to the global electricity demand (see Fig. 1b). In seven countries, solar PV was expected to contribute more than 10% to the electricity demand by the end of 2021 with Australia in first place with a theoretical penetration of 15.5% [18].

The successful implementation of solar energy projects relies in part on the understanding of available solar resources and their spatiotemporal distribution characteristics. The characterization of solar resources is fundamental to determining solar technologies and project design, and indicates the largest source of uncertainty in the estimation of project power generation with a non-negligible impact on financing terms and returns on investments for solar project deployment [19]. Therefore, it is critical to conduct an accurate assessment of solar resources in order to successfully implement solar projects.

To achieve the ambitious global climate goals and increase the penetration of solar energy, a deeper understanding of the temporal and spatial availability of solar irradiance resources is necessary and critical for optimizing the deployment of solar energy projects. Driven by this growing demand, a large number of methods have been proposed for accurate solar irradiance estimation following the pioneering work of the Ångström-Prescott model [20,21]. Generally, these methods include observation-driven statistical models (e.g., empirical models [22], satellite-based remote sensing models [23,24], and machine learning models [25]) and physical-based numerical weather prediction (NWP) models [26,27]. Although empirical models are characterised by simplicity of form and ease of operation, they often fail to make accurate estimations [28]. Solar radiation derived from satellite data provides extensive spatiotemporal coverage with fast-advancing remote-sensing technologies [29]. A significant issue, however, is that the frequent updates of the versioning of gridded solar radiation products lead to seemingly never-ending validation efforts by data owners as well as third parties [30]. NWP models are considered the backbone of state-of-the-art solar radiation estimation [31]. It cannot be neglected that despite recent improvements in NWP estimations of solar radiation,

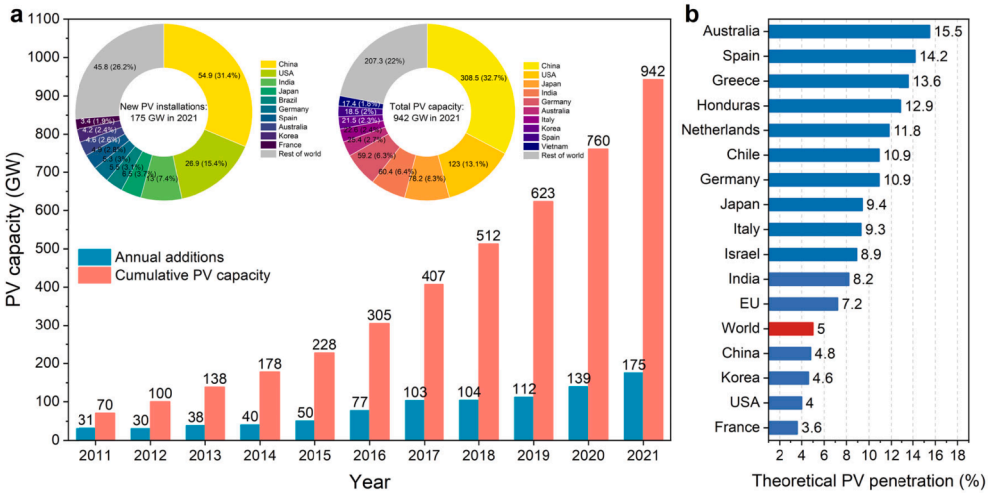


Fig. 1. Statistics of the global major solar PV markets for 2021 [3,18]. (a) Annual solar PV capacity installations and total accumulative solar PV capacity. (b) Theoretical PV penetration.

unavoidable errors associated with physical parameterizations and uncertain initial conditions result in relatively inaccurate estimations in most cases, even for the most advanced ones [32]. Also, NWP usually requires a combination of post-processing approaches, which makes NWP models more complex. Existing studies have demonstrated the promising potential of machine learning models for estimating solar radiation [25]. Machine learning models have the ability to learn complex non-linear features from various predictors such as meteorological parameters and sky images, thus building relatively better estimations [28]. Thus, further research on advanced machine learning techniques for solar radiation estimation is warranted, especially in regions with a thriving solar sector. Table 1 summarizes recent studies on solar radiation estimation based on machine learning methods. Although various methods based on different machine learning algorithms have been used for solar radiation estimation, the merits and limitations of each individual model may differ significantly. Therefore, hybrid machine learning models in which multiple algorithms work in concert to complement and augment each other are becoming a prominent practice in the solar radiation modeling community [33].

During the past years, China's installed PV power capacity has been increasing at an unprecedented pace with cost decline, solar PV technological improvement, and policy support [45,46]. It is expected that the cumulative PV power capacity in China will reach more than 600 GW by 2030, 1000 GW by 2040, and up to 1500 GW by 2060, contributing almost 25% of the total energy demand [47]. This will drive the achievement of China's goal to hit peak carbon emissions before 2030 and achieve carbon neutrality by 2060 [48,49]. On the other hand, China still faces a serious shortage of solar radiation observation data [50]. The coverage of the national solar radiation monitoring network is rather sparse, compared to the national surface meteorological stations. During the past years, many models have been developed for estimating solar radiation based on machine learning methods in China, in addition to several studies listed in Table 1. For instance, Fan et al. [51] conducted a comparison of SVM and XGBoost models and recommended the use of XGBoost for estimating daily solar radiation in China. The study reported by Zang et al. [52] demonstrated the superb accuracy and climate adaptability of the adaptive neuro-fuzzy inference system (ANFIS) model optimized by the whale optimization algorithm or the chaotic firefly algorithm for solar radiation estimation. However, it is unfortunate that despite the fact that models with promising potential have been reported for solar radiation estimations in China, there are few studies that discuss their applications in a broader context. Due to the reasons outlined above, studies focusing on China's solar radiation resources and solar energy potential based on observational data tend to be restricted to a limited spatial and temporal horizon, which may introduce significant uncertainties in the results. Also, further discussions of the model's potential practical applications considering the actual PV deployment scenario in China have usually been neglected in past studies, resulting in unclear engineering implications of the findings or conclusions. Additionally, it is also important to recognize that even though machine learning-based solar radiation models have been shown to deliver superior results, they are complex 'black-box' models that lack transparency, which casts doubt on their accuracy and reliability. Consequently, the importance of the ability to interpret the generated estimations has received increasing attention in recent work [53,54].

To this end, this study aims to provide an accurate and reliable strategy to address the current sparse coverage of solar radiation measurements in China, as well as to produce a long-term solar radiation dataset for assessing and understanding the national solar radiation resources and PV power potential. To fulfill these objectives, this study developed a hybrid machine learning model that allows for fully transparent model interpretation, contributing to the construction of global solar radiation datasets for locations without solar radiation measurements. In this method, the advanced machine learning techniques XGBoost, particle swarm optimization (PSO), and Shapley additive explanations (SHAP) were integrated as the model framework. XGBoost was selected because of its high efficiency and accuracy, and strong generalization ability, as well as being computationally less complex and less costly as compared to other algorithms. Accordingly,

Table 1
Summary of the literature studies on solar radiation estimation based on machine learning methods.

Method	Inputs	Modeling scale	Study area	Results/conclusions	Ref.
CNN-XGBoost-RF	meteorological parameters from observation and climate models	daily	Queensland, Australia	the proposed hybrid model outperformed ANN, DBN, DNN, ELM, and MARS models	[33]
ANN	satellite-derived land-surface temperature	monthly	Queensland, Australia	the preciseness of ANN for estimating GSR was established	[34]
MLP, SVM, and RF	air temperature and extraterrestrial solar radiation	daily	Spain and the United States	MLP for medium aridity areas, while SVM and RF for the aridest and most humid sites	[35]
LSTM	sky image	minutely	Turkey	the proposed sky image-based LSTM model outperformed most reported methods on a minute scale	[36]
iterative RF	Geostationary Operational Environmental Satellites −16 satellite data	half-hourly and daily	the United States	the iterative RF model outperformed traditional RF and MLR models	[37]
RLMD-LSTM	GSR 30-minute interval time series	half-hourly	Vietnam	the RLMD-LSTM model performed a generally stable capability	[38]
RRE	GSR time series	hourly	Brazil and Spain	RRE efficiently improved the accuracy of isolated models and is applicable to different sites worldwide	[39]
Generalized XGBoost	air temperature, sunshine duration, and geographical data	daily	China	the generalized XGBoost model performed better than the empirical model, and was recommended to estimate GSR for sites without solar radiation records	[40]
hybrid deep model based on encoder–decoder	air temperature, pressure, wind speed, and humidity	hourly	the United States and Australia	the proposed hybrid model is a reliable alternative with accurate and robust performance	[41]
SVM	ground meteorological and Himawari-7 satellite data	daily	China	SVM models with more complex inputs performed more accurate estimation	[42]
hybrid mind evolutionary algorithm and ANN	air temperature	daily	China	the proposed hybrid model was suggested for GSR estimation in temperate continental regions	[43]
CNN-LSTM	GSR time series	half-hourly	Australia	the hybrid CNN-LSTM model outperformed other benchmark models	[44]

Abbreviations: artificial neural network (ANN), convolutional neural network (CNN), deep belief network (DBN), deep neural network (DNN), extreme learning machine (ELM), extreme gradient boosting (XGBoost), long-short term memory (LSTM), multivariate auto-regressive spline (MARS), multilayer perceptron (MLP), multiple linear regression (MLR), random forest (RF), robust version of local mean decomposition (RLMD), ridge regression ensemble (RRE), support vector machine (SVM).

the PSO algorithm in this proposed method can significantly improve the efficiency of hyperparameter optimization, allowing XGBoost to achieve improved model performance. As an innovative method, the game theory-based SHAP method, the state-of-the-art in machine learning explainability, was introduced to integrate with the PSO-XGBoost model to provide visual explanations for the output, eliminating doubts about the accuracy and reliability of the estimation resulting from the use of 'black box' models. Most importantly, this study examined the spatiotemporal patterns of solar radiation resources and PV power potential across China with the constructed solar radiation dataset. Furthermore, this study discussed the issue of regional matching between PV deployment and theoretical PV power potential in China, providing policy recommendations for the deployment of long-term regional solar projects.

2. Methodology

2.1. Dataset description and data pre-processing

Daily measurements of global horizontal solar radiation (GSR), sunshine duration (SSD), average/maximum/minimum air temperature (TEM/TEM_{max}/TEM_{min}), air pressure (PRS), relative humidity (RHU), wind speed (WIN), average/maximum/minimum ground surface temperature (GST/GST_{max}/GST_{min}) in 2474 meteorological stations along with geographical information, including longitude, latitude, and altitude (ALT), from 1961 to 2016 were collected from the National Meteorological Information Centre, China Meteorological Administration (https://data.cma.cn). In addition, extraterrestrial solar radiation (ESR) is expressed as Eq. (1) was considered in this study [56]. In total, there are only 130 solar radiation stations, out of which 91 started recording before 1990, 26 ceased measurements before 1995, and 39 were established after 1990 [50]. Fig. 2 shows the geographical distribution of these stations across seven regions, i.e., Northern China (NC), Northeastern China (NEC), Eastern China (EC), Central China (CC), Southern China (SC), Southwestern China (SWC), and Northwestern China (NWC). Table S1 and Fig. S1 provide more details about the solar radiation stations and information on the seven regions.

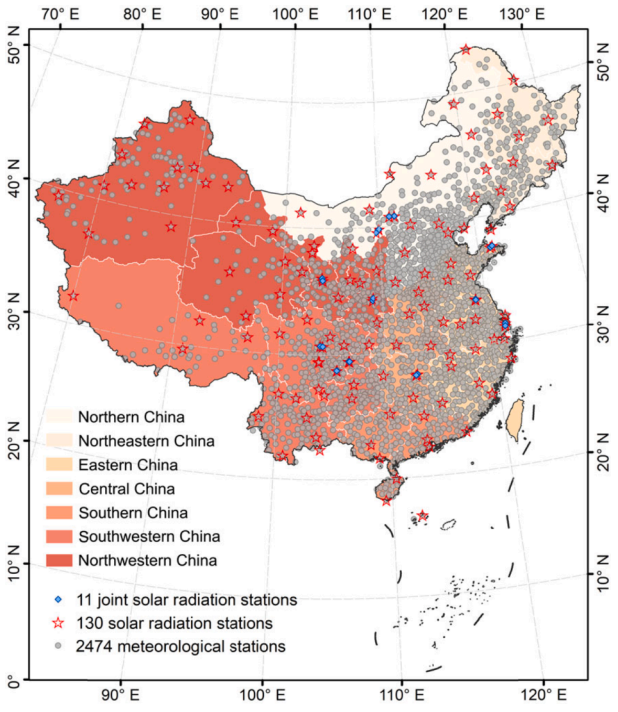


Fig. 2. The geographical distribution of 130 solar radiation and 2474 meteorological stations, as well as the seven geographical regions of China.

$$ESR = \frac{24 \times 3600}{\pi} I_{SC}\left(1 + 0.033cos\frac{360n_{day}}{365}\right)\left(cos\varphi cos\delta sin\omega_s + \frac{\pi\omega_s}{180}sin\varphi sin\delta\right) \quad (1)$$

where I_{SC} is the solar constant (equals 1367 $W{\cdot}m^{-2}$); n_{day} is the number of days in the year, starting with 1 on 1 January; φ, δ, and ω_s refer to the site latitude, solar declination, and sunset hour angle, respectively, in degrees.

