BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

ARTICLE

Evaluating the Generalization of an Ensemble Learning Model for Global Horizontal Irradiance Estimation in Guangxi Province Using FY-4A Satellite Data

Jiaqiu Hu [1], Yiming Qin [1], Qian Ye [2*], Kui Huang [1], Houjian Zhan [1], Jian Tang [1], Jie Lin [1], Yixin Zhuo [1], Huanxing Qi [1]

[1] *Power Dispatching Control Center of Guangxi Power Grid, Nanning 530012, China*

[2] *National Satellite Meteorological Center, China Meteorological Administration, Beijing 100081, China*

ABSTRACT

This study investigates the application of the Extreme Gradient Boosting (XGBoost) ensemble learning algorithm for estimating global horizontal irradiance (GHI) based on satellite data in Guangxi Province, China. By synergistically integrating top-of-atmosphere (TOA) reflectance and brightness temperature data from 14 spectral bands of the Fengyun-4A (FY-4A) Advanced Geosynchronous Radiation Imager (AGRI) and European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) reanalysis meteorological variables—including relative humidity, planetary boundary layer height, and surface pressure—we developed an all-sky model to predict hourly surface solar radiation. The model was trained on data from 159 ground stations across China in 2018 and incorporates 31 features covering satellite observations, geographical parameters, and meteorological variables. Validation was conducted using independent observational data from three additional ground stations in Guangxi (Guilin, Nanning, and Beihai) that were withheld from training, yielding Root Mean Square Error (RMSE) of approximately 126–150 W/m^2 and Correlation Coefficient (CC) of 0.80–0.84, confirming strong spatial generalization. Seasonal analysis revealed that the model performed best in winter and least accurately in summer, attributable to the complexity of convective cloud dynamics in the subtropical monsoon climate of Guangxi. Feature importance analysis identified brightness temperature at 7.42 μm, solar zenith angle, relative humidity, and TOA reflectance

*CORRESPONDING AUTHOR:

Qian Ye, National Satellite Meteorological Center, China Meteorological Administration, Beijing 100081, China; Email: yeqian@cma.gov.cn

ARTICLE INFO

Received: 10 January 2026 | Revised: 8 March 2026 | Accepted: 11 March 2026 | Published Online: 15 April 2026
DOI: https://doi.org/10.30564/jees.v8i4.13005

CITATION

Hu, J., Qin, Y., Ye, Q., et al., 2026. Evaluating the Generalization of an Ensemble Learning Model for Global Horizontal Irradiance Estimation in Guangxi Province Using FY-4A Satellite Data. Journal of Environmental & Earth Sciences. 8(4): 143–156.
DOI: https://doi.org/10.30564/jees.v8i4.13005

at 0.47 μm as the most influential predictors, consistent with the physical mechanisms governing atmospheric transmissivity. These findings demonstrate that the direct data-driven satellite–machine learning framework offers a computationally efficient and scalable alternative to semi-empirical approaches for regional solar resource assessment.

Keywords: Global Horizontal Irradiance; FY-4A; XGBoost; Data Validation

1. Introduction

Global Horizontal Irradiance (GHI) refers to the total amount of solar radiation received per unit time on a specific area of the Earth's surface, oriented perpendicular to the sun's rays. The measurement of GHI includes both direct and diffuse radiation and serves as a crucial indicator for assessing solar energy resources, climate change, and their impacts on ecosystems. Variations in GHI are influenced by multiple factors, including atmospheric composition, cloud cover, topography, and seasonal changes, all of which collectively determine the intensity and distribution characteristics of radiation in different regions [1–4].

Globally, GHI data is essential for the development of renewable energy, particularly in the field of solar power generation. Accurate GHI observations can help decision-makers optimize the siting and design of solar energy facilities, improving energy utilization efficiency. Furthermore, with the intensification of climate change, long-term monitoring of GHI also provides foundational data for studying climate patterns and assessing environmental changes [5]. Therefore, establishing effective observation networks and data analysis methods to enhance the accuracy of GHI measurements has become an important task in current scientific research [6].

Relying solely on sparse ground observations for regional solar radiation distribution presents significant limitations. Ground-based measurements are often limited in spatial coverage, which can lead to an incomplete understanding of solar irradiance patterns across diverse geographical areas. For instance, regions with few monitoring stations may not accurately represent local variations in solar radiation due to topographical differences, atmospheric conditions, and seasonal changes. This lack of comprehensive data can result in misleading conclusions about solar energy potential and hinder effective planning for solar energy projects. Furthermore, ground observations are susceptible to measurement errors and environmental factors that can affect their reliability, such as shading from nearby structures or vegetation. To address these challenges, integrating satellite-based data and advanced modeling techniques is essential for producing more accurate and high-resolution assessments of solar radiation distribution across larger regions [7,8]. Satellite-derived products have been increasingly adopted to overcome the spatial limitations of ground networks. Studies utilizing Meteosat, Geostationary Operational Environmental Satellite (GOES), and Himawari satellites have demonstrated that geostationary platforms can provide consistent, high-temporal-resolution GHI estimates over large domains [9,10].

