Article | Received 23 June 2025; Revised 12 September 2025; Accepted 16 October 2025; Published 22 October 2025
https://doi.org/10.55092/sc20250026

A minute-level interpretable solar irradiance machine learning prediction method for low-carbon building calculation

Yanyun Zhang[1,2], Runze Shi[2], Yupeng Wu[3] and Peng Xue[1,2,4,*]

[1] Beijing Key Laboratory of Green Built Environment and Energy Efficient Technology, Beijing University of Technology, Beijing, China
[2] College of Architecture and Civil Engineering, Beijing University of Technology, Beijing, China
[3] Faculty of Engineering, University of Nottingham, Nottingham, UK
[4] Chongqing Research Institute of Beijing University of Technology, Chongqing, China

* Correspondence author; E-mail: xp@bjut.edu.cn.

Highlights:

- An ensemble learning approach for solar irradiance prediction is proposed.
- Easily accessible meteorological parameters are used as inputs to the model.
- The contribution of the input features is discussed using interpretation techniques.
- The Bayesian optimization algorithm is used to tune the model hyperparameters.
- The proposed model is verified to outperform other state-of-the-art models.

Abstract: High-temporal-resolution solar irradiance data are essential for calculating and assessing high-performance buildings. However, limited access to measurement equipment often restricts the availability of such data. To address this challenge, this study proposed a highly accurate, interpretable, and convenient method for ultra-short-term global horizontal irradiance (GHI) prediction. Firstly, a dataset containing six types of conventional meteorological parameters and corresponding irradiance values was prepared, and its feasibility for model development was investigated through correlation analysis. Then, the eXtreme Gradient Boosting (XGBoost) model, combined with Bayesian optimization (BO) algorithm, was developed to predict 1-minute GHI based on the selected meteorological parameters. Finally, the prediction mechanism was revealed by analyzing the feature importance and the effect of key features using interpretation techniques. The results show that the BO-XGBoost model outperforms the other state-of-the-art models, with coefficient of determination (R^2) of 0.907 and root mean square error ($RMSE$) of 76.199, especially in clear sky conditions, the R^2 and $RMSE$ can be 0.990 and 24.077. The model interpretation results further indicate that GHI prediction heavily relies on the solar elevation angle and relative humidity, along with their interactions. This study provides a cost-effective solution for obtaining irradiance data critical for designing and optimizing solar-based low-carbon buildings.

Keywords: solar irradiance; ultra-short-term; interpretable machine learning; prediction model; building calculation; Bayesian optimization

1. Introduction

1.1. Significance of solar irradiance prediction

As global warming intensifies, countries have ramped up efforts to tackle climate change driven by carbon emissions. Many countries have pledged to achieve net-zero carbon emissions by 2050. Among various sectors, the building sector accounts for approximately 40% of global energy consumption and about 30% of carbon dioxide emissions [1]. Promoting energy management in buildings has therefore become a key strategy for carbon reduction and decarbonization.

Solar energy is one of the most promising clean energy sources for integrating renewables into buildings, with the potential to reduce electricity consumption by 20% to 50% [2]. Reliable solar irradiance data is essential for evaluating solar-driven building performance and for predicting and managing energy utilization [3]. Such data are typically obtained from typical meteorological year (TMY) datasets or standardized meteorological files [4,5]. However, due to the significant spatiotemporal variability of solar irradiance, these data files often fail to accurately reflect local irradiation conditions [5], potentially leading to significant errors in building energy calculations. In addition, with the promotion of net-zero energy buildings, especially with the widespread deployment of building integrated photovoltaic (BIPV) technology [6], building calculations are placing greater demands on the temporal resolution of radiation data. Consequently, acquiring accurate, localized, and ultra-short-term solar irradiance data has become critical for optimizing building energy performance and advancing low-carbon, zero-carbon, and smart building technologies.

Despite this need, the large-scale deployment of solar irradiance measurement equipment remains limited due to high costs and complex maintenance requirements, especially in remote or rural regions [7,8]. In China, only a small number of meteorological stations record solar radiation data [9,10], and even in developed countries, stations capable of high-temporal-resolution radiation monitoring are relatively scarce [11]. As a result, the development of irradiance prediction models has emerged as an effective and economical alternative.

1.2. Available prediction methodologies

Solar radiation variability is influenced by a multitude of dynamic meteorological and environmental factors, which are usually adequately incorporated into corresponding prediction models. These models are categorized into three types based on different classification criteria.

1.2.1. Input type-based classification

Solar irradiance prediction methods can be classified into four categories based on input factors: satellite or sky imagery-based methods, historical irradiance-based methods, meteorological data-based methods, and methods combining multiple data sources from these categories. The first usually requires specialized equipment for image capture, followed by image analysis [12]. As presented in studies [13–15], the models primarily utilize sky images as input variables, sometimes supplemented with additional data sources.

However, the complexity of cloud properties, the cost of image acquisition, and the tediousness of the image data processing steps pose challenges to this method [16], limiting regional applicability and predict accuracy enhancement. The second relies on extensive historical irradiance data across various time scales [17,18], occasionally integrating other inputs such as sky images [19,20]. Availability of historical data is constrained by equipment limitations and other factors, restricting their widespread application. In contrast, meteorological data-based methods leverage real-time, detailed, and accessible meteorological information, making them particularly valuable in resource-constrained environments [7,21]. For instance, Urraca *et al.* achieved one-hour-ahead solar irradiance predictions using meteorological records and solar calculations [12]. Despite achieving a root mean square error ($RMSE$) of approximately 100 W/m^2, there remains room for accuracy improvement. Li *et al.* developed BP and CNN models using meteorological parameters, achieving $RMSE$ values below 70 W/m^2 for irradiance predictions in Shanghai and Xi'an [9]. The accuracy of these models has also been validated in simulations of building energy consumption and PV power generation.

1.2.2. Method-based classification

In terms of prediction methods, models can be broadly classified into statistical, physical, machine learning. Statistical models usually leverage historical time series data and are adept at capturing linear relationships [11]. However, their performance diminishes when dealing with problems that are nonlinear or influenced by external factors [22]. Physical models, also known as radiative transfer models [8], involve complex and rigorous radiative transfer boundaries that require substantial computational resources and time [23]. Numerical weather prediction models, a type of radiative transfer model [24], are primarily suitable for longer-term forecasts (greater than 6 hours and up to several days) [21]. With the advancement of artificial intelligence, machine learning methods have gained increasing importance in various prediction tasks. These methods are capable of learning from datasets and establishing nonlinear mappings between input and output variables without the need for explicit programming [21], thereby reducing prediction errors and enhancing fitting accuracy. In particular, ensemble algorithms have contributed to the development of increasingly robust and efficient models that achieve a favorable balance between simplicity and performance [22]. For instance, Fan *et al.* found that the XGBoost model excels in estimating daily irradiance based on temperature and precipitation data, particularly in humid subtropical climates [25]. Hassan *et al.* demonstrated that using gradient boosting (GB), random Forest (RF), and bagging to estimate daily global horizontal irradiance (GHI) in the Middle East and North Africa resulted in more stable and accurate performance compared to support vector regression (SVR) and artificial neural networks (ANN) [11]. Lee *et al.* compared several ensemble learning (EL) algorithms, including boosted trees (BS), bagged trees (BG), RF, and generalized random forest (GRF), with single regression methods such as support vector machine (SVM) and Gaussian process regression (GPR), concluding that the ensemble methods provided better prediction performance [26].

1.2.3. Time interval-based classification

Solar irradiance prediction models are classified based on different prediction intervals, ranging from ultra-short-term to long-term predictions. Ultra-short-term predictions cover intervals from 1 second to 1 hour, focusing on capturing rapid fluctuations in irradiance [27]. These predictions are crucial for

applications requiring real-time adjustments, such as PV system control and grid management [28,29] and building energy simulation [3]. Despite their importance, research in this area remains limited compared to longer-term predictions [30,31]. Few studies have explored predictions at hourly [9,17] and intra-hourly [32,33] scales, highlighting challenges in achieving high accuracy due to the rapid variability of solar energy. For instance, Bhatt *et al.* developed three deep learning (DL) models for solar irradiance prediction with lead times ranging from 15 minutes to 1 hour and 30 minutes [34]. Nunes Maciel *et al.* utilized ANN and LightGBM models to predict GHI at intervals ranging from 1 to 60 minutes, revealing limitations in accurately predicting GHI at 1-minute intervals [13].

1.3. Objectives of this study

Solar irradiance prediction is essential for enhancing the precision of low-carbon building performance calculations. Although existing research has made notable progress, several key challenges remain insufficiently addressed: (i) Ultra-short-term irradiance prediction models with high temporal resolution still require further development and improvement; (ii) Existing approaches often rely on costly or complex data sources, while the potential of using only conventional meteorological parameters remains underexplored; (iii) Model interpretability remains insufficiently addressed, with few studies focused on explaining the underlying prediction mechanisms or identifying the most influential input variables.

To address the above research gaps, this study proposed a high-accuracy and interpretable model for ultra-short-term solar irradiance prediction. The main contributions of this study are as follows: (i) Minute-level global horizontal irradiance (GHI) forecasting is conducted by applying a Bayesian optimization strategy to the XGBoost model, improving prediction accuracy at this fine time scale compared with methods primarily focusing on hourly or longer intervals; (ii) The framework relies exclusively on six conventional meteorological variables that are widely available across most regions, thereby demonstrating a practical and cost-effective alternative to approaches requiring costly or complex data sources; (iii) Integrated interpretability techniques were employed to enhance the transparency and credibility of the model. Analyses based on SHAP and PDP provided new insights into the relative importance of meteorological variables. These contributions not only improve the accuracy of minute-level irradiance forecasting but also provide practical insights into variable importance, thereby supporting solar-based low-carbon building design and energy system optimization.

2. Methodology

This section systematically introduces the methods for developing, evaluating and interpreting the prediction models, including data collection and analysis, ensemble learning (EL) algorithms, hyperparameter optimization algorithm, model performance metrics, and interpretation methods. The framework and detailed workflow for this study are shown in Figure 1.