The raw GSR data from 130 solar radiation stations were then processed and quality controlled to filter out physically incorrect and suspicious measurements. Specifically, the pre-1993 data were excluded from adoption for modeling purposes, considering the high uncertainty of the measurements before 1990, as well as the updates of the solar radiation measurement network during 1990–1993 [50]. Also, quality control procedures were conducted to ensure that the daily global clearness index (i.e., the ratio of GSR to ESR) ranged between the lower limit of 0.015 and the upper limit of 1.0, and the sunshine fraction (i.e., the ratio of SSD to the maximum possible sunshine duration) was limited to between 0.0 and 1.0 [57]. Measurements outside the above range were excluded from the raw dataset. The quality-controlled GSR data and the corresponding meteorological parameters from 1993 to 2016 for all solar radiation stations were then used as the dataset for developing the model. During the modeling process, the dataset was randomly divided into two parts, in which 80% of the data was set as the training set, while the other 20% was treated as the testing set.

Feature scaling plays a critical role in data pre-processing for machine learning. In this study, the Z-score normalization technique was adopted to normalize the raw data used for GSR modeling, given as.

$$x_i^* = \frac{x_i - \mu}{\sigma} \quad (2)$$

where x_i and x_i^* are the raw data to be processed and the normalized data, respectively. μ is the arithmetic mean, and σ refers to the standard deviation of the raw data.

2.2. Optimized machine learning model development

2.2.1. Feature selection

Feature selection is a critical process in solar radiation model development because redundant variables reduce the generalization ability and overall accuracy of the model. Also, adding more variables to the model increases its complexity and computational cost. In addition, the explainability of the model decreases with too many features [58]. Therefore, it is necessary to eliminate potentially redundant features from the dataset through feature selection. Based on Pearson correlation coefficients, 12 features associated with GSR from the dataset were analyzed (see Fig. S2). It is observed that the features 'ALT' and 'PRS' show a strong correlation with the coefficient of –1, and thus it can keep solely the latter. Moreover, multicollinearity can be found between several features with a high correlation coefficient of more than 0.9, such as the variables of temperature information (i.e., TEM, TEM_{max}, TEM_{min}, GST, GST_{max}, and GST_{min}). In order to eliminate the effect of multicollinearity on the model stability, 'TEM_{max}' and 'TEM_{min}' were substituted with a new feature named air temperature difference (ΔTEM) by calculating the absolute value of the difference between them. Similarly, another new feature called the ground surface temperature difference (ΔGST) was introduced to substitute 'GST_{max}' and 'GST_{min}'. In addition, the feature 'WIN' was deleted for simplicity since it has a weak correlation coefficient of 0.08 with GSR. As a result of the feature selection, eight definitive features are listed in Table 2 with their statistical descriptions, and the Pearson correlation coefficient matrix between these features is shown in Fig. 3, along with the correlation with GSR. Clearly, SSD, TEM, GST, ESR, ΔTEM, and ΔGST have a positive correlation with GSR, while SSD shows the largest positive correlation of 0.82. Negative correlation coefficients between GSR and

Table 2
Statistical characteristics for the determined features.

Feature	SSD	TEM	PRS	RHU	GST	ESR	ΔTEM	ΔGST
Unit	h	°C	hPa	%	°C	$MJ{\cdot}m^{-2}{\cdot}day^{-1}$	°C	°C
Mean	6.41	12.10	920.32	62.22	14.76	29.71	10.78	26.81
Standard deviation	4.02	12.42	111.56	19.66	13.59	9.37	4.83	13.10
Minimum value	0.00	−41.70	572.40	2.00	−43.70	5.73	0.30	0.00
Maximum value	16.40	38.90	1051.40	100.00	47.90	41.94	34.50	76.10

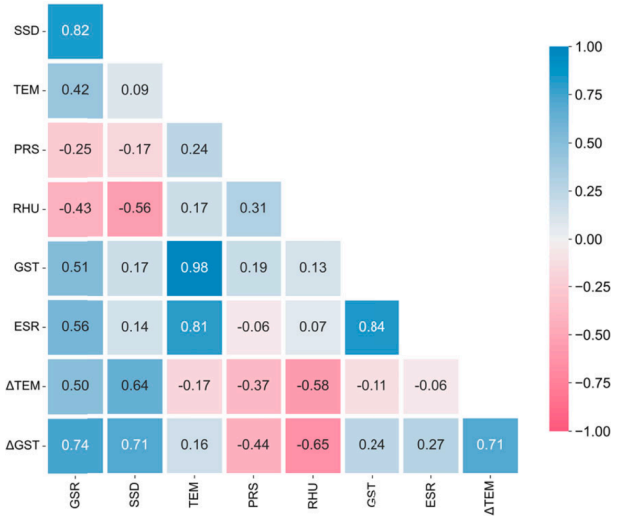


Fig. 3. Correlation coefficient matrix of GSR and eight features.

features demonstrate moderately smaller values compared to positive values. The largest negative value found between GSR and RHU is −0.43. Besides, there are a couple of noteworthy correlation coefficients observed between features. For example, the correlation between ESR and TEM is 0.81 and it increases to 0.84 between ESR and GST. Besides, although there was a large positive correlation between TEM and GST, both were retained to reduce information loss due to their fairly substantial correlation with GSR.

2.2.2. PSO-XGBoost model

XGBoost is an improved distributed gradient boosting technique [55]. Unlike the traditional gradient boosting (GBDT) algorithm, XGBoost adopts second-order Taylor's series expansion to optimize its loss function. Moreover, a regularization term is added to the objective function of XGBoost to reduce tree complexity, thus making the model less likely to be overfitted. Assuming that $\widehat{y}_i^{(t)}$ is the prediction $\widehat{y}_i$ for the t^{th} iteration, the regularized objective function ($\mathscr{L}$) of XGBoost can be defined as a composition of a differentiable convex loss function (l) and a regularization term (Ω), given as.

$$\mathscr{L}^{(j)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(j-1)} + f_j(x_i)\right) + \Omega(f_j) \tag{3}$$

$$\Omega(f_j) = \gamma\mathscr{T} + \frac{1}{2}\lambda\|\mathscr{W}\|^2 \tag{4}$$

where f_j is an independent tree structure with leaf scores, $\mathscr{T}$ is the number of leaves, $\mathscr{W}$ is the score on each leaf in the tree, γ and λ are the regularization factors that control the penalty related to $\mathscr{T}$ and $\mathscr{W}$.

PSO is a population-based stochastic optimization algorithm that imitates the social behavior of swarms of animals such as birds and insects [59,60]. The main idea of the PSO algorithm is to utilize the sharing of information among the particles in the swarm in every generation to move the whole swarm produce an evolutionary process from disorder to order in the solution space and finally find the optimal solution. Assuming that the population size is N, the position (θ) and velocity ($\mathscr{V}$) of each particle are D-dimension vectors, then the dynamics of each particle at $(k+1)^{\text{th}}$ iteration can be expressed as:

$$\mathscr{V}_{i,j}(k+1) = \omega\mathscr{V}_{i,j}(k) + c_1 r_1\left(pbest_{i,j} - \theta_{i,j}(k)\right) + c_2 r_2\left(gbest_j - \theta_{i,j}(k)\right) \tag{5}$$

$$\theta_{i,j}(k+1) = \theta_{i,j}(k) + \mathscr{V}_{i,j}(k+1) \tag{6}$$

where ω is the inertia weight, which adopted the linear decrease strategy in this study to control the global and local search ability of PSO. c_1 and c_2 are the acceleration constant for particles. r_1 and r_2 are random numbers in uniform distribution. *pbest* refers to the best-evaluated position obtained by the particle i (the self-knowledge part), while *gbest* corresponds to the best-evaluated position over the swarm (the social part).

The workflow of creating the PSO-XGBoost model is described in Fig. 4. In this study, seven important hyperparameters of XGBoost, i.e., learning_rate, subsample, max_depth, n_estimators, min_child_weight, colsample_bytree, and gamma, were selected for optimization with the PSO algorithm. Accordingly, a 7-dimensional vector space was created to represent the PSO particles, where each vector refers to one of the hyperparameters of the XGBoost model. The parameters of PSO were then determined. During the optimization process, the root mean square error of the XGBoost model was used as the fitness function in the PSO algorithm. Following that, the positions and velocities of the particles were initialized, and the fitness values of each particle were calculated. By comparison, the global and local optimal values of the whole particle swarm were found. Before reaching the maximum number of iterations, the position, velocity and inertia weight of each particle were updated, and the updated fitness values of each particle were calculated. Next, determine which local optima of the particles need to be updated and whether to update the global optima by comparing with historical fitness values. When the algorithm meets the maximum number of

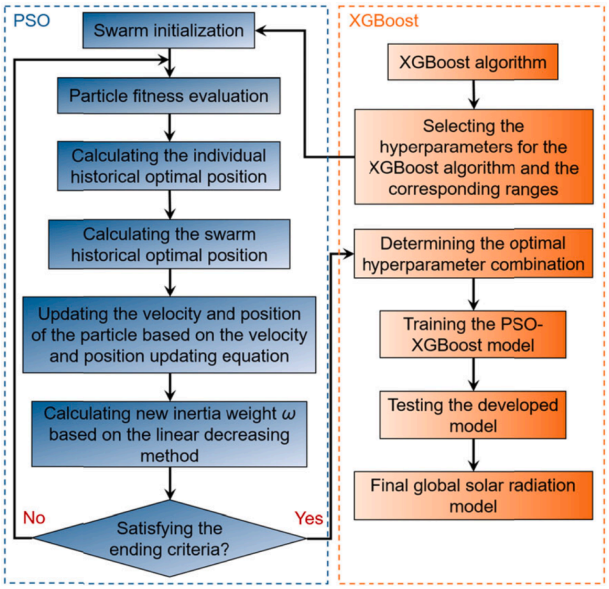


Fig. 4. The flowchart of the PSO-XGBoost algorithm.

iterations, the iteration process is ended and the optimal fitness value and the corresponding optimal position (i.e., the optimal values of the hyperparameters) are output. The parameter settings of the PSO-XGBoost model are listed in Table S2, and the fitness curve of the PSO-XGBoost model with iterations is shown in Fig. S3.

2.2.3. Evaluation metrics

In this study, the most frequently used statistical metrics, i.e., the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) were adopted to evaluate the model performance, given as:

$$R^2\left(GSR_{i,m}, GSR_{i,e}\right) = 1 - \frac{\sum_{i=1}^{n}\left(GSR_{i,e} - GSR_{i,m}\right)^2}{\sum_{i=1}^{n}\left(GSR_{i,m} - \overline{GSR}_{i,m}\right)^2} \tag{7}$$

$$\mathrm{RMSE}\left(GSR_{i,m}, GSR_{i,e}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(GSR_{i,e} - GSR_{i,m}\right)^2} \tag{8}$$

$$\mathrm{MAE}\left(GSR_{i,m}, GSR_{i,e}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|GSR_{i,e} - GSR_{i,m}\right| \tag{9}$$

$$\mathrm{MAPE}\left(GSR_{i,m}, GSR_{i,e}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{GSR_{i,e} - GSR_{i,m}}{GSR_{i,m}}\right| \times 100\% \tag{10}$$

where $GSR_{i,e}$ and $GSR_{i,m}$ are estimated and measured GSR values, respectively. $\overline{GSR}_{i,m}$ is the average value of measured GSR.

2.2.4. Shapley additive explanations (SHAP)

The machine learning model should be transparent, interpretable, and trustworthy. Understanding the reasons why a model generates a certain output is important to explain results, interpret the process being modeled, and assess how well the users understand the phenomenon under study, and is a critical step in building confidence in the model results [61,62]. Broadly, the machine learning models can be interpreted on three scales, i.e., (i) global level to evaluate the whole model performance; (ii) feature level to examine the importance of features as well as the interactions between them; and (iii) data point level to explain the behavior of a model at an individual point [63]. In this study, the model interpretation focused on the feature level and data point level to reveal the role played by input features in the model output using the SHAP method.

SHAP provides explanations for the output of any machine learning models with the theoretical guarantee of methodological consistency based on the Shapley values from game theory [64,65]. It shows better adaptation to human intuition by calculating SHAP values to assess the local accuracy, consistency, and missingness of the input features. The SHAP value is defined as the solution to Eq. (11), which indicates the Shapley value of an expectation function of the original machine learning model f conditioned on a feature subset $\mathscr{S} \subseteq \mathscr{U}$ ($\mathscr{U}$ is the set of all features), i.e., $\mathbb{E}[f(x)|x_{\mathscr{S}}]$. Readers are referred to Ref. [64] for more details on SHAP values.

$$\phi_i = \sum_{\mathscr{S}\subseteq\mathscr{U}\setminus\{i\}} \frac{|\mathscr{S}|!(|\mathscr{U}| - |\mathscr{S}| - 1)!}{|\mathscr{U}|!}\left[f_{\mathscr{S}\bigcup\{i\}}\left(x_{\mathscr{S}\bigcup\{i\}}\right) - f_{\mathscr{S}}\left(x_{\mathscr{S}}\right)\right] \tag{11}$$

where ϕ_i indicates an effect of a linear function of binary variables to each feature ($\phi_i \in \mathbb{R}$). $f_{\mathscr{S}\bigcup\{i\}}$ and $f_{\mathscr{S}}$ are the models that trained with the feature present and withheld, respectively, to compute the effect ϕ_i. $x_{\mathscr{S}}$ refers to the values of the input features in subset $\mathscr{S}$.