Recent studies have increasingly focused on estimating global horizontal irradiance (GHI) using machine learning algorithms, highlighting their potential to enhance solar energy forecasting. One notable research effort constructed twelve distinct machine learning models to predict daily and monthly solar radiation values, employing techniques such as gradient boosting regression trees and random forests. The study emphasized the significance of meteorological factors like sunshine duration and land surface temperature in improving prediction accuracy [11]. Another investigation utilized deep learning methods, including convolutional neural networks (CNN) and long short-term memory (LSTM) networks, to predict short-term solar irradiance, demonstrating that hybrid models can effectively leverage the strengths of various algorithms for enhanced accuracy [12–14]. Moreover, a comparative analysis of different machine learning approaches revealed that models like the gradient boosting regressor and random forest regressor excelled in capturing non-linear patterns in solar radiation data [11]. These advancements underscore the growing importance of machine learning in optimizing solar energy management and forecasting. Beyond ground-based approaches, satellite-derived GHI estimation using geostationary platforms has gained increasing traction. Tan et al. [15] demonstrated that multi-spectral TOA reflectance from the Himawari-8 satellite, when used as the sole input to machine learning models including XGBoost and random forest, can achieve near-real-time GHI es-

timation that outperforms the official Himawari-8 shortwave radiation product, highlighting the potential of direct data-driven satellite approaches. More recently, Suwanwimolkul et al.[16] developed a near-real-time GHI mapping system over Thailand using Himawari-8 imagery and advanced deep learning architectures, finding that LightGBM achieved the best overall performance. These parallel developments on comparable geostationary platforms underscore the broader applicability of the satellite–machine learning framework adopted in the present study.

However, despite these promising developments, several shortcomings remain. Many of these studies primarily rely on ground-based meteorological measurements, which may not capture the full spatial variability of cloud cover and atmospheric conditions. Recent literature suggests that integrating satellite-derived data—which provides comprehensive spatial coverage—with advanced AI techniques can significantly enhance GHI forecasting performance[17]. Such hybrid approaches are necessary to overcome the limitations of existing models by fusing complementary information from both ground and remote sensing sources, ultimately leading to more robust and accurate solar energy predictions.

In this study, we developed an all-sky model utilizing the XGBoost machine learning algorithm to estimate hourly solar radiation distribution across China. To accomplish this, we incorporated top-of-atmosphere (TOA) reflectance data from the FY-4A satellite, along with ERA5 reanalysis meteorological data and geographical information during 2018. To validate the model's performance and generalization capabilities, we assessed the estimation accuracy using data from Guangxi Province, which was not included in the training dataset. Detailed information regarding the data and methodologies employed is presented in Section 2. Section 3 outlines the validation results, while Section 4 provides a comprehensive analysis and discussion, with key findings summarized in Section 5.

2. Data and Method

2.1. FY4A Data

China's second-generation geostationary meteorological satellite, Fengyun-4A (FY-4A), was successfully launched on December 11, 2016, and is positioned at 104.7° E. It is equipped with four payloads: the Advanced Geosynchronous Radiation Imager (AGRI)[18–20], the Geostationary Interferometric Infrared Sounder (GIIRS), the Lightning Mapping Imager (LMI), and a space environment package (SEP). The AGRI features 14 spectral bands ranging from 0.47 μm in the visible spectrum to 13.8 μm in the infrared, with varying spatial resolutions of 1 km at nadir in the visible, 2 km in the near-infrared, and 4 km in the infrared[21]. The increased number of spectral bands compared to the five bands of the previous FY-2 series enables more effective aerosol retrieval from geostationary satellites. The high temporal resolution of FY-4A is particularly advantageous for air quality monitoring and modeling, enhancing the detection and tracking of haze.

Top-of-atmosphere (TOA) reflectance at a given wavelength arises from both surface-reflected radiation and atmospheric scattering—without any direct surface interaction. This angular spectral reflectance is closely linked to solar zenith angles (SZA) and aerosol optical properties. In remote sensing studies, the SZA is crucial because larger angles lead to longer atmospheric paths, which enhance scattering and absorption[22]. These effects can diminish the radiation reaching the Earth's surface and alter the reflected signal detected by sensors. To ensure high data quality and reliable surface reflectance measurements, samples with SZAs less than 72° are typically selected. Building on this framework, we analyzed 14 bands of reflectance along with corresponding SZA data, supplemented by land cover and surface elevation information. Moreover, since TOA reflectance data are only available during daylight, our study focused exclusively on daytime samples with solar zenith angles below 72°.