Firstly, the meteorological and irradiance data were prepared, and their temporal distribution patterns and correlations were analyzed. Secondly, the prediction models were developed using EL algorithms coupled with Bayesian optimization with meteorological data as inputs and irradiance as outputs, and the performance of the proposed model was evaluated and compared. Finally, the interpretation techniques were employed to identify and determine the importance and effect of input features in the prediction process.

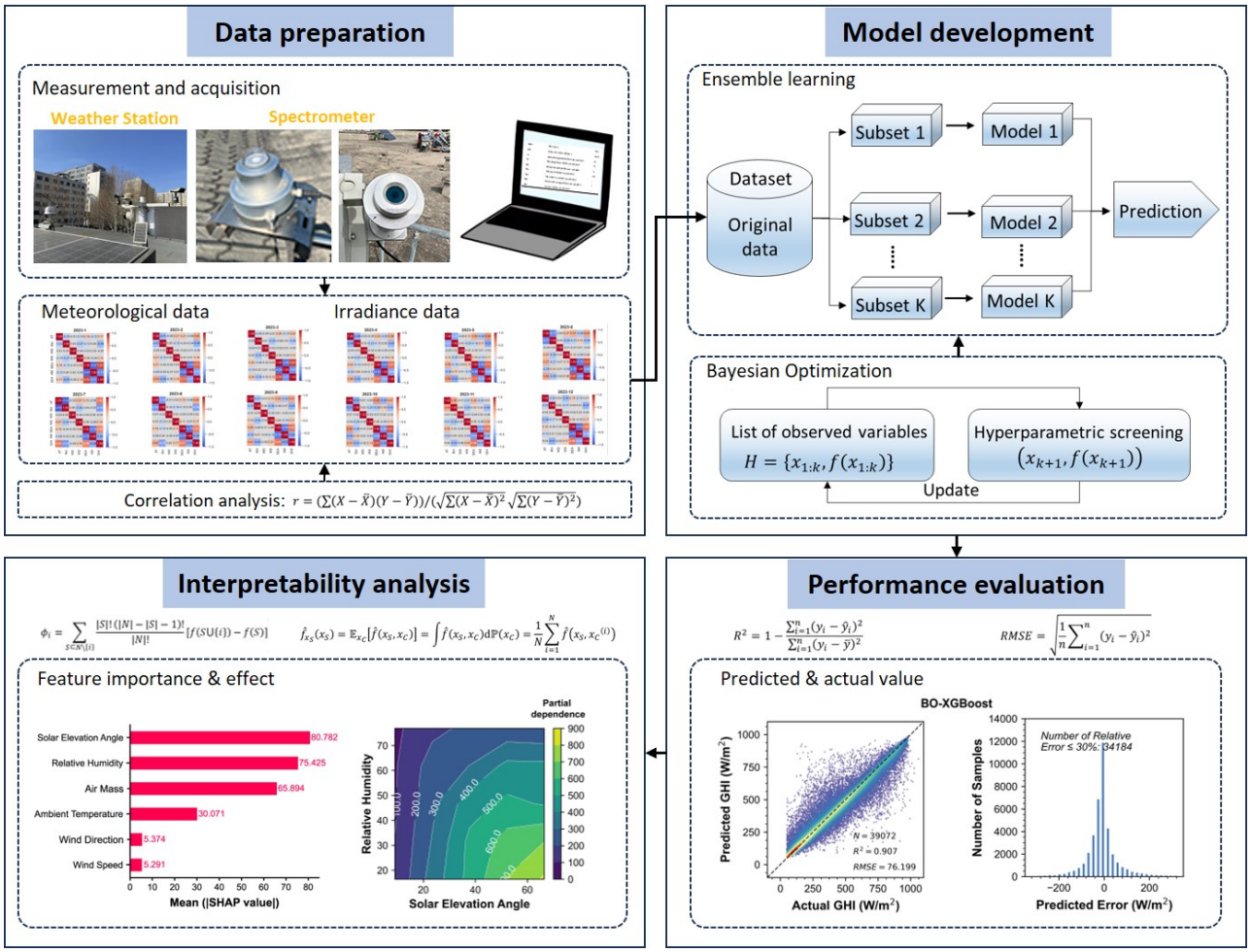

Figure 1. The framework and workflow of this study.

2.1. Dataset description

The measurement site is located in Beijing, situated in the northern part of China's North China Plain, with a temperate monsoon climate and four distinct seasons [19]. The dataset spans from January to December 2023 with a sampling interval of 1 minute, comprising meteorological data and global horizontal irradiance (GHI) data, continuously recorded throughout the year. To ensure data quality, irradiance values below 50 W/m² were excluded, as they mainly occurred during sunrise, sunset, and extreme weather conditions, when the measurements are often affected by higher signal-to-noise ratios. After filtering, a total of 195,357 GHI samples and their corresponding meteorological parameters were retained for model construction.

In this study, the spectrally integrated GHI over the 280–2500 nm range was defined as the model output, as this wavelength interval contains the vast majority of solar energy relevant to photovoltaic performance and building applications. The model inputs consisted of six conventional meteorological parameters: ambient temperature (AT), relative humidity (RH), wind speed (WS), wind direction (WD), solar elevation angle (SEA), and air mass (AM) [21,26,35–38]. Among these, AT, RH, WS, and WD were directly measured, while SEA and AM were calculated from geographic and temporal information. During the modelling process, 80% of the entire dataset was randomly selected for model training and the remaining 20% for testing [10,11,35].

Before formally determining these as effective variables for model development, Pearson correlation analysis was conducted to assess the correlations between the input variables and between the input and output variables, as shown in Equation (1). After performing the correlation analysis, the variables will be used for model development, as shown in Table 1 and Figure 2a.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}} \quad (1)$$

where X and Y represent the values of the two variables, $\bar{X}$ and $\bar{Y}$ represents the mean value of X and Y, respectively.

Table 1. Input and output variables for prediction models.

Role	Name	Description	Unit
Input	AT	Ambient temperature	°C
	RH	Relative humidity	%
	WS	Wind speed	m/s
	WD	Wind direction	/
	SEA	Solar elevation angle	°
	AM	Air mass	/
Output	GHI	Global horizontal irradiance	W/m^2

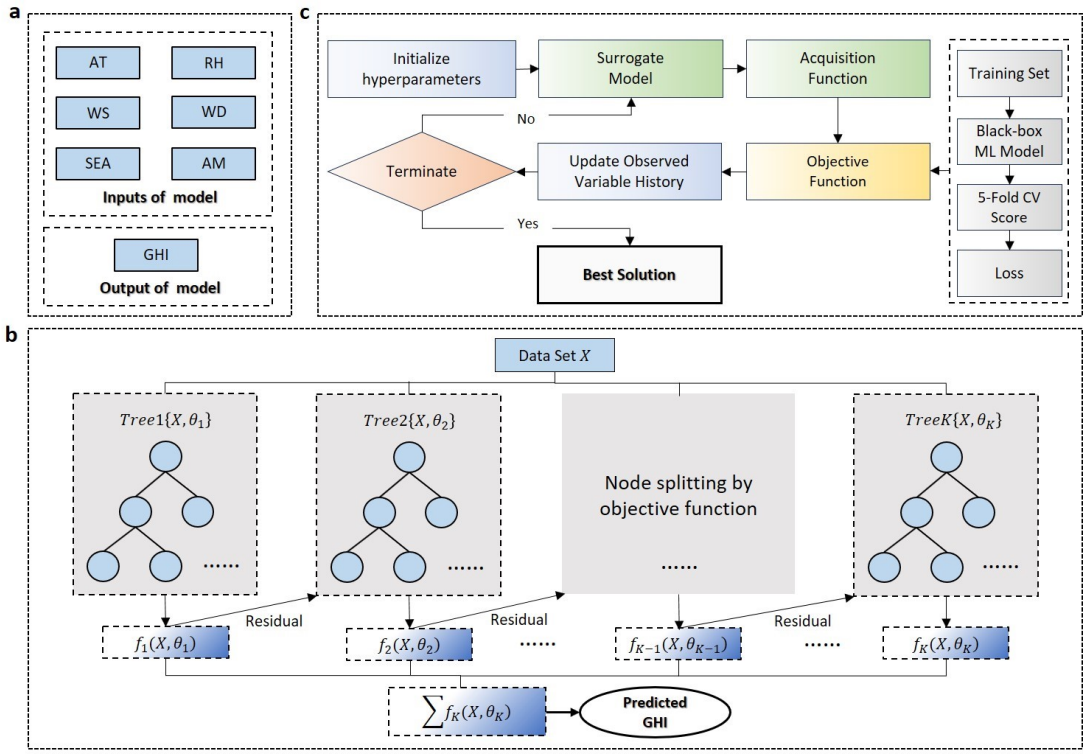

Figure 2. Flowchart of the proposed BO-XGBoost model.

2.2. Ensemble learning algorithm

Ensemble learning (EL) enhances model prediction performance by combining multiple weak learners to form a robust model structure [36]. In this study, the eXtreme Gradient Boosting (XGBoost) was selected as the primary model due to its strong capability in handling nonlinear relationships and high computational efficiency, which are particularly important for ultra-short-term irradiance forecasting, as

illustrated in Figure 2b. For comparison, three other representative ensemble learning algorithms, Gradient Boosting Decision Tree (GBDT), Random Forest (RF), and Categorical Boosting (CatBoost) prediction models, were also implemented to benchmark the proposed approach.

The GBDT algorithm, proposed by Friedman, is a boosting-based method that iteratively reduces prediction residuals through gradient descent optimization, and has been widely recognized for its strong predictive performance [11]. Random Forest (RF), introduced by Breiman, is based on bagging and random feature selection, which reduces correlation between trees and improves generalization [39]. CatBoost, developed by Yandex, is an advanced variant of gradient boosting that incorporates ordered boosting and efficient handling of categorical variables to avoid overfitting and target leakage [22]. XGBoost, proposed by Chen, is a scalable and efficient boosting algorithm that supports parallel tree construction and regularization, enabling fast training and robust prediction [40].