2.3. Solar photovoltaic power model

The operating temperature of a PV module (T_c) has a significant effect on its electrical efficiency (η_c). In this study, the general empirical model for PV module efficiency as a function of operating temperature was adopted to estimate the PV power potential, see Eq. (8) [66].

$$\eta_c = \eta_{ref}\left[1 - \beta_{ref}\left(T_c - T_{ref}\right) + \gamma\log_{10}GSR_c\right] \tag{12}$$

where η_{ref} is the reference electrical efficiency of a PV module at reference temperature $T_{ref} = 25$ °C and solar irradiance of 1000 W·m^{-2}. β_{ref} and γ refer to the efficiency correction coefficient for temperature and the efficiency correction coefficient for solar irradiance, respectively. For silicon PV modules, $\beta_{ref} = 0.0045$ K^{-1}, and $\gamma = 0.12$ [67,68]. However, the latter (i.e., $\gamma\log_{10}GSR_c$) is generally taken as 0 [68]. GSR_c is the solar irradiance on the PV module plane. In this study, GSR_c is assumed to be equal to the global horizontal solar irradiance, i.e., the PV tilt angle is 0°. T_c is calculated based on the energy balance on the PV module, as expressed below [68]:

$$T_c = T_a + \tau\alpha\frac{\mathrm{GSR}_c}{U_L}\left(1 - \frac{\eta_c}{\tau\alpha}\right) \tag{13}$$

where T_a is the ambient temperature. α and τ are the PV module surface absorbance and transmittance, respectively. U_L is the overall thermal loss coefficient, which is assumed constant and obtained from the nominal operating cell temperature (NOCT) test. Then, Eq. (13) can be written as the expression below [68]:

$$T_c = T_a + GSR_c\frac{NOCT - (T_a)_{NTE}}{(GSR_c)_{NTE}}\left(1 - \frac{\eta_c}{\tau\alpha}\right) \tag{14}$$

where the so-called NOCT is defined as the temperature of a PV module at the nominal terrestrial environment (NTE) conditions at the ambient temperature of 20 °C, solar irradiance of 800 W·m^{-2}, average wind speed of 1 m·s^{-1}, and no load operation (meaning open circuit). It is assumed that $\tau\alpha$ in Eq. (14) equals 0.9 [56,69].

For a certain period (t), the PV power output (P_{PV}) and energy output (E_{PV}) are given respectively by.

$$P_{PV} = GSR_c\eta_c \tag{15}$$

$$E_{PV} = P_{PV}t \tag{16}$$

assuming 365 days for a year.

2.4. Inverse distance weighted (IDW) interpolation

The inverse distance weighted (IDW) interpolation method determines unknown cell values based on a linearly weighted combination of a set of sample points, which is one of the most commonly used deterministic models in multivariate interpolation with scattered points [70]. Briefly, the general idea of IDW is to assume that the attribute values of the unsampled points are weighted averages of the known values within the adjacent area, with the weights inversely proportional to the distance between the predicted and sampled locations [71]. In this study, the solar radiation and PV energy output data generated for each weather station based on the proposed model were interpolated into grids with 50 km by 50 km spatial resolution using the IDW method to produce maps of national solar radiation resources, as well as PV power potential. Moreover, the gridded data were also used for spatial and temporal analyses.

3. Results and discussion

3.1. Comparison of different models

Several machine learning methods, including XGBoost, GBDT, light gradient boosting machine (LightGBM), K-nearest neighbor (KNN), multilayer perceptron (MLP), and support vector machine with the radial basis function kernel (RBF-SVM), as well as the linear Ångström-Prescott (A-P) model were also adopted to build models for comparison with the developed PSO-XGBoost model. Table 3 gives the overall performance of these models. The training scores of R^2, RMSE, MAE, and MAPE for the PSO-XGBoost model are 0.956, 1.555 MJ·m^{-2}·day^{-1},

Table 3
Comparison between the performance of the developed PSO-XGBoost model with several machine learning models and the A-P model.

Model	R^2		RMSE ($MJ·m^{-2}·day^{-1}$)		MAE ($MJ·m^{-2}·day^{-1}$)		MAPE (%)	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
PSO-XGBoost	0.956	0.943	1.555	1.754	1.107	1.261	10.097	12.337
XGBoost	0.943	0.940	1.770	1.804	1.276	1.304	12.534	12.882
GBDT	0.932	0.932	1.926	1.927	1.409	1.411	14.844	14.789
LightGBM	0.939	0.938	1.827	1.837	1.326	1.334	13.646	13.653
KNN	0.942	0.912	1.776	2.182	1.322	1.628	13.095	16.124
MLP	0.936	0.936	1.866	1.870	1.358	1.365	13.439	14.708
RBF-SVM	0.893	0.894	2.415	2.407	1.841	1.836	22.687	22.428
A-P model	0.897	0.897	2.367	2.363	1.780	1.775	23.376	23.222

1.107 $MJ·m^{-2}·day^{-1}$, and 10.097%, respectively, improving by 1.38%, 12.15%, 13.24%, and 19.44% compared to the XGBoost model. More significant improvements can be observed when comparing the performance of the PSO-XGBoost model with models based on GBDT, LightGBM, KNN, MLP, RBF-SVM, and the A-P model, with improvements between 1.49–7.05%, 12.44–35.61%, 16.26–39.87%, and 22.89–56.81% observed for R^2, RMSE, MAE, and MAPE, respectively. For the testing set, the score of R^2 for the PSO-XGBoost model is 0.934 with RMSE, MAE, and MAPE values of 1.754 $MJ·m^{-2}·day^{-1}$, 1.261 $MJ·m^{-2}·day^{-1}$, and 12.337%, respectively, which provides the highest performance with the least statistical error for the whole dataset.

Moreover, the performance of the developed models in this study at the individual site demonstrates the superior performance of the PSO-XGBoost model in terms of accuracy and stability in GSR estimation, compared to other machine learning models as well as the A-P model (see Fig. 5). Specifically, the ranges of R^2, RMSE, MAE, and MAPE of the PSO-XGBoost model are 0.891–0.975, 1.109–2.099 $MJ·m^{-2}·day^{-1}$, 0.816–1.508 $MJ·m^{-2}·day^{-1}$, and 4.990–18.610%, respectively, with average values of 0.946, 1.580 $MJ·m^{-2}·day^{-1}$, 1.137 $MJ·m^{-2}·day^{-1}$, and 10.571%. Also, the narrower range and smaller values of statistical errors of the PSO-XGBoost model compared to the XGBoost model indicate the effective implementation of the PSO optimization algorithm. On the contrary, the RBF-SVM and A-P models showed inferior model performance. For example, the values of R^2, RMSE, MAE, and MAPE for the A-P model vary in the range of 0.576–0.955, 1.552–3.740 $MJ·m^{-2}·day^{-1}$, 1.125–3.201 $MJ·m^{-2}·day^{-1}$, and 8.720–68.980%, respectively. This large range in statistical error values indicates the instability of the performance of the linear A-P model.

The comparison between the GSR from measurements and estimations of different models at all sites is given in Fig. 6. For both the RBF-SVM model and the linear A-P model, neither generated acceptable performance in estimating daily GSR, with R^2 below 0.9, as well as larger values of RMSE, MAE, and MAPE. Turning to models based on PSO-XGBoost, XGBoost, GBDT, LightGBM, KNN, and MLP, significant improvements were found in terms of statistical errors. Furthermore, the PSO-XGBoost model provided even better performance with R^2, RMSE, MAE, and MAPE of 0.953, 1.597 $MJ·m^{-2}·day^{-1}$, 1.138 $MJ·m^{-2}·day^{-1}$, and 10.500%, respectively, improving by 1.17–6.72%, 10.28–33.79%, 11.37–38.12%, and 16.67–54.94%. Therefore, it was suggested to adopt the PSO-XGBoost model to estimate daily GSR in China.

3.2. Spatiotemporal performance of the PSO-XGBoost model

After discussing the superior performance of the PSO-XGB model compared to other models, its spatiotemporal performance in estimating GSR is explored in this section. Fig. 7a depicts the statistical error of the PSO-XGBoost model for each GSR station using all data after 1993, and its overall performance for seven regions is shown in Fig. 7b. Generally, the results indicated that the PSO-XGBoost model performed well in GSR estimation across seven regions in China with reasonable average statistical errors between 0.929–0.955, 1.498–1.705 $MJ·m^{-2}·day^{-1}$, 1.054–1.205 $MJ·m^{-2}·day^{-1}$, and 9.547–15.066% for R^2, RMSE, MAE, and MAPE, respectively. Moreover, in terms of spatial differences in model performance, it can be found that the PSO-XGBoost model generated better estimations in the NEC, NC, and EC regions with a more concentrated error distribution. Their average R^2, RMSE, MAE, and MAPE are approximately 0.95, 1.5 $MJ·m^{-2}·day^{-1}$, 1.1 $MJ·m^{-2}·day^{-1}$, and 10%, respectively. In contrast, although the PSO-XGBoost model produced acceptable average estimation errors in the NWC region, there are large differences between stations within this region. It is also the case for the SWC region. This variation was considered mainly to be a result of the complexity of the internal climate [72].

In addition, the performance of the PSO-XGBoost model on both daily and monthly mean daily temporal scales was assessed (see Fig. 8). Overall, the PSO-XGBoost model underestimated the daily and monthly mean daily GSR in China with a relative error of less than 1% for most days and months. For the majority of the year, underestimations of no more than 3% occurred in the NEC, NC, EC, SC, and SWC regions, with a relative error of about −4% in the NEC and SC regions for a few daily estimations. On the contrary, the GSR in the NWC region was

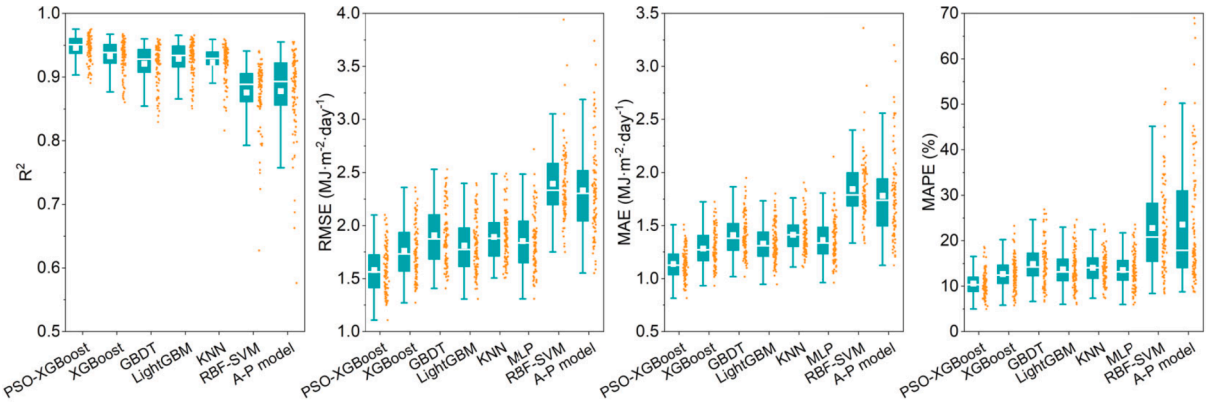


Fig. 5. Boxplot of the performance of different models at each of the sites used for modeling.

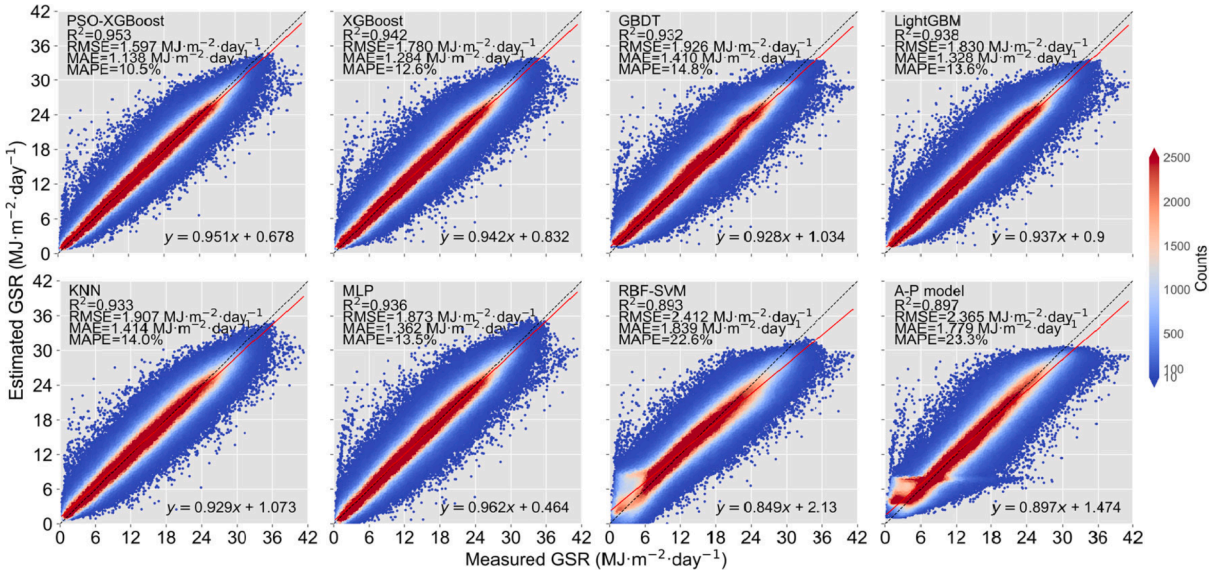


Fig. 6. GSR comparison between estimated values and measurements at all sites used for modeling.