While we utilize the same input variables for both cloudy and clear sky conditions, cloud masking is essential during model construction. In this study, we employ the Level-2 cloud mask product as a cloud indicator. More details of FY-4A could be found at the data server website (http://satellite.nsmc.org.cn).

2.2. Ground-Based Solar Radiation Measurements

The China Meteorological Administration (CMA) provides ground radiation measurements of 162 stations in China in 2018. Thermoelectric pyranometers, characterized by a spectral response spanning the range of 0.3 to 3.0 microme-

ters, have been deployed across the CMA network. These pyranometers have an uncertainty of ±3%[8,23]. The CMA surface observations include three basic physical variables: the global horizontal irradiance (GHI), direct solar radiation, and diffuse radiation. This study uses the CMA GHI product with a temporal resolution of 1 hour. The distribution of CMA radiation stations is shown in **Figure 1**. 3 stations in Guangxi are left for validation, and the remaining 159 stations are all used in model training. As solar irradiance is zero at night, only daytime data during 00:00–09:00 Coordinated Universal Time (UTC) were used. Finally, we obtained 443,933 valid samples across 2018.

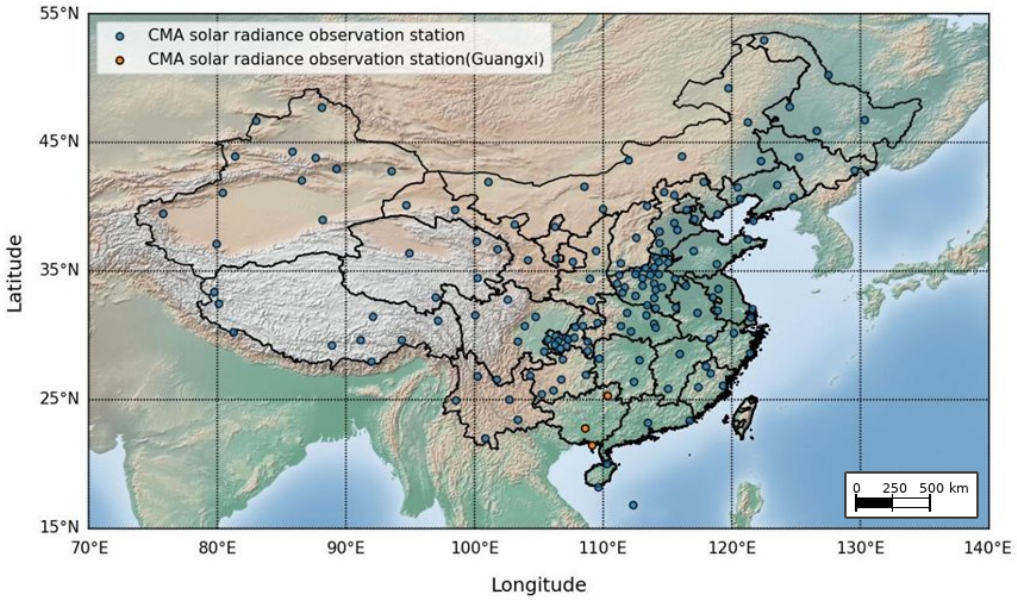

Figure 1. Geographical distribution of the CMA stations.

2.3. Meteorological Variables

ERA5 is the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for global climate and weather research, which assimilates as many observations as possible in the upper air and near surface. It provides a large number of hourly atmospheric variables at a resolution of 0.25 degrees.

Many previous studies have revealed that meteorological factors have a significant influence on ground GHI[24–26]. The vertical mixing of aerosols—which directly affects the attenuation of solar radiation—is largely determined by the planetary boundary layer height (PBLH). A higher PBLH indicates stronger turbulent mixing and a deeper layer for pollutant dispersion, which generally reduces the columnar aerosol load, whereas a lower PBLH can lead to elevated aerosol concentrations that diminish the solar irradiance reaching the surface. Similarly, relative humidity (RH) is critical because it governs aerosol hygroscopic growth and alters their optical properties. As RH increases, aerosols absorb more water, which enhances their scattering and absorption capabilities, thereby intensifying the extinction of incoming radiation. Consequently, incorporating PBLH and RH as predictor variables helps capture both the vertical distribution of aerosols and their moisture-dependent optical effects—factors that are crucial for accurately estimating global horizontal irradiance (GHI). For these reasons, the surface atmospheric pressure (P, hPa), total column water (TCW, kg m^{-2}), 10-m u-wind (U10) and v-wind (V10) component, air temperature at an altitude of 2 m (T, K), total column ozone (kg m^{-2}), relative humidity (RH, %), and planetary boundary layer height (PBLH, m) were chosen for GHI estimation. The P, T, PBLH, RH, U10, V10, TCW, and ozone from ERA5 were used and resampled to match the FY-4A data.