2.3. Hyperparameter optimization

For machine learning models, especially tree-based and neural network algorithms, hyperparameters play a critical role in degerming model performance. To enhance model performance, it is essential to identify the optimal hyperparameters for each model. This is achieved through hyperparameter tuning and optimization, a process involving the exploration and testing the hyperparameter space [36].

Most existing studies on solar energy prediction have adopted grid search or random search for hyperparameter tuning. However, these methods are computationally expensive and often fail to ensure global optimality, especially when dealing with complex models containing a large number of hyperparameters. This limitation is even more pronounced in minute-level forecasting, where models must rapidly adapt to highly dynamic irradiance fluctuations. To address this issue, this study employs the Bayesian optimization (BO) algorithm, as shown in Figure 2c, to efficiently identify the optimal hyperparameter configurations for XGBoost and other ensemble models. BO iteratively proposes new candidate configurations based on prior evaluations, thereby enabling efficient exploration of the hyperparameter space. Compared with grid search and random search, BO significantly reduces computational cost and improves search efficiency, resulting in more effective and stable tuning. The core framework of the BO algorithm is sequential model-based optimization (SMBO), with the surrogate model and acquisition function as the two key components that distinguish different BO methods. The overall steps for performing SMBO are as follows [41]:

Step 1. Based on the existing tuning history $H = \{x_{1:k}, f(x_{1:k})\}$, a probabilistic model $y(x)$ is established. Specifically, in a given domain space, a random selection of hyperparameters $x_{1:k}$ and their associated objective function values $f(x_{1:k})$ is used to form the observation variables list $H = \{x_{1:k}, f(x_{1:k})\}$. Subsequently, the probability distribution of the observation variables H is modeled using the Tree Parzen Estimator (TPE) algorithm. For the observation points x on either side of a specific threshold y^*, different probability distributions are constructed, as shown in Equation (2).

$$p(x|y) = \begin{cases} l(x), y < y^* \\ g(x), y \geq y^* \end{cases} \tag{2}$$

where y^* is the threshold corresponding to the quantile γ of y ($p(y < y^*) = \gamma$), and $l(x)$ and $g(x)$ represent the probability distributions of the hyperparameters for the objective function values below and above the threshold y^*, respectively.

Step 2. The next hyperparameter combination x_{k+1} is selected according to the acquisition function. Utilizing the hyperparameter probability distributions described above, the Expected Improvement (EI) strategy is employed to find the next optimal hyperparameter combination. The calculation process is shown in Equation (3).

$$EI_{y^*}(x) = \frac{\int_{-\infty}^{y^*}(y^* - y)p(y)dy}{\gamma + (1-\gamma)\frac{g(x)}{l(x)}} \tag{3}$$

where the value of $EI_{y^*}(x)$ depends solely on the ratio of the probabilities $\frac{g(x)}{l(x)}$ once γ is determined ($EI_{y^*}(x) \propto \left(\gamma + (1-\gamma)\frac{g(x)}{l(x)}\right)^{-1}$). This means that in the next round of search, the goal is to find the candidate hyperparameter combination x_{k+1} that minimizes this ratio.

Step 3. Update the existing observation variables list H. Calculate the value of the objective function $f(x_{k+1})$ corresponding to the hyperparameters x_{k+1} and add the new observation $\left(x_{k+1}, f(x_{k+1})\right)$ to the existing observations list H.

Step 4. Repeat Steps 1 to 3 with the updated observations list H until the predefined number of iterations is completed.

2.4. Performance metrics

Errors are an inevitable issue of prediction models. In this study, the coefficient of determination (R^2) and the root mean square error ($RMSE$) were selected to evaluate the model performance. R^2 represents the strength of the linear relationship between the actual and predicted values and is a widely used goodness-of-fit statistic. $RMSE$, which is the square root of the average of the squared differences between the actual and predicted value, helps to identify and eliminate outliers in the data. It is one of the most reliable and popular performance evaluation metrics. In general, good model performance is indicated by higher R^2 and lower $RMSE$. The calculations for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of all actual values.

2.5. Interpretation techniques

To enhance model interpretability, this study adopted interpretable machine learning techniques to explain the model's decision-making process and prediction results. By understanding the model's prediction mechanisms and revealing potential adversarial perturbations that may alter the model output, the credibility and controllability of the model can be improved [42].

2.5.1. SHapley Additive exPlanations

The SHapley Additive exPlanations (SHAP) technique, proposed by Lundberg and Lee, is an interpretable machine learning method [43] rooted in the principles of game theory. By constructing an

additive explanation model, SHAP fairly distributes feature contributions, overcoming the issue of dependence on feature order. This method quantifies the importance of each input variable in making predictions, enabling the interpretation of outputs from any machine learning model. By calculating SHAP values, individual sample prediction, overall model prediction, and feature selection can be explained. The calculation is based on Equation (6).

$$\phi_i = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{6}$$

where i represents the feature of a given sample vector, ϕ_i is the importance value of feature i to the model, N is the set of all features, $N \backslash \{i\}$ is the set excluding feature i, S is a subset containing any number of features, $|S|$ is the number of features in subset S, and $f(*)$ is the output of the model.

2.5.2. Partial Dependence Plots

The Partial Dependence Plots (PDP) technique, proposed by Friedman *et al.*, is an interpretable machine learning method that reveals the linear or nonlinear relationships between one or more feature inputs and the model output, thereby helping to explain the behavior of complex models. By conditionally averaging the target feature and marginalizing the non-target features, the PDP technique quantifies the average impact of the target feature on the model output.

In this study, one-dimensional PDP (1D-PDP) and two-dimensional PDP (2D-PDP) were used to analyze the effect of a single target feature and the interaction of two target features, respectively, on the output of the machine learning model. In PDP, the partial dependence function $\hat{f}_{x_S}(x_S)$ for the target feature x_S is defined as the expectation over the marginal distribution of the non-target features x_C, as shown in Equation (7).

$$\hat{f}_{x_S}(x_S) = \mathbb{E}_{x_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) \mathrm{d}\mathbb{P}(x_C) = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(x_S, {x_C}^{(i)}) \tag{7}$$

where N is the number of samples in the dataset, and ${x_C}^{(i)}$ is the i-th sample value of the non-target feature x_C.

3. Results and discussion

3.1. Data analysis

3.1.1. Data distribution

Table 2 presents the monthly means and standard deviations of meteorological inputs and GHI outputs used in the development of the subsequent prediction model. The data indicate that the ambient temperature (AT) is relatively low during the winter months (January to February, November to December) and increases gradually through the spring and summer (March to October). Relative humidity (RH) varies significantly between months but is generally higher in summer (July to September). Wind direction (WD) and wind speed (WS) show slight variations across different months but remain relatively stable overall. In addition, there are significant differences in solar elevation angle (SEA), air mass (AM) and global horizontal irradiance (GHI) values across the months.

Table 2. Monthly descriptive statistics for input and output variables.

Variable	Mean (Standard Deviation)												
	2023-01	2023-02	2023-03	2023-04	2023-05	2023-06	2023-07	2023-08	2023-09	2023-10	2023-11	2023-12	All months
AT	1.097	3.407	11.896	15.654	22.85	29.421	29.259	28.238	24.342	18.513	6.181	−1.347	17.533
	(4.594)	(3.692)	(5.295)	(4.494)	(4.229)	(5.114)	(5.026)	(4.265)	(3.886)	(3.546)	(6.88)	(6.693)	(11.327)
RH	28.433	36.211	27.946	36.044	43.917	37.195	53.93	57.571	55.548	40.958	37.882	38.572	41.318
	(9.069)	(15.962)	(14.448)	(17.005)	(17.474)	(16.239)	(20.109)	(16.314)	(16.143)	(14.481)	(15.102)	(14.231)	(18.48)
WD	167.29	167.279	163.435	156.91	159.184	155.754	155.091	154.608	155.413	155.823	158.319	158.388	158.557
	(114.1)	(103.273)	(104.759)	(98.486)	(96.73)	(101.438)	(105.199)	(107.814)	(102.686)	(111.129)	(115.915)	(111.644)	(105.394)
WS	1.248	1.188	1.575	1.623	1.299	1.381	1.325	1.17	1.054	1.13	1.577	1.275	1.326
	(1.321)	(1.116)	(1.39)	(1.481)	(1.151)	(1.132)	(1.116)	(0.909)	(0.876)	(0.971)	(1.502)	(1.305)	(1.208)
SEA	20.499	26.186	32.103	39.264	44.047	43.152	45.496	42.512	35.113	27.395	21.475	19.505	34.611
	(7.261)	(9.662)	(13.106)	(16.328)	(18.809)	(20.656)	(19.35)	(16.01)	(14.157)	(11.03)	(8.232)	(6.73)	(17.376)
AM	3.504	2.853	2.493	2.111	1.943	2.111	1.898	1.847	2.297	3.074	3.869	3.705	2.522
	(2.071)	(1.889)	(1.81)	(1.493)	(1.379)	(1.608)	(1.373)	(1.147)	(1.713)	(3.357)	(4.602)	(2.45)	(2.253)
GHI	302.443	339.51	419.835	417.186	461.43	508.247	435.366	436.332	384.197	386.862	308.117	283.69	403.748
	(143.843)	(195.19)	(222.428)	(269.314)	(284.186)	(295.997)	(296.453)	(270.081)	(228.242)	(197.815)	(165.176)	(146.813)	(249.576)

Figure 3 visually displays the GHI distribution for each month, revealing a strong correlation between GHI values and the time of day: values are lower in the morning, peak at noon, and gradually decrease in the afternoon. It is also evident that GHI values are lower in winter and higher in summer. These patterns in meteorological and irradiance data align with the seasonal climate characteristics of Beijing, which experiences cold and dry conditions in autumn and winter, and warm and humid conditions in spring and summer.