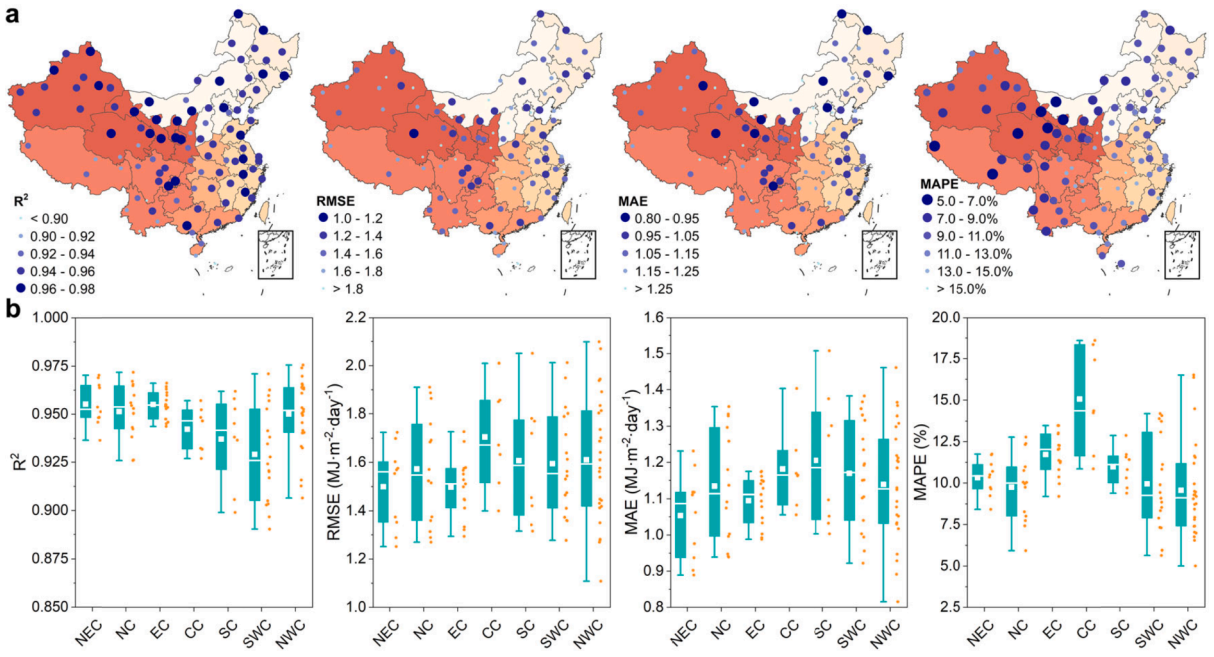


Fig. 7. Statistical errors of the PSO-XGBoost model for (a) all stations with GSR measurements after 1993 and (b) seven regions in China.

overestimated by less than 2% and 1% in most cases for daily and monthly mean daily GSR, respectively. In the CC region, underestimation and overestimation occurred roughly at the same frequency, with absolute values of relative errors of less than 4% and 1%, respectively, for daily and monthly mean daily GSR. In addition, it is noteworthy that the relative errors tend to be larger in autumn (September to November) for both daily and monthly mean daily estimations, whereas the estimations tend to be more accurate in the spring (March to May). Fig. S4 shows the frequency distribution of the absolute values of the relative errors for the daily GSR estimations. For instance, approximately 85% of the GSR estimations in the NEC region in the spring had an absolute error of less than 1%, and only 1% of days had an error greater than 2%. As a comparison, about 34% of the GSR estimations for the autumn had an absolute error of more than 2%.

3.3. *Model interpretation based on SHAP*

As described in Section 2.2.4, SHAP values were applied to interpret the PSO-XGBoost model. The feature importance of the PSO-XGBoost model for GSR estimation is shown in Fig. 9a. The mean absolute SHAP values indicated that SSD was the most important feature for the estimation of GSR, followed by ESR, in agreement with the claims made in the literature [72,73]. On the other hand, ΔTEM, RHU, and PRS had less influence on the model output. The distribution of the SHAP values of all instances for each feature is depicted by a SHAP summary plot as shown in Fig. 9b, which illustrates the instances' individual and oriented effect on the model output. A smooth change in the coloring of the points can be found for all instances of SSD and ESR, indicating a monotonic positive correlation between these two dominant features and GSR. In addition, temperature-based solar radiation models tend to choose air

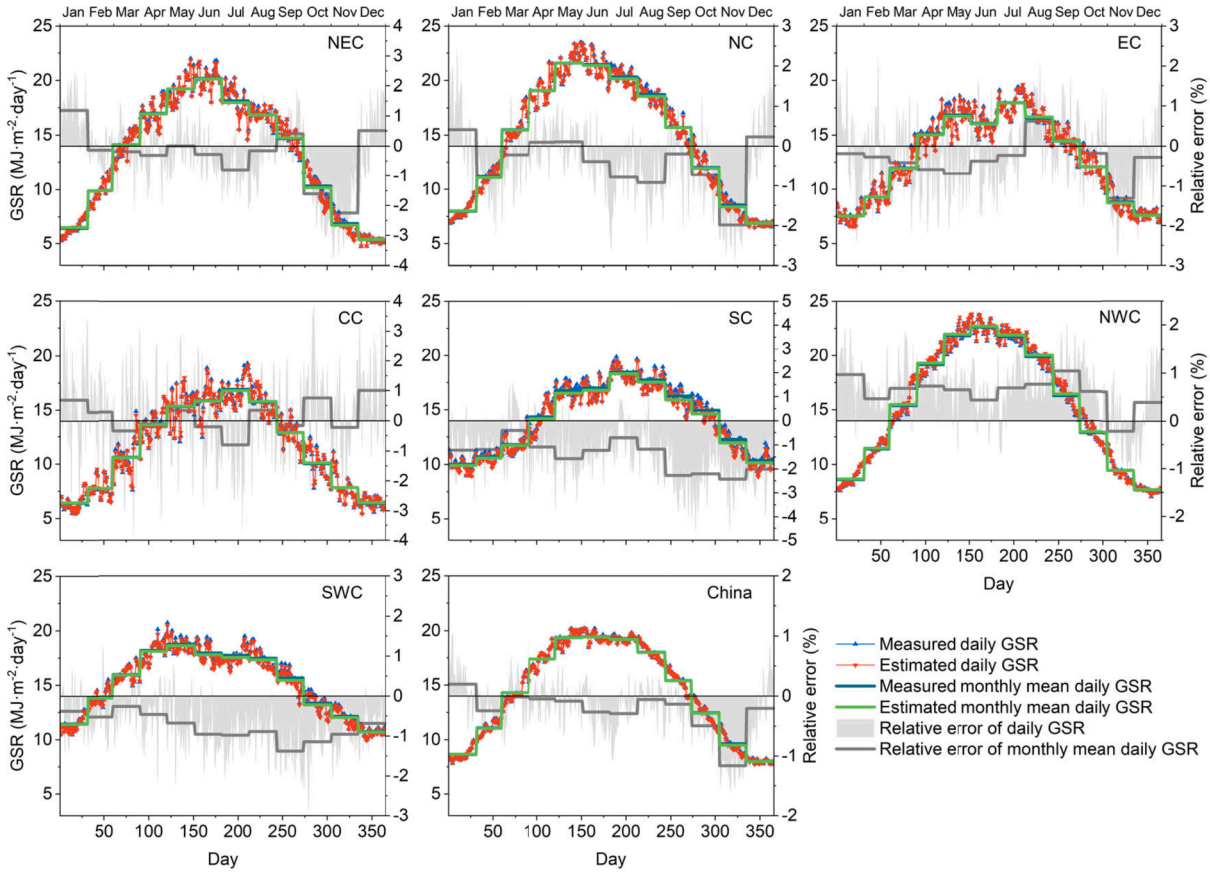


Fig. 8. Comparison between the average measured GSR and the values estimated by the PSO-XGBoost model on daily and monthly mean daily scales.

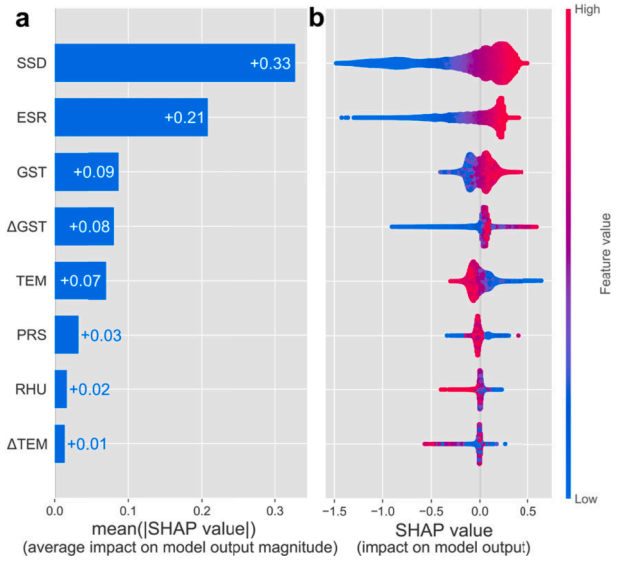


Fig. 9. SHAP feature importance of the PSO-XGBoost model. (a) The mean absolute SHAP values of features. (b) SHAP plot summarizing all of the features for all instances.

temperature rather than ground surface temperature when considering the use of temperature as an input variable [74,75]. However, the distribution of SHAP values in Fig. 9b suggests that GST is more significant than TEM, despite the shorter tail. As a result, although GST had a moderate effect on the output of the model, it was experienced by most instances.

Apart from the global explanation of features, Fig. 10 provides two samples for local interpretation obtained from the SHAP waterfall plot, which visually explains the effect of each feature on the individual output. The length of each bar indicates the SHAP value for that feature, and the red and blue denote that the feature affected the model output to move higher and lower, respectively. It should be noted that waterfall plots are limited to interpreting individual output without being applicable to drawing general inferences about model behavior. As shown in Fig. 10, SSD and ESR were the most significant features affecting the model output in both samples, while RHU, ΔTEM, and PRS contributed negligibly to the final output. When comparing these two samples, a larger SSD obtained a higher SHAP value and tended to push the GSR upwards as well. Moreover, the lower ESR and GST in sample #1 in winter both produced negative SHAP values, pushing GSR to a lower value, whereas their higher values in sample #2 (summer) drove the model to output a larger GSR.

3.4. Solar radiation dataset reconstruction and photovoltaic power assessment

With the developed PSO-XGBoost model, it was possible to estimate daily GSR at locations without solar radiation measurements, thereby reconstructing the long-term solar radiation dataset for 2474 meteorological stations in China between 1961 and 2016. Furthermore, the dataset was used to assess the long-term PV power potential, i.e., the estimated PV energy output E_{PV}, in China based on a PV power model.

3.4.1. Map of solar radiation resources

In this study, the theoretical solar radiation resources in China were assessed based on the assumption of the long-term availability of solar radiation at any site and indicated in terms of global horizontal

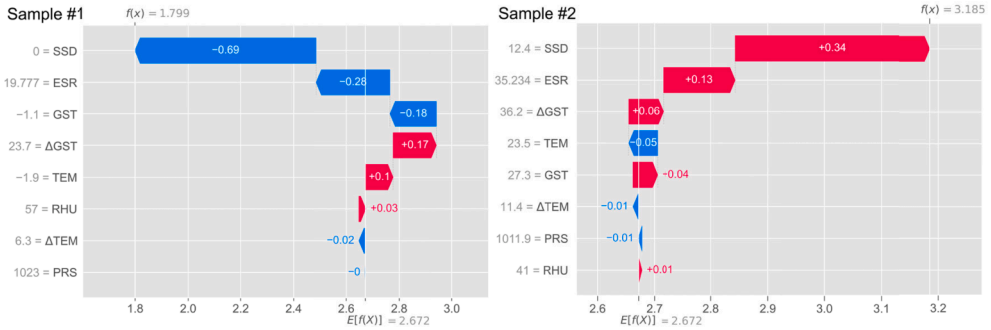


Fig. 10. Example SHAP waterfall plots for two estimations in Beijing on 12 February 2012 (sample #1), and 22 August 2012 (sample #2).

irradiation, i.e., GSR. Fig. 11a shows the spatial distribution of the annual mean GSR from 1961 to 2016 in China. Solar radiation resources are plentiful in the Qinghai-Tibet Plateau, the western and northern regions of China. The southeastern provinces of China are characterised by relatively low solar radiation resources, especially in the Sichuan Basin, Chongqing, and northern Guizhou. The names of each province of China are given in Fig. S1. Specifically, the descriptions of solar radiation resources for each province and region are illustrated in Fig. 11b and 11c. The long-term annual mean GSR in China was estimated at 174.36 $W{\cdot}m^{-2}$. The largest potentials of solar radiation were concentrated in NWC with an annual mean GSR of 188.07 $W{\cdot}m^{-2}$, accounting for about 34.36% of the national resources. On the other hand, the annual mean GSR of the EC, CC, and SC regions was all below 150 $W{\cdot}m^{-2}$. Moreover, due to its high altitude, Tibet, located in the SWC region, had the most abundant solar radiation resources of 218.71 $W{\cdot}m^{-2}$, representing 15.90% of the nation's solar radiation resources. In contrast, Chongqing is the province that has the lowest solar radiation resources in China with a long-term annual mean GSR of 112.33 $W{\cdot}m^{-2}$. In general, the main reasons for this difference in the spatial pattern of solar radiation resources can be attributed to the higher altitudes of the western and northern provinces, as well as their arid climate with lower cloud cover than the humid southeastern provinces.