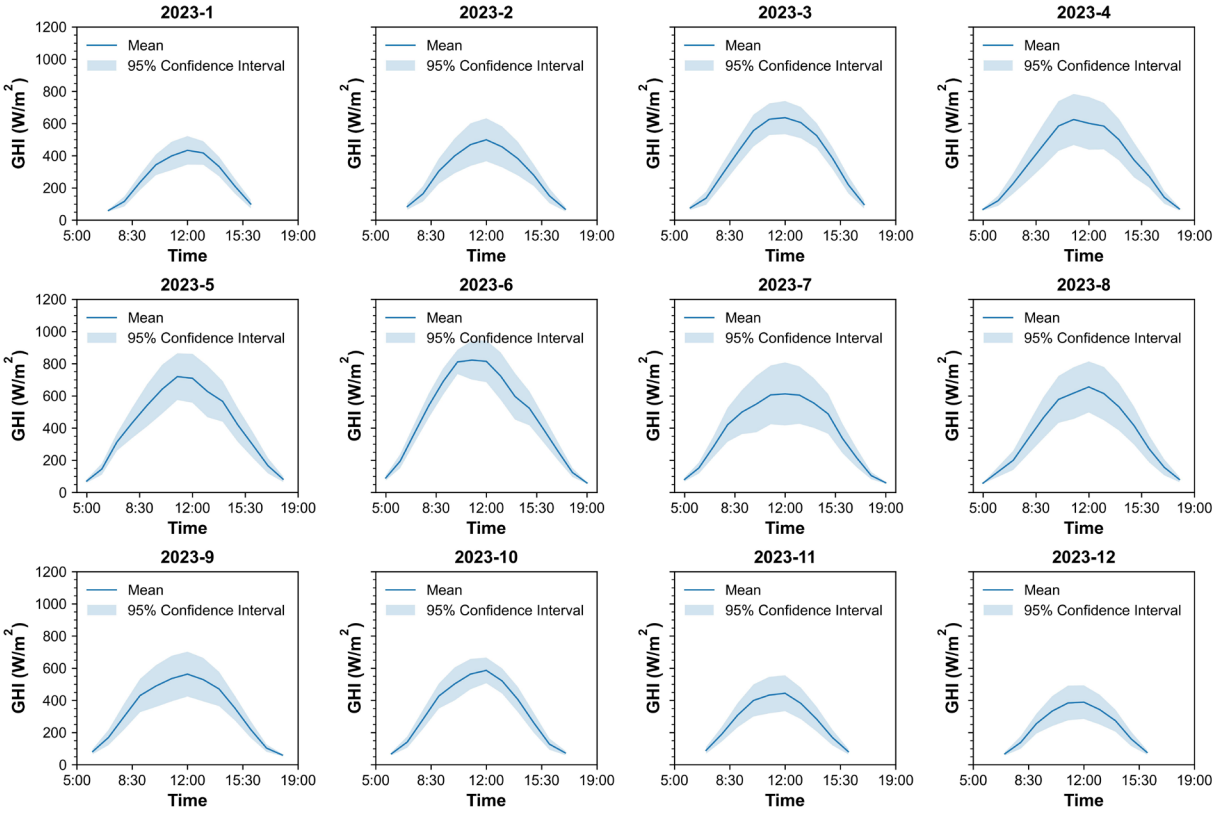

Figure 3. Distribution of GHI at 1-minute intervals throughout the year.

3.1.2. Correlation between data variables

This section conducts an exploratory analysis of the correlations between each input as well as between the inputs and outputs. This analysis helps to determine whether solar irradiance can be predicted from a single weather indicator and how it depends on the combination of multiple weather indicators [44].

Figure 4 shows the Pearson correlation coefficients between the six types of meteorological variables and GHI, as well as among the meteorological variables themselves. It can be observed that the GHI is positively correlated with SEA, AT and WS and negatively correlated with RH, AM and WD across different months. Although the absolute values of the linear correlations between GHI and the WD are relatively low, there may be non-linear relationships between them. Therefore, this does not necessarily mean that WD is more independent and can be ignored. The complex relationships between the meteorological variables and GHI shown in Figure 4 indicate that these six types of meteorological variables are potentially useful features in the prediction process, leading to the development, optimization, and evaluation of the prediction model in the next section [24,44].

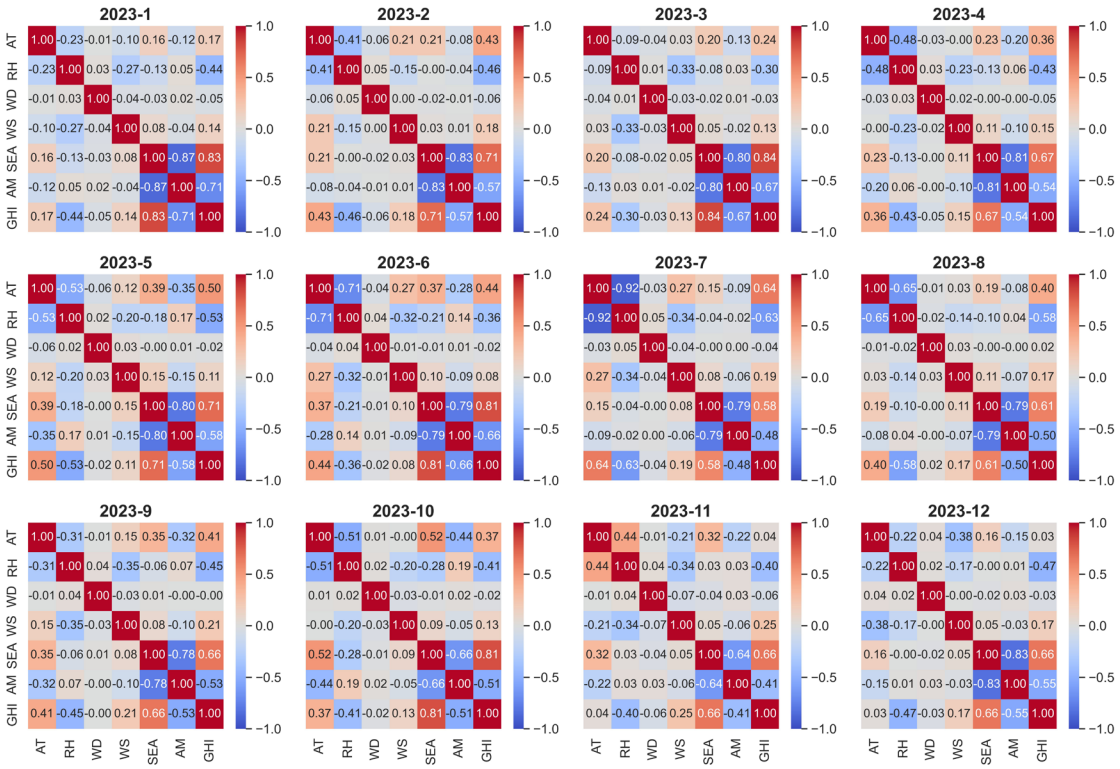

Figure 4. Heatmap of correlations among inputs and between inputs and the output.

3.2. Model development and performance comparison

3.2.1. Model hyperparameter optimization

Although ensemble tree algorithms generally share the same hyperparameters, the default values and importance of these parameters may vary between different algorithms due to the differing principles and assumptions on which each algorithm is based. Consequently, the importance of the same hyperparameters and the strategies for tuning them may differ between algorithms.

Figure 5 visualizes the Bayesian optimization (BO) process of the hyperparameters of the XGBoost model proposed in this study, illustrating the impact of different hyperparameter combinations on the cross-validation R^2 and gradually revealing the interactions and progressive relationships among the hyperparameters. This approach provides a comprehensive understanding of the optimization strategies under different hyperparameter combinations. The white asterisks in each heatmap marks the point with the highest R^2, indicating the best performance of the model under the specific parameter combination. Specifically, the combination of a high subsample ratio (0.881), a larger tree depth (18), a moderate number of estimators (635), and a low learning rate (0.012) yields the best model performance. This result provides important guidance for hyperparameter tuning, helping to find the optimal balance between model complexity and generalization ability, thereby improving the model's prediction performance and robustness.

In addition, the hyperparameter tuning information and cross-validation R^2 values for the XGBoost model and three other comparison models under BO are shown in Table 3. The hyperparameter tuning ranges, optimal values, and performance metrics for each model are listed. The comparison

reveals differences in hyperparameter optimization among the models. Despite the differences in the complexity and selection of hyperparameters, all models identified optimal values within their respective hyperparameter spaces and demonstrated high predictive ability in cross-validation. Notably, the XGBoost exhibited highest cross-validation R^2 after optimization, indicating a clear advantage in handling the current dataset.

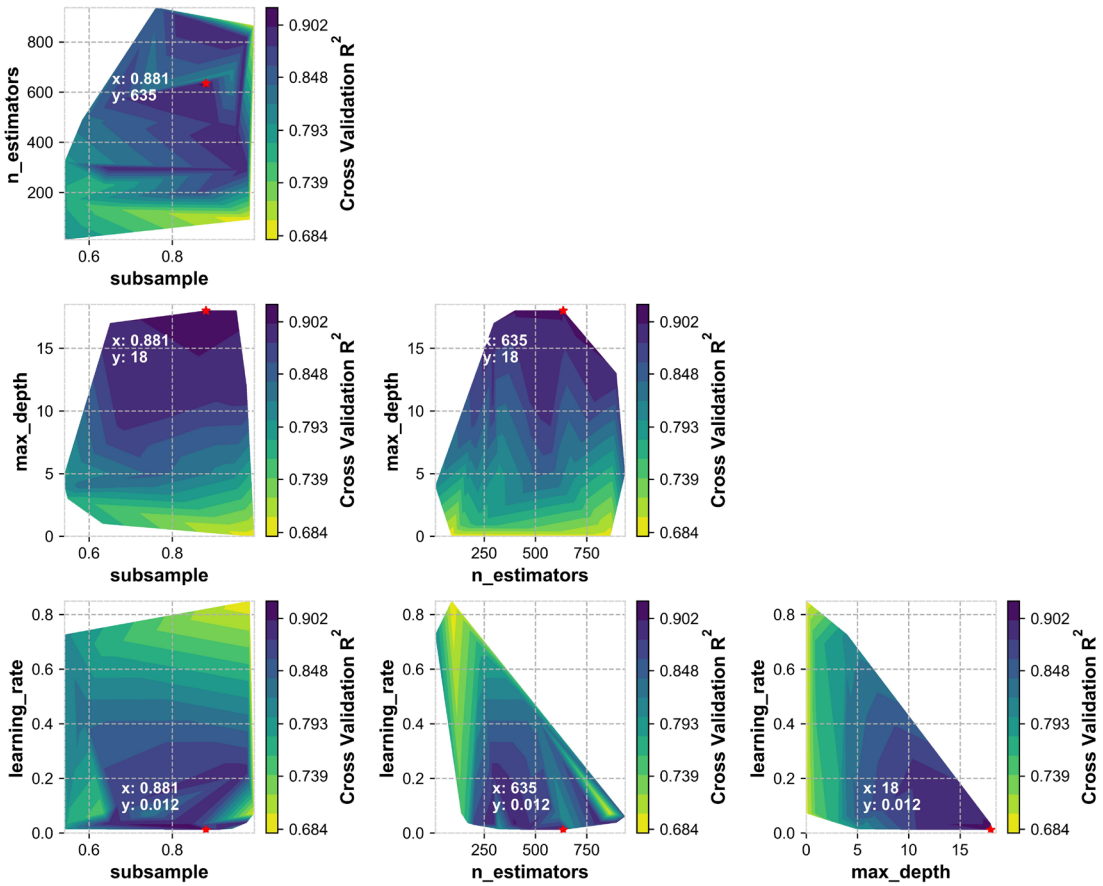

Figure 5. Hyperparameter optimization process of the XGBoost model.

Table 3. Results of hyperparameter optimization of the models.

Model	Hyperparameter	Tuning space	Optimal value	Cross validation R^2
BO-GBDT	n_estimators	[50, 1000]	632	0.891
	learning_rate	[0.00001, 1]	0.026	
	max_depth	[1, 20]	16	
BO-RF	n_estimators	[50, 1000]	718	0.894
	max_depth	[1, 20]	18	
BO-CatBoost	border_count	[1, 255]	90	0.851
	iterations	[50, 1000]	955	
	l2_leaf_reg	[0.00001, 100]	0.029	
	learning_rate	[0.00001, 1]	0.454	
BO-XGBoost	n_estimators	[50, 1000]	635	0.902
	max_depth	[1, 20]	18	
	learning_rate	[0.00001, 1]	0.012	
	subsample	[0.5, 1.0]	0.881	

3.2.2. Model performance

The determination of model performance is primarily based on R^2 and $RMSE$. The prediction performance of the proposed BO-XGBoost model was compared with the simultaneously developed BO-GBDT, BO-RF, and BO-CatBoost models, as shown in Figure 6.

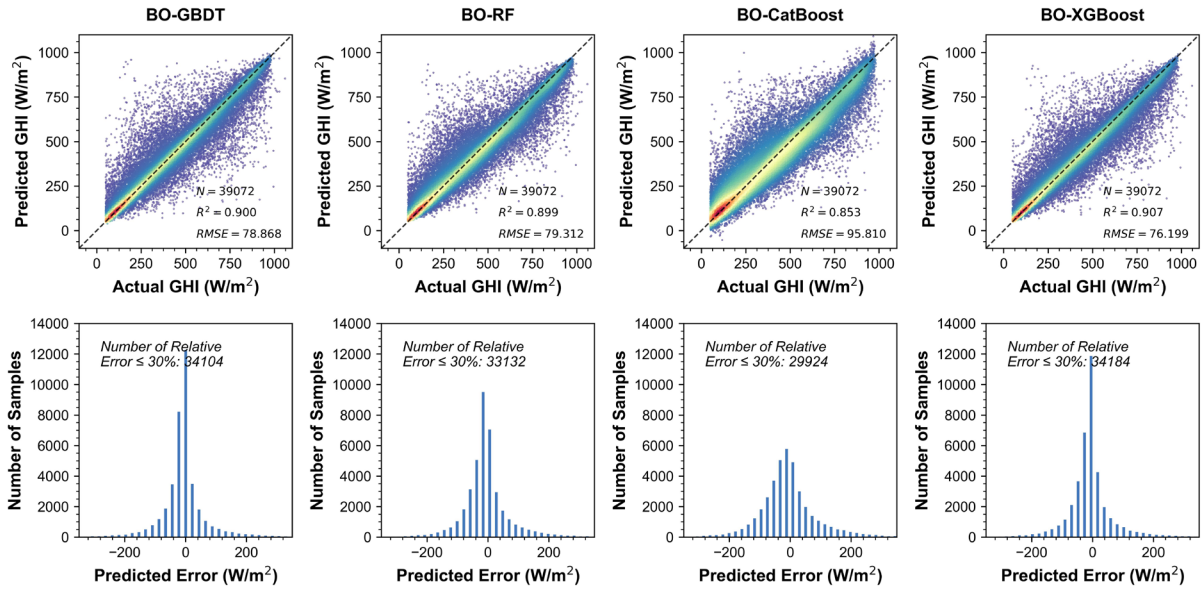

Figure 6. Scatter plot of predicted *versus* actual values and error distribution.

The upper part of the Figure 6 shows the scatter plots of the predicted and actual GHI values for each model and their fitting results, while the lower part displays the corresponding histograms of the prediction error distributions. It can be seen that the BO-XGBoost model can be identified as the best model as the values of R^2 and $RMSE$ in the test set are 0.907 and 76.199 respectively, with smaller errors and the most concentrated distribution. In contrast, the BO-GBDT and BO-RF models the next best, and the BO-CatBoost model has a more dispersed error distribution and larger errors, resulting in relatively poorer prediction accuracy. These results demonstrate that the BO-XGBoost model is particularly effective in capturing the rapid and high-frequency fluctuations of solar irradiance at the minute level, which is a key challenge often overlooked in previous irradiance prediction studies that mainly focus on hourly or longer intervals.

Furthermore, Figure 7 shows the distribution of the actual and predicted GHI values from four models (BO-GBDT, BO-RF, BO-CatBoost, BO-XGBoost) over different time periods to facilitate a deeper analysis of the accuracy of model predictions. It can be seen that all models capture the seasonal and daily patterns of GHI, but there are certain differences in predicting high GHI values. Among them, the predicted values of the BO-XGBoost model are very close to the actual values, especially in higher irradiance domains such as midday and spring-summer seasons.

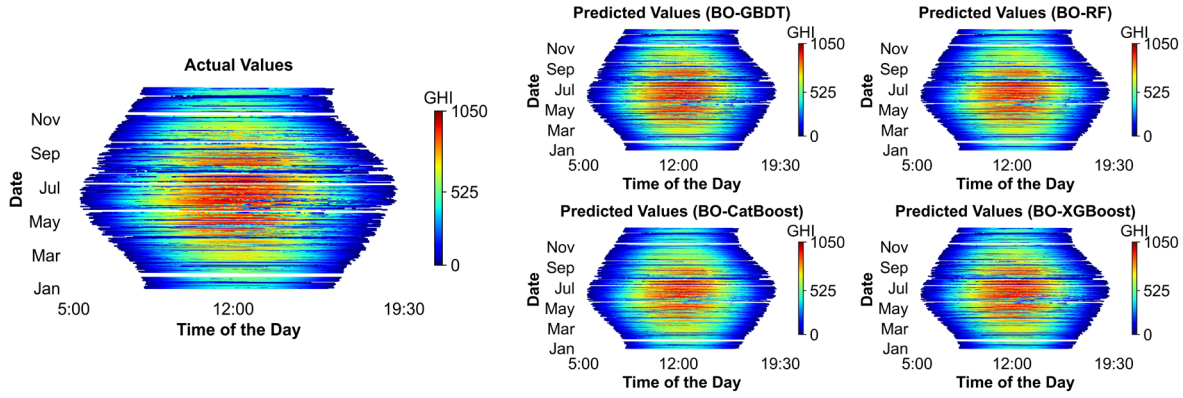

Figure 7. Comparison of the distribution of predicted *versus* actual values over time.

3.3. Interpretability for the proposed model

From the model performance comparison results, it is evident that the BO-XGBoost is the optimal algorithm for predicting GHI. However, the mechanisms by which meteorological parameters influence irradiance are still not fully understood. Therefore, this study analyzed the importance and effects of input features within the BO-XGBoost model. This analysis not only enhances the model's credibility but also provides a foundation for further improvements to the model.

3.3.1. Feature importance

Figure 8 presents a SHAP summary plot for the BO-XGBoost model, illustrating the overall importance of the model's input features.

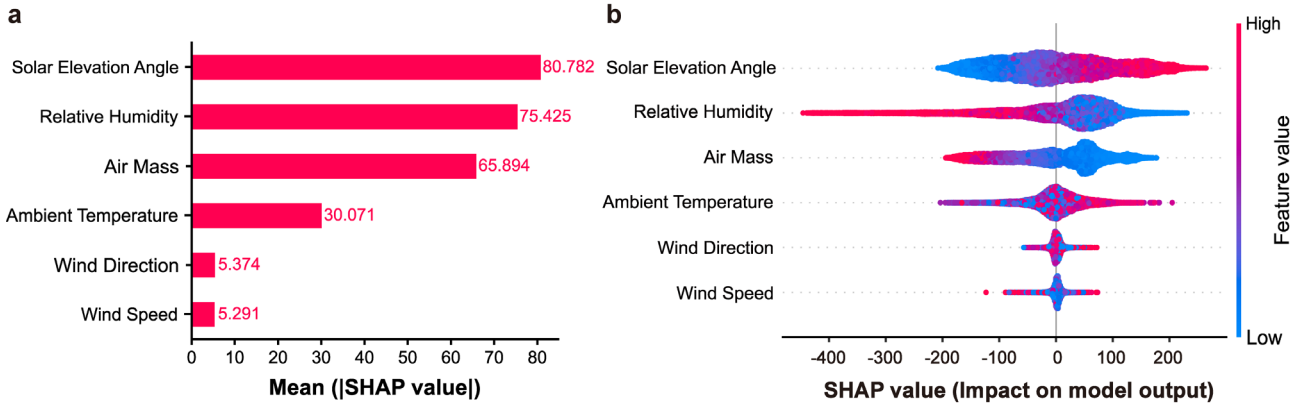

Figure 8. Input feature importance of the BO-XGBoost.