3.4.2. *Spatiotemporal variations of solar radiation resources*

For a better understanding of the likely changes in solar radiation resources, this section also provides analyses of the spatiotemporal variability of GSR, which is becoming increasingly important for the rapidly growing PV sector. Past studies have typically defined the

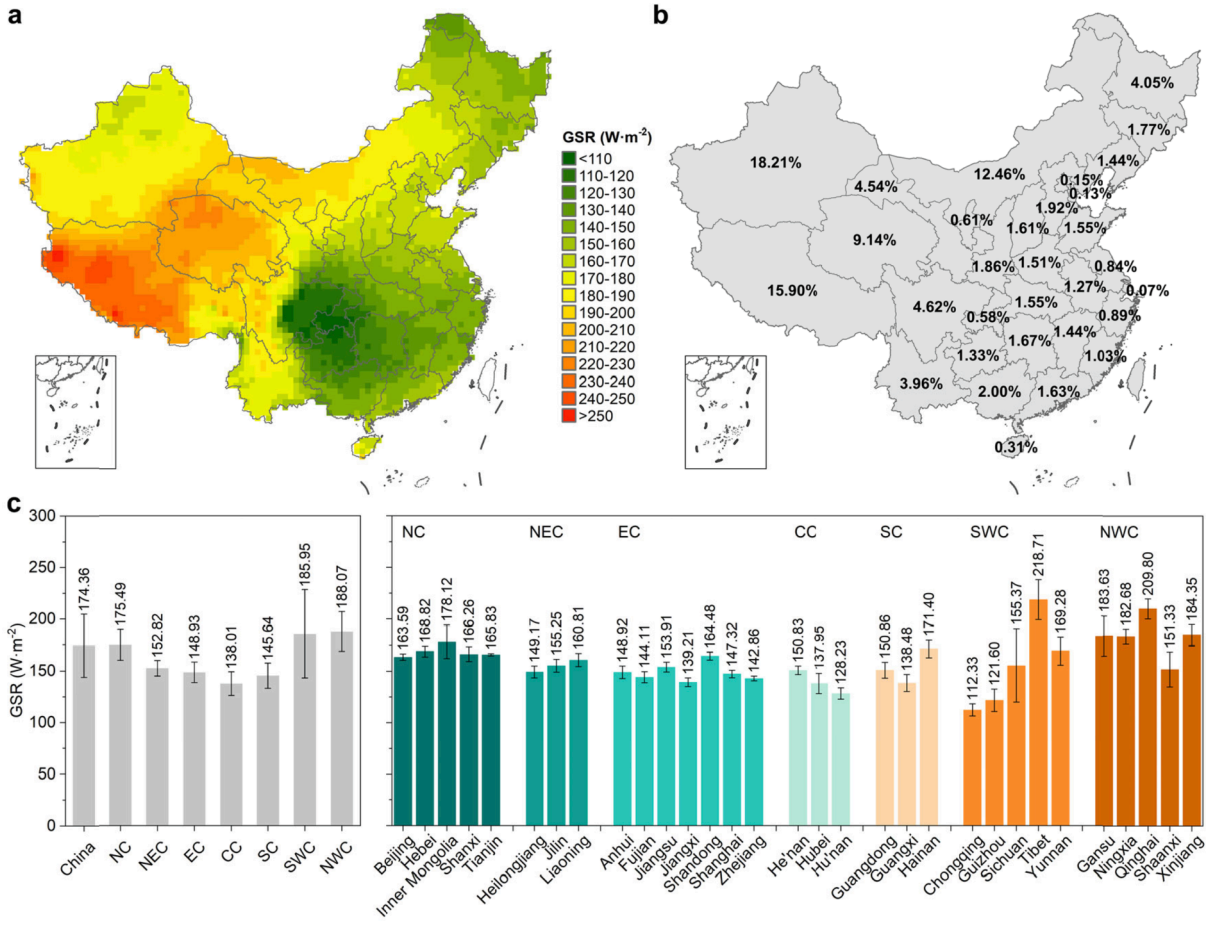


Fig. 11. Long-term solar radiation resources from 1961 to 2016 in China. (a) Spatial distribution of annual mean GSR across China. (b) Share of solar radiation resources in each province. (c) Average annual mean GSR on regional and provincial levels.

periods before and after the 1990s as 'global dimming' and 'global brightening', respectively [50,76]. Fig. 12 shows these two phases and the long-term GSR trends based on reconstructed and measured data from 130 solar radiation stations. During the dimming period (1961–1990), the decreasing trend of the reconstructed GSR was −2.68 ± 0.70 W·m^{-2}·decade^{-1} ($p < 0.01$), which is in agreement with the trend between −2.90 and −2.66 W·m^{-2}·decade^{-1} for the modified GSR derived from sunshine duration [77,78]. However, a downward trend of up to −8.04 ± 0.64 W·m^{-2}·decade^{-1} ($p < 0.01$) for the dimming period was observed based on the measured GSR. It is likely that this overestimation is due to the inhomogeneity of the measurements as a result of the strong sensitivity drift of the pyranometers used before the 1990s [50,76,77]. Following the 1990s, the trend of dimming diminished and a brightening tendency emerged in this century, with an increasing rate of 1.18 ± 0.50 W·m^{-2}·decade^{-1} ($p < 0.05$) and 2.10 ± 0.79 W·m^{-2}·decade^{-1} ($p < 0.05$) for the reconstructed and measured GSR, respectively, from 1990 to 2016. Overall, the long-term GSR illustrates a decreasing trend, with a decadal variability of −1.15 ± 0.26 W·m^{-2}·decade^{-1} ($p < 0.01$) for the reconstructed GSR during the period 1990–2016.

Given the sparse coverage of solar radiation stations, the estimated national GSR trends based on only the data from 130 stations appear to lack representativeness and credibility. Therefore, long-term trends were further analyzed using reconstructed solar radiation data from 2474 meteorological stations located throughout China. Fig. 13 depicts the spatial pattern of the GSR trends for the 'global dimming and brightening' periods and the long-term period from 1961 to 2016, and a detailed description of the GSR trends is provided in Table S3. For the dimming phase (see Fig. 13a), most of China showed a decreasing trend, with an average decline rate of −1.49 W·m^{-2}·decade^{-1}. It is estimated that the decadal decrease of GSR was up to 4.33 W·m^{-2}·decade^{-1} in EC, and more than 3.5 W·m^{-2}·decade^{-1} in the low-resource regions such as CC and SC. Turning to the brightening phase (see Fig. 13b), a fair number of provinces showed an increase in annual mean GSR. For instance, Yunnan showed the highest increasing trend of 3.83 W·m^{-2}·decade^{-1} from 1990 to 2016. There was also a recovery rate exceeding 1 W·m^{-2}·decade^{-1} in Inner Mongolia, Heilongjiang, Jilin, Xinjiang, Tibet, Sichuan, and Jiangxi. Conversely, the Beijing-Tianjin-Hebei region, Henan, and Shandong were still experiencing a declining trend of around −2 W·m^{-2}·decade^{-1} or higher. As a whole, the annual mean GSR in China has decreased consistently over the past decades with a decadal trend of −0.83 W·m^{-2}·decade^{-1}. The northwestern provinces show the least decrease, with an average reduction of 0.27 W·m^{-2}·decade^{-1} for NWC. The declining trend in the Beijing-Tianjin-Hebei region and its surrounding areas was the highest in China, with annual mean GSR reduced by 2.89–3.98 W·m^{-2}·decade^{-1}. Interestingly, these areas are also characterized by the most severe air pollution, which has been suggested to be a primary cause for the reduction in solar radiation resources [79].

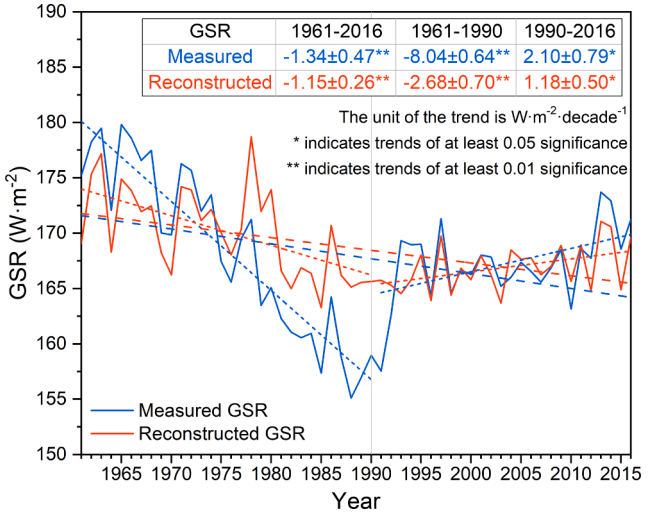


Fig. 12. Annual variations in average GSR from 1961 to 2016 at 130 solar radiation stations in China.

According to Fig. 14a, the spatial pattern of the seasonal mean GSR across China resembles that of the annual mean GSR, with higher resources distributed in the northwestern and northern provinces, and lower values distributed in the southeastern provinces. The average national GSR in spring, summer, autumn, and winter is 204.40, 229.25, 151.28, and 112.47 W·m^{-2}, respectively (see Fig. S5). Furthermore, the spatial patterns of seasonal GSR trends from 1961 to 2016 are shown in Fig. 14b. The GSR decreased by 1.83, 0.74, and 0.83 W·m^{-2}·decade^{-1} in summer, autumn, and winter, respectively, except for spring with an increasing trend of 0.36 W·m^{-2}·decade^{-1}. Besides, it is noteworthy that the GSR in the NWC region, as well as some provinces in the southeast, increased in the spring, while the Beijing-Tianjin-Hebei region and its neighboring provinces, like Shandong, Henan, and Shanxi, exhibited the highest decreasing tendency in the other three seasonal scenarios. More detailed information on the data is provided in Table S4.

3.4.3. Assessment of solar photovoltaic power potential

Fig. 15 depicts the spatial distribution of average yearly PV power potential from 1961 to 2016 as well as statistics by region and province. Generally, the long-term average yearly PV power potential shows a similar spatial pattern with high levels in the northwestern and northern provinces while low levels in the southeastern provinces as solar radiation resources (see Fig. 15a). As illustrated in Fig. 15b and 15c, the over-50-year average result shows a yearly PV power potential of 285.00 kWh·m^{-2}, with a maximum of 308.84 kWh·m^{-2} in the NWC region, representing 34.52% of the national potential. Benefiting from the high altitude, Tibet is at the top of the provinces with 365.37 kWh·m^{-2} of yearly PV power potential, which accounts for 16.25% of the total PV power potential in China. Moreover, Xinjiang holds the largest share at 18.06% with a yearly PV power potential of 289.90 kWh·m^{-2}, as a result of its vast land areas and abundant solar radiation resources. On the other hand, humid and cloudy provinces such as Chongqing and Guizhou have the lowest PV power potential of fewer than 200 kWh·m^{-2}, which is up to 40% or more below the northwestern provinces.

A description of the PV sector deployment status by the province in 2021 is shown in Fig. S6. In 2021, China added around 54.88 GW of on-grid solar PV capacity, of which about 10.71 GW (19.51%) was contributed by Shandong, followed by Hebei and Henan contributed 7.30 GW (13.30%) and 3.81 GW (6.94%), respectively [80]. The leading province for total solar PV capacity in 2021 remained Shandong, with a cumulative capacity of 33.43 GW, accounting for 10.93% of the national PV installations. The next four provinces in 2021 were Hebei (9.55%), Jiangsu (6.26%), Zhejiang (6.02%), and Anhui (5.58%). However, Tibet, Qinghai, Xinjiang, Inner Mongolia, and Gansu, with the largest PV power potentials and land areas, contributed only 8.17% of the nation's PV additions in 2021, less than half of Shandong. Also, their cumulative PV installations were much lower than that of low-potential eastern provinces, like Shandong and Jiangsu, indicating a regional mismatch in the development of the solar PV sector in China. Therefore, the holistic layout of solar PV under China's 2060 carbon neutrality target requires further consideration of the spatial pattern of PV power potential. A geographically targeted development strategy adapted to the local environmental, economic, and resource characteristics is imperative for leading the future solar PV industry in China onto a healthy and sustainable development track. Specifically, for Xinjiang, Inner Mongolia, Qinghai, Gansu, and Tibet with huge PV power potential and sparse populations, it is most appropriate to prioritize the construction of large-scale centralized PV power plants to fully exploit the solar energy of the region, while the southeastern provinces should focus on developing

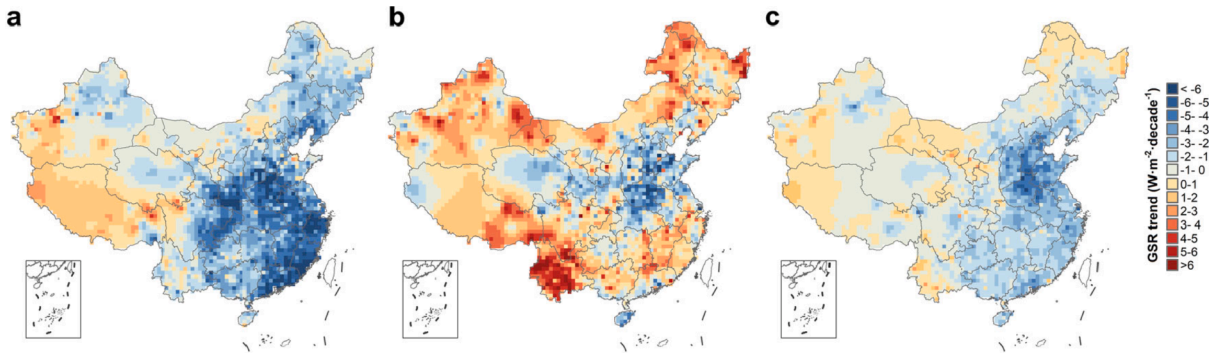


Fig. 13. Spatial pattern of GSR trends in China for (a) 1961–1990, (b) 1990–2016, and (c) 1961–2016.

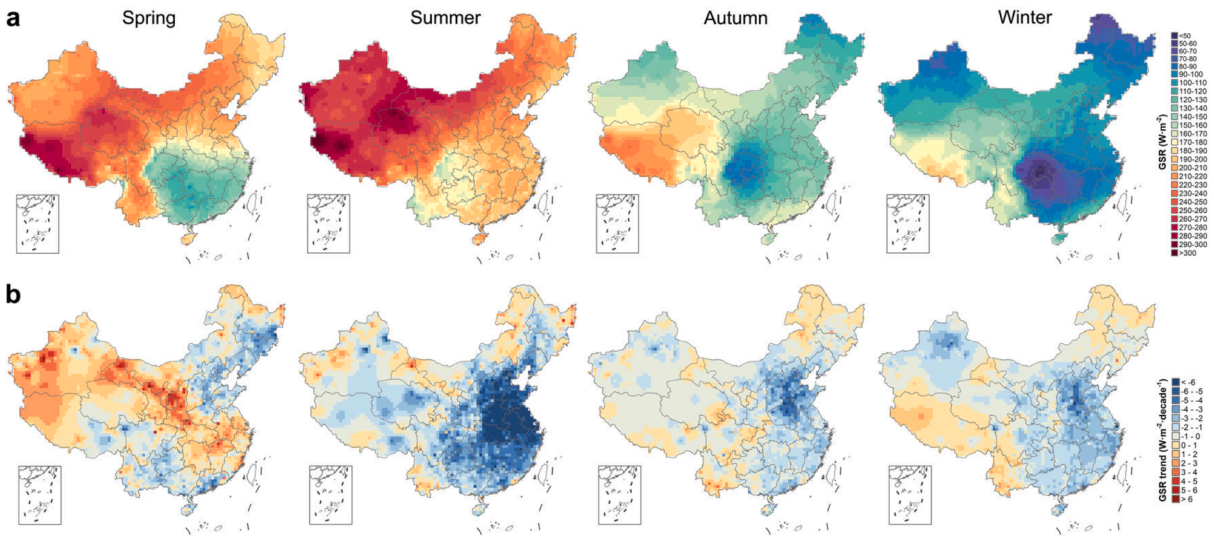


Fig. 14. The long-term seasonal pattern of mean GSR from 1961 to 2016 in China. (a) Spatial distribution of seasonal mean GSR. (b) Spatial distribution of the seasonal GSR trends for the period from 1961 to 2016.

distributed PV. Moreover, on a long-term policy-oriented level, government policy guidance and support are crucial for developing an effective solar PV power market, such as the interconnection of regional grids and PV power trading between regional grids, as China's regional economic development and energy demand show a spatial pattern opposite to the distribution of PV power potential.

3.4.4. Spatiotemporal pattern of solar photovoltaic power potential

Fig. 16 depicts the spatial distribution of the trends in PV power potential in China over the three periods, indicating a similar pattern to that of GSR since PV power output is primarily determined by solar radiation. As illustrated in Fig. 16a, the national PV power potential decreased by an average of -2.54 kWh·m^{-2}·decade^{-1} between 1961 and 1990. All provinces except Tibet were characterized by a predominantly downward trend, especially in the low-potential regions. For example, provinces in EC, CC, and SC saw average PV power potential reductions of -6.81, -6.09, and -5.82 kWh·m^{-2}·decade^{-1}, respectively. On the other hand, the downward trend in the high-potential NEC was relatively minor (-1.33 kWh·m^{-2}·decade^{-1}), with Xinjiang trending at -0.65 kWh·m^{-2}·decade^{-1} during this period. Although the national PV power potential indicated an increasing trend of 0.68 kWh·m^{-2}·decade^{-1} after the 1990s, the average yearly potential in China still decreased by -1.69 kWh·m^{-2}·decade^{-1} in the long term from 1961 to 2016, as illustrated in Fig. 16b and 16c (see Table S5 for a more detailed description of the PV power potential tendencies). In addition, the attenuation in the long-term PV power potential in heavily polluted regions such as the Beijing-Tianjin-Hebei region was up to -6.66 to -4.88 kWh·m^{-2}·decade^{-1} during the period from 1961 to 2016.

Furthermore, Fig. 17 compares the PV power potential of China in the 1960s and 2010s. It is found that the overall spatial pattern of PV power potential in the 2010s scenario is similar to that of the 1960s conditions (see Fig. 17a). However, as shown in Fig. 17b, the national PV power potential in the 2010s shows a general decline compared to the 1960s, with greater percentage changes observed in the high-polluted northern, eastern, and southern regions, increasing the difference between the high-potential northwestern regions and the low-potential southeastern regions. Fig. 17c illustrates the comparison of the average PV power potential in different regions under the two scenarios, as well as by province. Specifically, the PV power potential of China decreased by 2.88% from 287.55 (the 1960s) to 280.21 (the 2010s) kWh·m^{-2}. EC saw the largest average reduction of 8.64%, followed by SC and CC, both of which experienced approximately 8% losses in PV power potential. As compared to the 1960s, the average PV power potential in the five most severely affected provinces (i.e., Henan, Zhejiang, Tianjin, Fujian, and Shanghai) decreased by 9.23–10.27% during the 2010s. Besides, a small positive change of 0.75% was observed in Yunnan.

Fig. 18 shows the spatial pattern of the seasonal PV power potential in China, as well as its trends from 1961 to 2016. As illustrated in Fig. 18a, the distribution of seasonal PV power potential is similar to the yearly potential, showing a spatial pattern of higher potential in the northwestern provinces, with the highest in Tibet, and lower potential in the southeastern provinces, with the lowest in Chongqing and the Sichuan basin. The average PV power potential in China for spring,

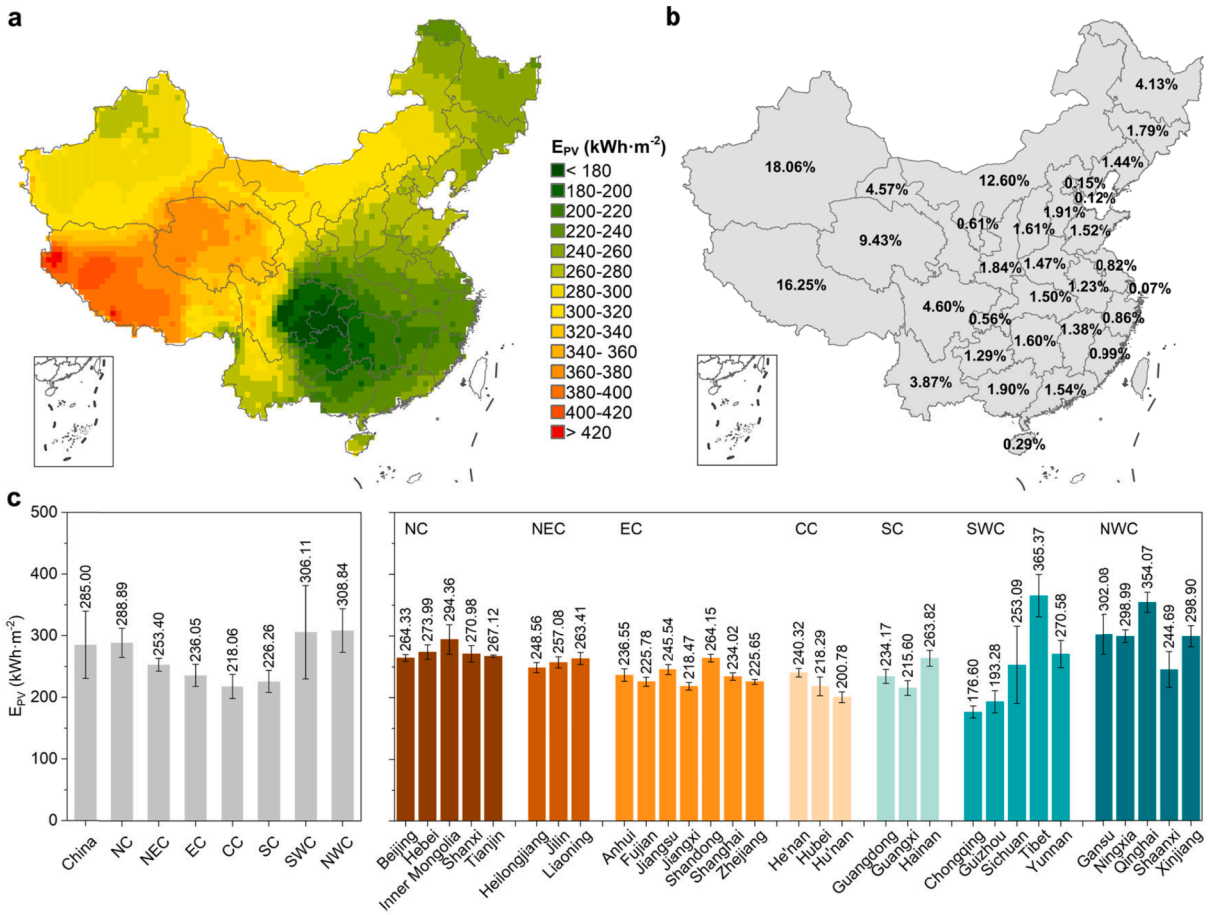


Fig. 15. Long-term PV power potential from 1961 to 2016 in China. (a) Spatial distribution of average yearly PV power potential across China. (b) Share of PV power potential in each province. (c) Average yearly PV power potential on regional and provincial levels.

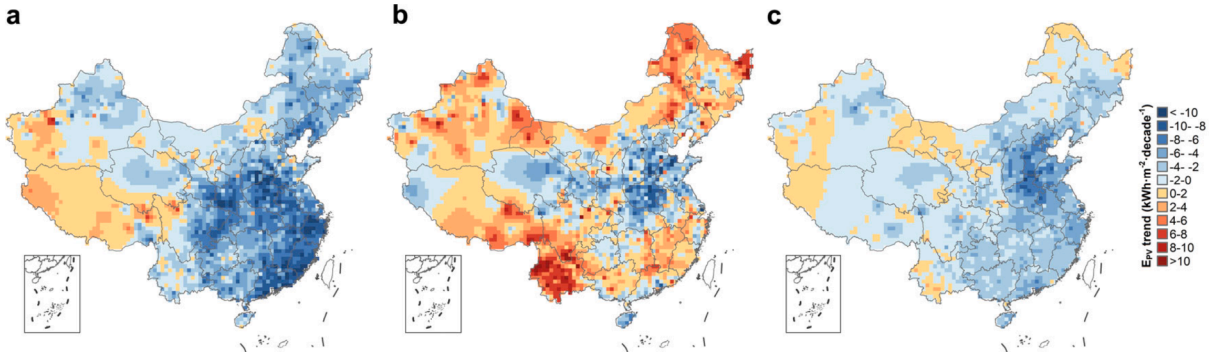


Fig. 16. Spatial pattern of PV power potential in China for (a) 1961–1990, (b) 1990–2016, and (c) 1961–2016.

summer, autumn, and winter was estimated to be 84.43, 90.12, 62.21, and 48.51 kWh·m^{-2}, respectively (see Fig. S7). For the long-term trend in seasonal PV power potential, the northwestern and eastern provinces show a clear increasing trend in the spring, with a slight nationwide increase of 0.03 kWh·m^{-2}·decade^{-1}. However, the national trend reversed in summer, autumn, and winter with decreases of −0.80, −0.38, and −0.43 kWh·m^{-2}·decade^{-1}, respectively (see Fig. 18b and Table S6). Also, most of the provinces within the NC, CC, and EC regions showed the most significant decreases in the summer, with a downward trend of more than −2 kWh·m^{-2}·decade^{-1}.