As depicted in the mean bar chart in Figure 8a, regardless of whether the values are positive or negative, higher absolute value indicates greater importance of the corresponding feature to the GHI prediction. The y-axis represents the input features, ranked from top to bottom in terms of importance as SEA, RH, AM, AT, WD and WS with importance percentages of 30.735%, 28.696%, 25.070%, 11.441%, 2.045% and 2.013%, respectively. Among these features, the SHAP values for SEA and RH are the highest, indicating that SEA and RH are the two most important features in the prediction model. This finding is consistent with physical principles: SEA directly determines the solar incidence angle

and atmospheric path length, thereby strongly affecting irradiance intensity, while RH influences scattering and absorption processes, reducing direct irradiance. While AM and AT also contribute notably, the importance of WD and WS is negligible, consistent with previous studies [4,45]. As shown in the beeswarm plot in Figure 8b, the SHAP values are assigned to different features for the predictions, displaying the directionality of the features. By examining the horizontal color distribution of each feature variable, it is possible to identify whether each feature responds positively (positive values on the x-axis) or negatively (negative values on the x-axis) to the prediction values. Unlike other features, higher RH and AM have greater negative contributions to the GHI prediction. The dominance of SEA and RH, together with the considerable contributions of AM and AT, provides new insights into the meteorological drivers of minute-level irradiance variability, an aspect of interpretability at this temporal scale that has rarely been explored in previous studies.

3.3.2. Feature effect

After identifying the key features influencing GHI prediction, it is necessary to further determine their relationships with the prediction outcomes. To achieve this, one-dimensional partial dependence plots (1D-PDP) were generated for the key features (*i.e.*, the two most important ones, as noted above), as shown in Figure 9a,b, along with a two-dimensional partial dependence plot (2D-PDP), as shown in Figure 9c.

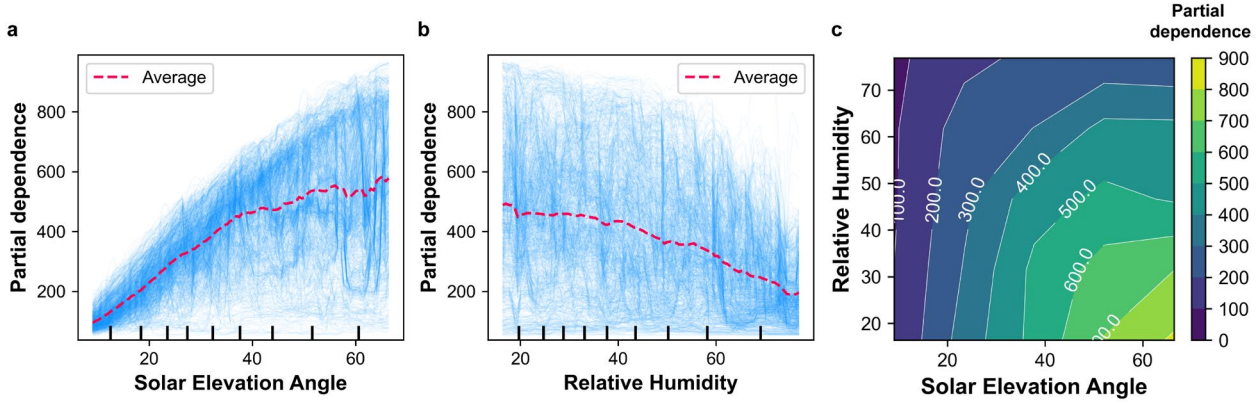

Figure 9. The individual and interactive effects of two key features on GHI prediction.

It is evident that changes in SEA have a significant effect on the response-based GHI prediction, which is positively correlated with the contribution to prediction, as shown in Figure 9a. As illustrated in Figure 9b, RH, identified as the second most important feature, also exerts a considerable influence on GHI prediction, predominantly exhibiting a negative correlation pattern with pronounced fluctuations. Due to the potential for complex interactions between different features, analyzing the PDP for a single feature may not be comprehensive. PDPs with two features can more visualize the effect of feature interactions on the prediction results. As shown in Figure 9c, the two-dimensional partial dependence plots (2D-PDP) illustrate the dependence of the GHI prediction on the interactions between SEA and RH. This means that the same SEA results in greater GHI at lower RH compared to higher RH. When SEA exceeds approximately 20° and RH less than approximately 60%, the interaction between these two features significantly affects the prediction. When RH remains constant, the contribution of SEA to

GHI prediction is almost entirely positive. This is particularly evident at lower RH values, where the gradient of partial dependence changes more significantly. The above PDP analysis results help to understand the specific mechanisms underlying meteorological response-based solar irradiance prediction, facilitating the further development and application of solar spectrum prediction models. These patterns demonstrate threshold-like behavior, whereby RH levels above approximately 60% substantially limit irradiance even under high SEA conditions. Such nonlinear responses highlight the importance of explicitly considering humidity effects in forecasting models. The PDP results further demonstrate pronounced nonlinear and interactive effects of SEA and RH on GHI prediction, underscoring that ultra-short-term irradiance forecasts are highly sensitive to compound meteorological conditions. These results advance the understanding of irradiance dynamics at fine temporal resolutions.

Overall, the integration of SHAP and PDP not only confirms physical intuition but also provides practical insights into microscale meteorological response mechanisms, significantly enhancing the transparency and reliability of the proposed model. In building energy simulations and PV grid integration, these insights can inform computational calibration and operational planning.

3.4. Impact of factors on prediction accuracy

3.4.1. Time intervals

To further assess the sensitivity of the model to prediction time intervals, this subsection analyzes the performance of the proposed BO-XGBoost model and the three comparison models at prediction time intervals of 5, 10, 20, 30 and 60 minutes, respectively, as shown in Figure 10.

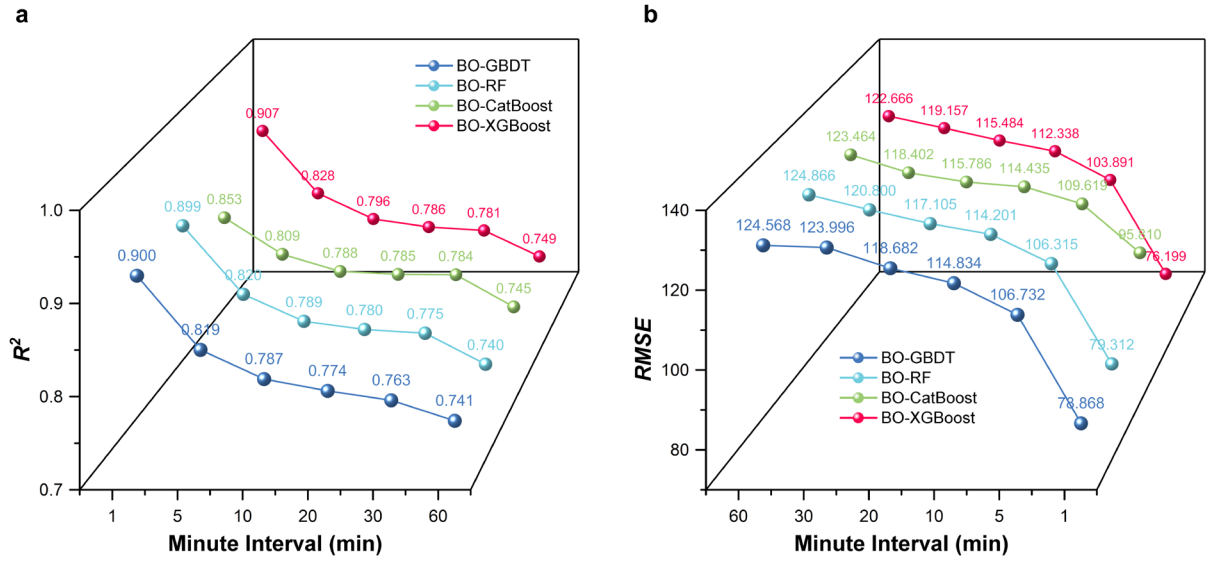

Figure 10. Variation of model performance metrics at different minute intervals.

As shown in Figure 10, the performance of each model is affected differently by the time interval, with a general trend that larger time intervals result in greater prediction errors, indicating strong sensitivity to the prediction time [46]. In this study, the BO-XGBoost model with a 1-minute prediction interval performs the best, which is consistent with previous findings that shorter prediction intervals lead to higher model performance [24,47].

3.4.2. Weather conditions

Solar irradiance is influenced not only by the time of day and season but also by weather conditions. Therefore, it is essential to further evaluate the effectiveness of the developed model in predicting extreme irradiance, particularly under cloudy and rainy conditions. This study categorizes the entire dataset into three weather types: clear sky, partly clear sky, and non-clear sky. The applicability of the developed model is then tested under these three different weather conditions. The clear sky refers to weather condition where sunlight is direct and abundant throughout the entire day. The partly clear sky refers to conditions where the weather alternates between sunny and cloudy or rainy periods, with intervals of sunlight and clouds. The non-clear sky refers to consistently overcast or rainy conditions throughout the entire day, resulting in dim light.

Figure 11 shows the distribution of actual and predicted GHI for nine randomly selected discrete days. There are small deviations between the actual and predicted values for each prediction model, particularly with larger variations in sky conditions leading to larger deviations. However, compared to BO-GBDT, BO-RF and BO-CatBoost, the BO-XGBoost model proposed in this study demonstrates superior capability in tracking the actual irradiance trend.

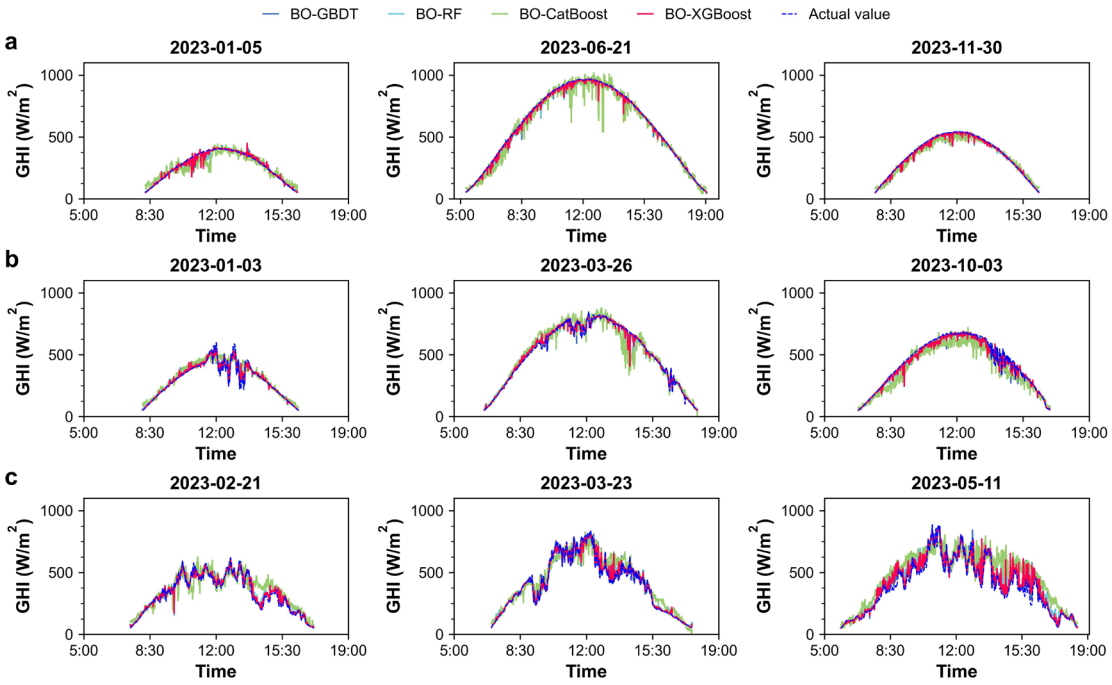

Figure 11. Representative examples under different weather conditions: **(a)** clear sky; **(b)** partly clear sky and **(c)** non-clear sky.

On days 1–3 with clear skies, as shown in Figure 11a, the actual GHI values exhibit a smooth, hump-shaped distribution over time. The predicted values generated from the BO-XGBoost model closely align with the actual values in real-time, effectively capturing the dynamic changes in GHI throughout the day. On days 4–5 with partly clear skies, as shown in Figure 11b, the actual GHI values display a relatively smooth curve during sunny periods and more fluctuations during other complex sky conditions. Notably, compared to clear sky conditions, the performance of the models is almost

unaffected under this weather condition, with the BO-XGBoost model consistently maintaining high prediction accuracy. On days 6–9 with non-clear skies, as shown in Figure 11c, the actual GHI values show irregular changes and significant fluctuations throughout the day due to meteorological factors. However, compared to the other three models, the BO-XGBoost model proposed in this study still demonstrates a stronger competitive advantage.

Additionally, the average R^2 and average $RMSE$ for the four prediction models under the three types of weather conditions are shown in Figure 12. The results highlight that all models are influenced by weather variations, with higher prediction accuracy under clear sky conditions compared to partly clear sky and non-clear sky conditions. This discrepancy underscores the increased difficulty in accurately predicting solar irradiance under complex weather conditions, aligning with the conclusions of existing studies [14,15,17,24,44,48,49]. However, the BO-XGBoost model proposed in this study consistently exhibits superior performance across all prediction scenarios, including clear sky, partly clear sky and non-clear sky. Specifically, under clear sky conditions, the BO-XGBoost achieves an R^2 of 0.990 and an $RMSE$ of 24.077, while under non-clear sky conditions, it still maintains strong prediction accuracy with an R^2 of 0.971 and an $RMSE$ of 42.031. It is worth noting that although the BO-GBDT presents comparable prediction results, requires significantly more computational time compared to BO-XGBoost. Therefore, the BO-XGBoost model proposed in this study can be considered the optimal choice. Therefore, the BO-XGBoost model proposed in this study can be considered the optimal choice.

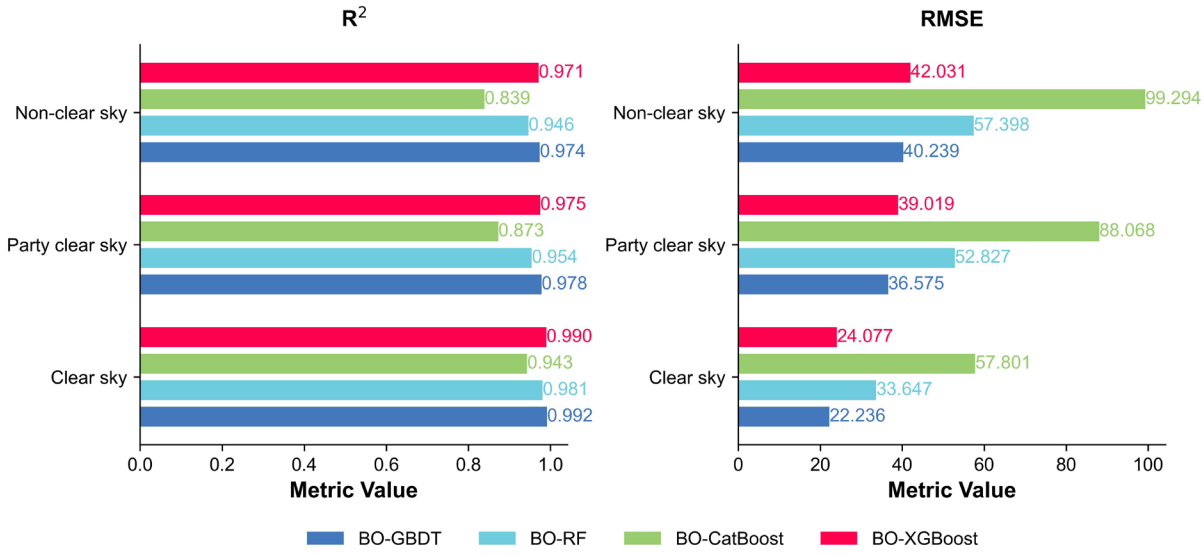

Figure 12. Model performance metrics in different weather conditions.

While the weather type classification in this study may differ somewhat from other published research, the comparative analysis remains illustrative in a broader context. Table 4 summarizes the model prediction results from this study and several other studies, highlighting the progress made by the proposed model in this study in overcoming the challenges posed by complex weather conditions.

Table 4. R^2 and $RMSE/nRMSE$ of different models in different sky conditions.

Reference	Model type	Clear sky			Partly clear sky			Non-clear sky		
		R^2	$RMSE$	$nRMSE$	R^2	$RMSE$	$nRMSE$	R^2	$RMSE$	$nRMSE$
Hou *et al.* [48]	CNN-A-LSTM	0.966	/	0.054	0.956	/	0.087	/	/	/
Huang *et al.* [20]	LSTM-MLP	/	25.229	0.047	/	141.827	0.359	/	26.629	0.463
YU *et al.* [49]	LSTM	0.990	27.610	/	0.920	70.810	/	0.933	43.280	/
Kumari *et al.* [50]	LSTM-CNN	0.979	47.524	/	0.932	79.851	/	0.967	53.257	/
Papachristopoulou *et al.* [51]	NWC-SAF	0.960	72.900	0.115	0.840	135.600	0.379	0.730	138.000	1.008
Bae *et al.* [35]	SVM	0.969	49.260	/	0.945	62.570	/	0.9182	57.870	/
This study	GBDT	0.992	22.236	/	0.978	36.575	/	0.974	40.239	/
	RF	0.981	33.647	/	0.954	52.827	/	0.946	57.398	/
	CatBoost	0.943	57.801	/	0.873	88.068	/	0.839	99.294	/
	XGBoost	0.990	24.077	/	0.975	39.019	/	0.971	42.031	/

3.5. Limitations and future work

The proposed BO-XGBoost framework achieves a favorable balance between prediction accuracy and computational efficiency, significantly reducing the cost of minute-level irradiance modeling and demonstrating strong potential for engineering applications. Nevertheless, its generalizability remains limited. The model was developed and validated using data from a single site in Beijing, which represents a typical temperate continental climate. Although the dataset spans a full year and reflects pronounced seasonal variability, its applicability to other climate zones still requires further validation. For regions with climatic conditions similar to Beijing, the framework has strong potential for transferability and application.

Future research should extend the analysis to multi-site and multi-climate datasets to assess scalability and identify transferable patterns. In addition, incorporating numerical weather forecast data as model inputs represents an effective means of enabling real-time predictions at larger spatial scales, thereby supporting practical applications in power system operation, PV grid integration, and building energy management.

Overall, by relying solely on conventional meteorological parameters, the proposed framework offers good practicality, cost-effectiveness, and scalability, making it well suited for deployment in real engineering contexts.

4. Conclusions

To obtain accurate ultra-short-term irradiance data with extremely limited measurement equipment resources, this study proposes a prediction framework using the XGBoost model combined with BO algorithm, based on six types of conventional meteorological parameters. The prediction mechanism is further interpreted using SHAP and PDP techniques to enhance the model's credibility. The following conclusions can be drawn:

(i) The BO-XGBoost model exhibits significant advantages in irradiance prediction, achieving the highest accuracy compared to the other three state-of-the-art models with R^2 of 0.907 and $RMSE$ of 76.199. Particularly under clear sky conditions, the model's R^2 and $RMSE$ further improve to 0.990 and 24.077, respectively.

(ii) Among all input features, the solar elevation angle (SEA) and relative humidity (RH) are identified as the top two features influencing model predictions, contributing 30.735% and 28.696% respectively. GHI predictions are positively correlated with SEA and negatively correlated with RH. Additionally, the interactions between SEA and RH further impact the GHI predictions.

(iii) Longer prediction time intervals and more complex weather conditions reduce the model's prediction accuracy. However, even under complex weather conditions, the BO-XGBoost model maintains a high level of prediction accuracy, demonstrating a strong competitive advantage.