4. Conclusions

In-depth knowledge of solar radiation resources and assessment of solar PV potential is important for the implementation of solar energy projects. In this study, an interpretable machine learning model based on extreme gradient boosting optimized by the particle swarm optimization algorithm (PSO-XGBoost) was developed to estimate the global solar irradiance at locations without solar radiation measurements in China. Using the dataset reconstructed based on the PSO-XGBoost model, combined with GIS-based approaches and a general solar PV power model, comprehensive assessments of solar radiation resources and PV power potential in China were conducted, while their spatial patterns and spatiotemporal variability characteristics were elucidated.

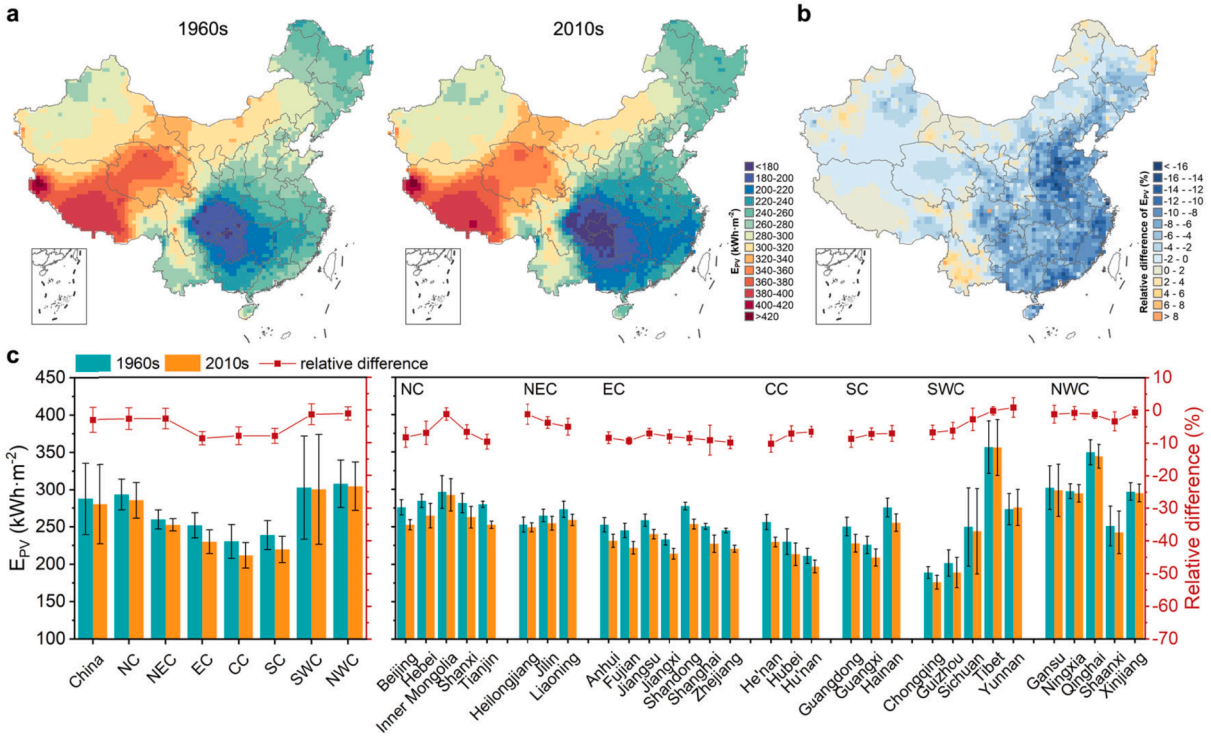


Fig. 17. PV power potentials in the 1960s and 2010s. (a) Spatial distribution of PV power potential. (b) The percentage difference between the 1960s and 2010s. (c) The regional and provincial averages in the 1960s and 2010s, as well as the relative differences.

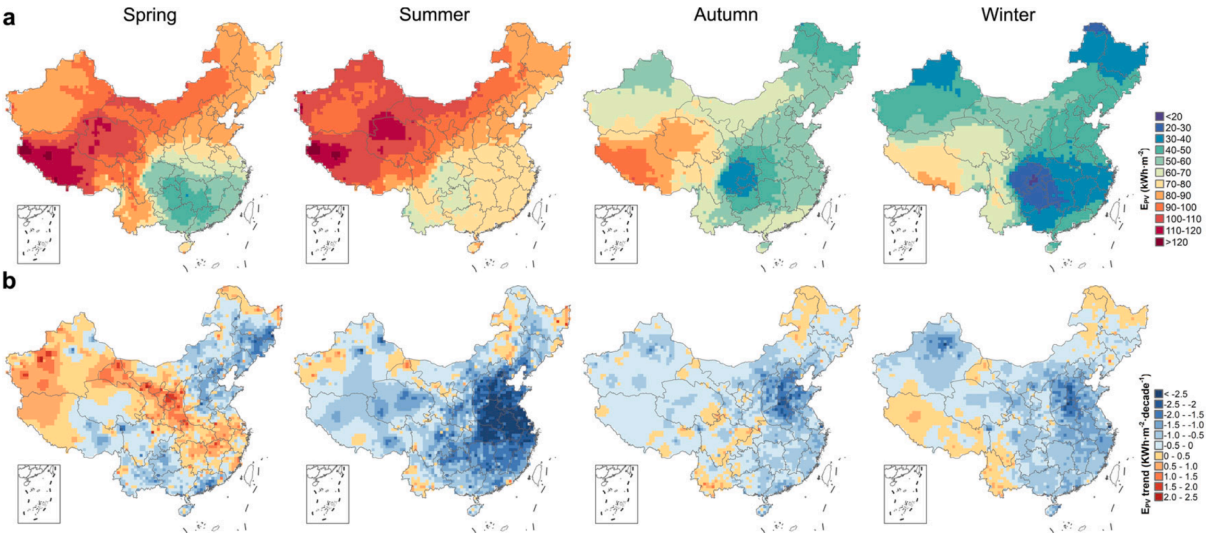


Fig. 18. The long-term seasonal pattern of PV power potential from 1961 to 2016 in China. (a) Spatial distribution of seasonal PV power potential. (b) Spatial distribution of the seasonal trends of PV power potential.

The main findings are summarized as follows:

(1) The developed PSO-XGBoost model showed the most superior accuracy and stability with overall R^2, RMSE, MAE, and MAPE of 0.953, 1.597 MJ·m^{-2}·day^{-1}, 1.138 MJ·m^{-2}·day^{-1}, and 10.500%, respectively, improving by 1.17–6.72%, 10.28–33.79%, 11.37–38.12%, and 16.67–54.94% compared with several other well-known machine learning models and Ångström-Prescott model. Besides, the Shapley additive explanations indicated that sunshine duration is the dominating feature and is linearly correlated with global horizontal solar radiation, as well as extraterrestrial solar radiation, as expected.

(2) The annual mean global solar radiation in China from 1961 to 2016 was estimated at 174.36 W·m^{-2}, with a decreasing trend of −0.83 W·m^{-2}·decade^{-1}. From the spatial pattern, northwestern provinces with high-quality solar radiation resources showed the least average decrease of −0.27 W·m^{-2}·decade^{-1}, while the provinces in and around the Beijing-Tianjin-Hebei region had the highest decreasing trend between −2.89 and −3.98 W·m^{-2}·decade^{-1}. Turning to seasonal variation, the most significant decreasing trend in solar radiation resources was observed in summer, reaching 1.83 W·m^{-2}·decade^{-1} during the period between 1961 and 2016.

(3) The average yearly potential for solar power generation in China from 1961 to 2016, assessed with global horizontal radiation data from the PSO-XGBoost model, reached 285.00 kWh·m^{-2}. Moreover, the PV power potential indicates a spatial pattern of higher potentials in the northwestern and northern provinces, with a maximum of 365.37 kWh·m^{-2} in Tibet, and lower values in the southeastern provinces, with a minimum of 176.60 kWh·m^{-2} in Chongqing. Xinjiang accounts for 18.06% of the national potential owing to plentiful solar resources and wide land areas.

(4) The yearly PV power potential in China decreased by 1.69 kWh·m^{-2}·decade^{-1} from 1961 to 2016. The provincial average PV power potential decreased by above 5 kWh·m^{-2}·decade^{-1} in the most heavily affected provinces such as Tianjin, Beijing, Henan, and Shandong. Also, the largest downward trend of −0.80 kWh·m^{-2}·decade^{-1} on average was observed in summer. Compared to the scenario of the 1960s, China's average yearly PV power potential in the 2010s was reduced by 2.88% from 287.55 to 280.21 kWh·m^{-2}, with 30 out of the 31 provinces showing a reduction between 0.25% and 10.27%. Eastern China saw the largest average reduction of 8.64%, followed by Southern China and Central China, both with about 8% losses in PV power potential.

(5) The PV sector in China faces a regional mismatch between PV power potential and installed PV capacity. Targeted development strategies adapted to local environmental, economic and resource characteristics are essential to put China's solar PV industry on a sustainable development track. For Xinjiang, Inner Mongolia, Qinghai, Gansu, and Tibet with huge PV power potential and sparse populations, it is most appropriate to prioritize the construction of large-scale centralized PV power plants, while the southeastern provinces should focus on developing distributed PV. Also, government policy guidance and support for regional grid interconnection and PV power trading between regional grids are crucial to developing an efficient solar PV power market.

In addition, this work still has limitations that should be addressed in future research. This study assumed that there would be long-term availability of solar resources at any location without considering the influences of geographical elements and engineering factors on solar radiation and PV power generation. Future works are expected to further assess the potential of PV power generation with high spatial and temporal resolution in engineering contexts. Moreover, this study aimed to assess China's long-term average solar radiation resources and PV power potential, as well as examine their spatiotemporal patterns. Considering future environmental changes and the increasing penetration of PV installations, China's future solar energy resources and PV power generation from a climate change perspective are worth further attention in future work to assist solar energy planners, policymakers and investors to make more informed decisions for long-term solar project deployment.

CRediT authorship contribution statement

Zhe Song: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Sunliang Cao:** Investigation, Writing – review & editing, Supervision. **Hongxing Yang:** Funding acquisition, Writing – review & editing, Supervision, Project administration.

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.

Data availability

The data that has been used is confidential.

Acknowledgments

The work described in this paper was financially supported by the research project (with project ID: P0035245) of The Hong Kong Polytechnic University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apenergy.2023.121005.

References

[1] IPCC. Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,. Cambridge, UK and New York, NY, USA: 2018. https://doi.org/10.1017/9781009157940.
[2] Liu J, Chen X, Yang H, Shan K. Hybrid renewable energy applications in zero-energy buildings and communities integrating battery and hydrogen vehicle storage. Appl Energy 2021;290:116733. https://doi.org/10.1016/j.apenergy.2021.116733.
[3] REN21. Renewables 2022 Global Status. Paris: 2022.
[4] IEA. World Energy Investment 2022. Paris, France: 2022.
[5] IEA. World Energy Outlook 2020. Paris: 2020.
[6] Liu J, Yang H, Zhou Y. Peer-to-peer energy trading of net-zero energy communities with renewable energy systems integrating hydrogen vehicle storage. Appl Energy 2021;298:117206. https://doi.org/10.1016/j.apenergy.2021.117206.
[7] Niveditha N, Rajan Singaravel MM. Optimal sizing of hybrid PV–Wind–Battery storage system for Net Zero Energy Buildings to reduce grid burden. Appl Energy 2022;324:119713. https://doi.org/10.1016/j.apenergy.2022.119713.
[8] LaPotin A, Schulte KL, Steiner MA, Buznitsky K, Kelsall CC, Friedman DJ, et al. Thermophotovoltaic efficiency of 40%. Nature 2022;604:287–91. https://doi.org/10.1038/s41586-022-04473-y.
[9] Zuo W, Chen Z, E J, Li Q, Zhang G, Huang Y. Effects of structure parameters of tube outlet on the performance of a hydrogen-fueled micro planar combustor for thermophotovoltaic applications. Energy 2023;266:126434. https://doi.org/https://doi.org/10.1016/j.energy.2022.126434.
[10] Zuo W, Wang Z, E J, Li Q, Cheng Q, Wu Y, et al. Numerical investigations on the performance of a hydrogen-fueled micro planar combustor with tube outlet for thermophotovoltaic applications. Energy 2023;263:125957. https://doi.org/https://doi.org/10.1016/j.energy.2022.125957.
[11] Liu J, Zhou Y, Yang H, Wu H. Net-zero energy management and optimization of commercial building sectors with hybrid renewable energy systems integrated with energy storage of pumped hydro and hydrogen taxis. Appl Energy 2022;321:119312. https://doi.org/10.1016/j.apenergy.2022.119312.
[12] Liu J, Ma T, Wu H, Yang H. Study on optimum energy fuel mix for urban cities integrated with pumped hydro storage and green vehicles. Appl Energy 2023;331:120399. https://doi.org/10.1016/j.apenergy.2022.120399.
[13] Perčić M, Vladimir N, Jovanović I, Koričan M. Application of fuel cells with zero-carbon fuels in short-sea shipping. Appl Energy 2022;309:118463. https://doi.org/10.1016/j.apenergy.2021.118463.
[14] Chen Z, Zuo W, Zhou K, Li Q, Huang Y, E j.. Multi-objective optimization of proton exchange membrane fuel cells by RSM and NSGA-II. Energy Convers Manag 2023;277:116691. https://doi.org/10.1016/j.enconman.2023.116691.
[15] Song Z, Liu J, Yang H. Air pollution and soiling implications for solar photovoltaic power generation: A comprehensive review. Appl Energy 2021;298:117247. https://doi.org/10.1016/j.apenergy.2021.117247.
[16] Song Z, Wang M, Yang H. Quantification of the Impact of Fine Particulate Matter on Solar Energy Resources and Energy Performance of Different Photovoltaic Technologies. ACS Environ Au 2022;2:275–86. https://doi.org/10.1021/acsenvironau.1c00048.
[17] IEA. World Energy Investment 2021. Paris, France: 2021.
[18] IEA. Snapshot of Global PV Markets 2022. Paris, France: 2022.
[19] Zell E, Gasim S, Wilcox S, Katamoura S, Stoffel T, Shibli H, et al. Assessment of solar radiation resources in Saudi Arabia. Sol Energy 2015;119:422–38. https://doi.org/10.1016/j.solener.2015.06.031.
[20] Solar AA. terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Q J R Meteorol Soc 1924;50:121–6. https://doi.org/10.1002/qj.49705021008.
[21] Prescott JA. Evaporation from a Water Surface in Relation to Solar Radiation. Trans R Soc South Aust 1940;46:114–8.
[22] Hassan GE, Youssef ME, Mohamed ZE, Ali MA, Hanafy AA. New Temperature-based Models for Predicting Global Solar Radiation. Appl Energy 2016;179:437–50. https://doi.org/10.1016/j.apenergy.2016.07.006.
[23] Deo RC, Şahin M, Adamowski JF, Mi J. Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over