This study advances the field of solar energy applications by introducing a cost-effective, efficient, and interpretable irradiance prediction framework. The generalizability of the model, however, remains limited by the current dataset, which is restricted to the Beijing area. Future work should incorporate multi-site and long-term datasets to further evaluate and extend the applicability of the proposed approach. Despite these limitations, the findings provide methodological and practical insights that can inform subsequent research and support the development of solar-driven building performance simulations and low-carbon energy system design.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 52378085, 12411530112), Beijing Capital Development Co., Ltd. (No. FZX230305), Chongqing Natural Science Foundation (No. CSTB2022NSCQ-MSX0937) and National Natural Science Foundation of China (No. 72301019).

Authors' contribution

Yanyun Zhang: conceptualization, methodology, investigation, software, visualization, writing—original draft, writing—review & editing. Runze Shi: investigation, data curation. Yupeng Wu: supervision, conceptualization, writing-review & editing. Peng Xue: conceptualization, funding acquisition, project administration, resources, supervision, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of interests

The authors declare no conflict of interest.

References

[1] Ürge-Vorsatz D, Chatterjee S, Cabeza LF, Molnár G. Global and regional estimation and evaluation of suitable roof area for solar and green roof applications. *Dev. Built. Environ.* 2025, 21:100607.

[2] Abouelaziz I, Jouane Y. Photogrammetry and deep learning for energy production prediction and building-integrated photovoltaics decarbonization. *Build. Simul.* 2024, 17:189–205.

[3] Yaman K, Arslan G. The impact of hourly solar radiation model on building energy analysis in different climatic regions of Turkey. *Build. Simul.* 2018, 11:483–495.
[4] Chen C, Duan Q, Feng Y, Wang J, Ghaeili Ardabili N, *et al.* Reconstruction of narrowband solar radiation for enhanced spectral selectivity in building-integrated solar energy simulations. *Renew. Energy* 2023, 219:119554.
[5] Vuckovic M, Hammerberg K, Mahdavi A. Urban weather modeling applications: a Vienna case study. *Build. Simul.* 2020, 13:99–111.
[6] Dong H, Xu C, Chen W. Modeling and configuration optimization of the rooftop photovoltaic with electric-hydrogen-thermal hybrid storage system for zero-energy buildings: consider a cumulative seasonal effect. *Build. Simul.* 2023, 16:1799–1819.
[7] Fan J, Wang X, Wu L, Zhou H, Zhang F, *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, 164:102–111.
[8] Hassan MA, Khalil A, Kaseb S, Kassem MA. Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Appl. Energy* 2017, 203:897–916.
[9] Li H, He X, Hu Y, Lv W, Yang L. Research on the generation method of missing hourly solar radiation data based on multiple neural network algorithm. *Energy* 2024, 287:129650.
[10] Song Z, Cao S, Yang H. Assessment of solar radiation resource and photovoltaic power potential across China based on optimized interpretable machine learning model and GIS-based approaches. *Appl. Energy* 2023, 339:121005.
[11] Hassan MA, Khalil A, Kaseb S, Kassem MA. Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renew. Energy* 2017, 111:52–62.
[12] Urraca R, Antonanzas J, Alia-Martinez M, Martinez-De-Pison FJ, Antonanzas-Torres F. Smart baseline models for solar irradiation forecasting. *Energy Convers. Manag.* 2016, 108:539–548.
[13] Maciel JN, Ledesma JJG, Junior OHA, Junior OHA. Hybrid prediction method of solar irradiance applied to short-term photovoltaic energy generation. *Renew. Sustain. Energy Rev.* 2024, 192:114185.
[14] Wang Z, Wang L, Huang C, Luo X. A hybrid ensemble learning model for short-term solar irradiance forecasting using historical observations and sky images. *IEEE Trans. Ind. Appl.* 2023, 59:2041–2049.
[15] Liu J, Zang H, Cheng L, Ding T, Wei Z, *et al*. A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting. *Appl. Energy* 2023, 342:121160.
[16] Nou J, Chauvin R, Eynard J, Thil S, Grieu S. Towards the intrahour forecasting of direct normal irradiance using sky-imaging data. *Heliyon* 2018, 4:e00598.
[17] Puah BK, Chong LW, Wong YW, Begam KM, Khan N, *et al.* A regression unsupervised incremental learning algorithm for solar irradiance prediction. *Renew. Energy* 2021, 164:908–925.
[18] Sharma A, Kakkar A. Forecasting daily global solar irradiance generation using machine learning. *Renew. Sustain. Energy Rev.* 2018, 82:2254–2269.
[19] Gao XY, Liu JM, Yuan Y, Tan HP. Global horizontal irradiance prediction model considering the effect of aerosol optical depth based on the Informer model. *Renew. Energy* 2024, 220:119671.

[20] Huang X, Zhang C, Li Q, Tai Y, Gao B, *et al.* A comparison of hour-ahead solar irradiance forecasting models based on LSTM network. *Math. Probl. Eng.* 2020, 2020:4251517.
[21] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* 2018, 148:461–468.
[22] Ahmed U, Khan AR, Mahmood A, Rafiq I, Ghannam R, *et al.* Short-term global horizontal irradiance forecasting using weather classified categorical boosting. *Appl. Soft Comput.* 2024, 155:111441.
[23] del Campo-Avila J, Pilougine M, Morales-Bueno R, Mora-López L. A data mining system for predicting solar global spectral irradiance. Performance assessment in the spectral response ranges of thin-film photovoltaic modules. *Renew. Energy* 2019, 133:828–839.
[24] Pereira S, Canhoto P, Salgado R. Development and assessment of artificial neural network models for direct normal solar irradiance forecasting using operational numerical weather prediction data. *Energy AI* 2024, 15:100314.
[25] Fan J, Wu L, Zhang F, Cai H, Zeng W, *et al.* Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. *Renew. Sustain. Energy Rev.* 2019, 100:186–212.
[26] Lee J, Wang W, Harrou F, Sun Y. Reliable solar irradiance prediction using ensemble learning-based models: a comparative study. *Energy Convers. Manag.* 2020, 208:112582.
[27] Azizi N, Yaghoubirad M, Farajollahi M, Ahmadi A. Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output. *Renew. Energy* 2023, 206:135–147.
[28] Verzijlbergh RA, Heijnen PW, de Roode SR, Los A, Jonker HJJ. Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications. *Solar Energy* 2015, 118:634–645.
[29] Segarra-Tamarit J, Pérez E, Moya E, Ayuso P, Beltran H. Deep learning-based forecasting of aggregated CSP production. *Math. Comput. Simul.* 2021, 184:306–318.
[30] Dou W, Wang K, Shan S, Li C, Wang Y, *et al.* Day-ahead numerical weather prediction solar irradiance correction using a clustering method based on weather conditions. *Appl. Energy* 2024, 365:123239.
[31] Qin S, Liu Z, Qiu R, Luo Y, Wu J, *et al.* Short–term global solar radiation forecasting based on an improved method for sunshine duration prediction and public weather forecasts. *Appl. Energy* 2023, 343:121205.
[32] Zhang L, Wilson R, Sumner M, Wu Y. Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: a gate and transformer mechanism approach. *Renew. Energy* 2023, 216:118952.
[33] Zuo HM, Qiu J, Li FF. Ultra-short-term forecasting of global horizontal irradiance (GHI) integrating all-sky images and historical sequences. *J. Renew. Sustain. Energy* 2023, 15(5):053701.
[34] Bhatt A, Ongsakul W, Singh JG. Sliding window approach with first-order differencing for very short-term solar irradiance forecasting using deep learning models. *Sustain. Energy Technol. Assess.* 2022, 50:101864.
[35] Bae KY, Jang HS, Sung DK. Hourly solar irradiance prediction based on support vector machine and its error analysis. *IEEE Trans. Power Syst.* 2017, 32:935–945.

[36] Allal Z, Noura HN, Chahine K. Machine learning algorithms for solar irradiance prediction: a recent comparative study. *e-Prime Adv. Electr. Eng. Electron. Energy* 2024, 7:100453.
[37] Singh N, Jena S, Panigrahi CK. A novel application of Decision Tree classifier in solar irradiance prediction. *Mater. Today Proc.* 2022, 58:316–323.
[38] Aggarwal SK, Saini LM. Solar energy prediction using linear and non-linear regularization models: a study on AMS (American Meteorological Society) 2013–2014 Solar Energy Prediction Contest. *Energy* 2014, 78:247–256.
[39] Schonlau M, Zou RY. The random forest algorithm for statistical learning. *Stata J.* 2020, 20:3–29.
[40] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, August 13–17, 2016, pp. 785–794.
[41] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In *25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, December 12–17, 2011, pp. 2546–2554.
[42] Tursunalieva A, Alexander DLJ, Dunne R, Li J, Riera L, *et al*. Making sense of machine learning: a review of interpretation techniques and their applications. *Appl. Sci. (Switzerland)* 2024, 14(2):496.
[43] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 4765–4774.
[44] Sharma N, Sharma P, Irwin D, Shenoy P. Predicting solar generation from weather forecasts using machine learning. In *IEEE International Conference on Smart Grid Communications,* Brisbane, Australia, October 17–20, 2011, pp. 528–533.
[45] 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.
[46] Engerer NA. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. *Solar Energy* 2015, 116:215–237.
[47] Cheng HY, Yu CC, Lin SJ. Bi-model short-term solar irradiance prediction using support vector regressors. *Energy* 2014, 70:121–127.
[48] Hou X, Ju C, Wang B. Prediction of solar irradiance using convolutional neural network and attention mechanism-based long short-term memory network based on similar day analysis and an attention mechanism. *Heliyon* 2023, 9:e21484.
[49] Yu Y, Cao J, Zhu J. An LSTM short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access* 2019, 7:145651–145666.
[50] Kumari P, Toshniwal D. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Appl. Energy* 2021, 295:117061.
[51] Papachristopoulou K, Fountoulakis I, Bais AF, Psiloglou BE, Papadimitriou N, *et al.* Effects of clouds and aerosols on downwelling surface solar irradiance nowcasting and short-term forecasting. *Atmos. Meas. Tech.* 2024, 17:1851–1877.