Australia to forecast global solar radiation: A new approach. Renew Sustain Energy Rev 2019;104:235–61. https://doi.org/10.1016/j.rser.2019.01.009.
[24] Polo J, Wilbert S, Ruiz-Arias JA, Meyer R, Gueymard C, Súri M, et al. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Sol Energy 2016;132:25–37. https://doi.org/10.1016/j.solener.2016.03.001.
[25] Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: A review. Renew Energy 2017;105:569–82. https://doi.org/10.1016/j.renene.2016.12.095.
[26] Perez R, Kivalov S, Schlemmer J, Hemker K, Renné D, Hoff TE. Validation of short and medium term operational solar radiation forecasts in the US. Sol Energy 2010;84:2161–72. https://doi.org/10.1016/j.solener.2010.08.014.
[27] Verbois H, Saint-Drenan Y-M, Thiery A, Blanc P. Statistical learning for NWP post-processing: A benchmark for solar irradiance forecasting. Sol Energy 2022;238:132–49. https://doi.org/10.1016/j.solener.2022.03.017.
[28] 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 Manag 2021;235:113960. https://doi.org/10.1016/j.enconman.2021.113960.
[29] Yang D. A correct validation of the National Solar Radiation Data Base (NSRDB). Renew Sustain Energy Rev 2018;97:152–5. https://doi.org/10.1016/j.rser.2018.08.023.
[30] Yang D, Bright JM. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. Sol Energy 2020;210:3–19. https://doi.org/10.1016/j.solener.2020.04.016.
[31] Wang W, Yang D, Hong T, Kleissl J. An archived dataset from the ECMWF Ensemble Prediction System for probabilistic solar power forecasting. Sol Energy 2022;248:64–75. https://doi.org/10.1016/j.solener.2022.10.062.
[32] Bakker K, Whan K, Knap W, Schmeits M. Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation. Sol Energy 2019;191:138–50. https://doi.org/10.1016/j.solener.2019.08.044.
[33] Ghimire S, Deo RC, Casillas-Pérez D, Salcedo-Sanz S. Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms. Appl Energy 2022;316:119063. https://doi.org/10.1016/j.apenergy.2022.119063.
[34] Deo RC, Şahin M. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renew Sustain Energy Rev 2017;72:828–48. https://doi.org/10.1016/j.rser.2017.01.114.
[35] Bellido-Jiménez JA, Estévez Gualda J, García-Marín AP. Assessing new intra-daily temperature-based machine learning models to outperform solar radiation predictions in different conditions. Appl Energy 2021;298. https://doi.org/10.1016/j.apenergy.2021.117211.
[36] Eşlik AH, Akarslan E, Hocaoğlu FO. Short-term solar radiation forecasting with a novel image processing-based deep learning approach. Renew Energy 2022;200:1490–505. https://doi.org/10.1016/j.renene.2022.10.063.
[37] Chen J, Zhu W, Yu Q. Estimating half-hourly solar radiation over the Continental United States using GOES-16 data with iterative random forest. Renew Energy 2021;178:916–29. https://doi.org/10.1016/j.renene.2021.06.129.
[38] Ngoc-Lan Huynh A, Deo RC, Ali M, Abdulla S, Raj N. Novel short-term solar radiation hybrid model: Long short-term memory network integrated with robust local mean decomposition. Appl Energy 2021;298:117193. https://doi.org/10.1016/j.apenergy.2021.117193.
[39] Carneiro TC, Rocha PAC, Carvalho PCM, Fernández-Ramírez LM. Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain. Appl Energy 2022;314:118936. https://doi.org/10.1016/j.apenergy.2022.118936.
[40] Qiu R, Liu C, Cui N, Gao Y, Li L, Wu Z, et al. Generalized Extreme Gradient Boosting model for predicting daily global solar radiation for locations without historical data. Energy Convers Manag 2022;258:115488. https://doi.org/10.1016/j.enconman.2022.115488.
[41] Tong J, Xie L, Fang S, Yang W, Zhang K. Hourly solar irradiance forecasting based on encoder–decoder model using series decomposition and dynamic error compensation. Energy Convers Manag 2022;270:116049. https://doi.org/10.1016/j.enconman.2022.116049.
[42] Zhao S, Wu L, Xiang Y, Dong J, Li Z, Liu X, et al. Coupling meteorological stations data and satellite data for prediction of global solar radiation with machine learning models. Renew Energy 2022;198:1049–64. https://doi.org/10.1016/j.renene.2022.08.111.
[43] Feng Y, Gong D, Zhang Q, Jiang S, Zhao L, Cui N. Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Convers Manag 2019;198:111780. https://doi.org/10.1016/j.enconman.2019.111780.
[44] Ghimire S, Deo RC, Raj N, Mi J. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Appl Energy 2019;253:113541. https://doi.org/10.1016/j.apenergy.2019.113541.
[45] He G, Kammen DM. Where, when and how much solar is available? A provincial-scale solar resource assessment for China. Renew Energy 2016;85:74–82. https://doi.org/10.1016/j.renene.2015.06.027.
[46] Yan J, Yang Y, Elia Campana P, He J. City-level analysis of subsidy-free solar photovoltaic electricity price, profits and grid parity in China. Nat Energy 2019;4:709–17. https://doi.org/10.1038/s41560-019-0441-z.
[47] Wang Y, He J, Chen W. Distributed solar photovoltaic development potential and a roadmap at the city level in China. Renew Sustain Energy Rev 2021;141:110772. https://doi.org/10.1016/j.rser.2021.110772.
[48] Chen F, Yang Q, Zheng N, Wang Y, Huang J, Xing L, et al. Assessment of concentrated solar power generation potential in China based on Geographic Information System (GIS). Appl Energy 2022;315:119045. https://doi.org/10.1016/j.apenergy.2022.119045.
[49] Liu L, Wang Y, Wang Z, Li S, Li J, He G, et al. Potential contributions of wind and solar power to China's carbon neutrality. Resour Conserv Recycl 2022;180:106155. https://doi.org/10.1016/j.resconrec.2022.106155.
[50] Yang S, Wang XL, Wild M. Homogenization and trend analysis of the 1958–2016 in situ surface solar radiation records in China. J Clim 2018;31:4529–41. https://doi.org/10.1175/JCLI-D-17-0891.1.
[51] Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, et al. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Convers Manag 2018;164:102–11. https://doi.org/10.1016/j.enconman.2018.02.087.
[52] Zang H, Cheng L, Ding T, Cheung KW, Wang M, Wei Z, et al. Estimation and validation of daily global solar radiation by day of the year-based models for different climates in China. Renew Energy 2019;135:984–1003. https://doi.org/10.1016/j.renene.2018.12.065.
[53] Ahmed R, Sreeram V, Mishra Y, Arif MD. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew Sustain Energy Rev 2020;124:109792. https://doi.org/10.1016/j.rser.2020.109792.
[54] Mitrentsis G, Lens H. An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Appl Energy 2022;309:118473. https://doi.org/10.1016/j.apenergy.2021.118473.
[55] Chen T, Xgboost GC. A scalable tree boosting system. Proc. 22nd acm sigkdd Int. Conf Knowl Discov data Min 2016:785–94.
[56] Duffie JA, Beckman WA. Solar Engineering of Thermal Processes. 4th ed. New York: John Wiley & Sons inc.; 2013.
[57] Hassan MA, Akoush BM, Abubakr M, Campana PE, Khalil A. High-resolution estimates of diffuse fraction based on dynamic definitions of sky conditions. Renew Energy 2021;169:641–59. https://doi.org/10.1016/j.renene.2021.01.066.
[58] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[59] Eberhart R, Kennedy J. A new optimizer using particle swarm theory. MHS'95. Proc. sixth Int. Symp micro Mach Hum Sci, Ieee 1995:39–43.
[60] Kennedy J, Eberhart R. Particle swarm optimization. Proc. ICNN'95-international Conf. neural networks. IEEE 1995;4:1942–8.
[61] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018;51:1–42.
[62] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res 2021;70:245–317.
[63] Python A, Bender A, Nandi AK, Hancock PA, Arambepola R, Brandsch J, et al. Predicting non-state terrorism worldwide. Sci Adv 2022;7:eabg4778. https://doi.org/10.1126/sciadv.abg4778.
[64] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[65] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67.
[66] Evans DL. Simplified method for predicting photovoltaic array output. Sol Energy 1981;27:555–60. https://doi.org/10.1016/0038-092X(81)90051-7.
[67] Notton G, Cristofari C, Mattei M, Poggi P. Modelling of a double-glass photovoltaic module using finite differences. Appl Therm Eng 2005;25:2854–77. https://doi.org/10.1016/j.applthermaleng.2005.02.008.
[68] Skoplaki E, Boudouvis AG, Palyvos JA. A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting. Sol Energy Mater Sol Cells 2008;92:1393–402. https://doi.org/10.1016/j.solmat.2008.05.016.
[69] Zhang Y, Ma T, Elia Campana P, Yamaguchi Y, Dai Y. A techno-economic sizing method for grid-connected household photovoltaic battery systems. Appl Energy 2020;269:115106. https://doi.org/10.1016/j.apenergy.2020.115106.
[70] Lu GY, Wong DW. An adaptive inverse-distance weighting spatial interpolation technique. Comput Geosci 2008;34:1044–55. https://doi.org/10.1016/j.cageo.2007.07.010.
[71] Burrough PA, McDonnell RA, Lloyd CD. Principles of geographical information systems. Oxford University Press; 2015.
[72] He C, Liu J, Xu F, Zhang T, Chen S, Sun Z, et al. Improving solar radiation estimation in China based on regional optimal combination of meteorological factors with machine learning methods. Energy Convers Manag 2020;220:113111. https://doi.org/10.1016/j.enconman.2020.113111.
[73] Liu F, Wang X, Sun F, Wang H. Correct and remap solar radiation and photovoltaic power in China based on machine learning models. Appl Energy 2022;312:118775. https://doi.org/10.1016/j.apenergy.2022.118775.
[74] Prieto J-I, García D. Modified temperature-based global solar radiation models for estimation in regions with scarce experimental data. Energy Convers Manag 2022;268:115950. https://doi.org/10.1016/j.enconman.2022.115950.
[75] El Mghouchi Y. On the prediction of daily global solar radiation using temperature as input. An application of hybrid machine learners to the six climatic Moroccan zones. Energy Convers Manag X 2022;13:100157. https://doi.org/10.1016/j.ecmx.2021.100157.
[76] Wang Y, Wild M. A new look at solar dimming and brightening in China. Geophys Res Lett 2016;43:11,711-777,785. https://doi.org/https://doi.org/10.1002/2016GL071009.
[77] Wang K, Ma Q, Li Z, Wang J. Decadal variability of surface incident solar radiation over China: Observations, satellite retrievals, and reanalyses. J Geophys Res Atmos 2015;120:6500–14. https://doi.org/10.1002/2015JD023420.

[78] He Y, Wang K, Zhou C, Wild M. A Revisit of Global Dimming and Brightening Based on the Sunshine Duration. Geophys Res Lett 2018;45:4281–9. https://doi.org/10.1029/2018GL077424.

[79] Li X, Wagner F, Peng W, Yang J, Mauzerall DL. Reduction of solar photovoltaic resources due to air pollution in China. Proc Natl Acad Sci U S A ;114:11867–72. https://doi.org/10.1073/pnas.1711462114.

[80] National Energy Administration of China. China's solar PV capacity and annual additions in 2021 n.d. http://www.nea.gov.cn/2022-03/09/c_1310508114.htm (accessed December 15, 2022).