2.4. XGBoost Algorithm

XGBoost is a novel algorithm introduced in 2016 by Chen and Guestrin[27]. Similar to the Random Forest (RF) algorithm, XGBoost is an ensemble method based on many weak learners. XGBoost employs a gradient boosting ensem-

ble technique, which fundamentally differs from the bagging approach adopted by random forests (RF) in terms of both model training and error correction mechanisms. Learners of RF are parallel, and share the same data distribution. However, learners of XGBoost are serial, and focus more on samples that are predicted incorrectly. XGBoost is very efficient in computation; the training time of the XGBoost model is 1/7 of that of the RF model with the same hyperparameter settings. The key hyperparameters of XGBoost include the number of trees and the maximum depth of the tree.

Feature importance is obtained from the impurity reduction. For a given node m with left and right child nodes, the impurity reduction $Gain_m$ is expressed as

$$Gain_m = i_m - (w_{left} \cdot i_{left} + w_{right} \cdot i_{right}) \quad (1)$$

where i_m, i_{left} and i_{right} are the impurity of node m, its left and right child nodes, respectively. w is the weight, and is defined as the share of the parent's examples in a child node (e.g., $w_{left} = N_{left}/N_m$, where N is the number of examples in a node or leaf). In order to derive the total impurity reduction of a given feature f in tree t, we need to sum across all nodes m$\in M_f^{(t)}$, which perform a split on that feature f and divide it by the total impurity reduction number of all nodes of that tree. Eventually, the total importance of a feature f is calculated across all trees t in the random forest with a total number of trees T, and expressed as

$$Importance_f = \frac{1}{T}\sum_{t=1}^{T} Importance_f^{(t)} \quad (2)$$

where $Importance_f^{(t)}$ is the importance of a given feature f in tree t, and expressed as

$$Importance_f^{(t)} = \frac{\sum_{m\in M_f^{(t)}} Gain_m}{\sum_f \sum_{m\in M_f^{(t)}} Gain_m} \quad (3)$$

As shown in **Table 1**, the XGBoost model used here was developed by incorporating a total of 31 features, including FY-4A observations at 13 wavelengths, NDVI index, 2 observation angles, elevation, land cover information, and 14 meteorological variables from the ERA-5 reanalysis.

Table 1. Description of the data sources used in this study.

Type	Parameters	Temporal Resolution	Spatial Resolution	Sources
Ground-level solar irradiance observation	GHI	hourly	-	
Satellites Datasets	TOA reflectance, Brightness Temperature, and angle information from FY-4A	hourly	4 km	http://satellite.nsmc.org.cn/portalsite/default.aspx
	NDVI	hourly	4 km	FY-4A
Meteorological data from numerical weather model	surface pressure (SP) temperature at 2 m (T2M) relative humidity (RH) total column water (TCW) total column ozone (TCO3) planet boundary layer height (PBLH) 10-m u-wind (U10) 10-m v-wind (V10)	hourly	0.25°	https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview

2.5. Validation Method

Three widely used statistical metrics were selected to evaluate the agreement between satellite-derived solar irradiance estimates and ground-based observations, providing a comprehensive assessment of model accuracy.

Correlation Coefficient (CC): Quantifies the linear relationship between satellite estimates and observations. Values close to 1 indicate a strong positive correlation [28].

$$CC = \frac{\left[\sum_{i=1}^{n}\left(S_i-\overline{S}\right)\cdot\left(G_i-\overline{G}\right)\right]^2}{\sum_{i=1}^{n}\left(S_i-\overline{S}\right)^2\cdot\sum_{i=1}^{n}\left(S_i-\overline{S}\right)\cdot\left(G_i-\overline{G}\right)^2} \quad (4)$$

Mean Absolute Error (MAE): Measures the average absolute difference between satellite estimates and observations, regardless of direction. Lower values indicate better accuracy [29].

$$\mathrm{MAE} = 1\sum_{i=1}^{n}|S_i - G_i| \quad (5)$$

Root Mean Square Error (RMSE): Penalizes larger deviations more heavily than ME or MAE, offering a stringent assessment of errors [30].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - G_i)^2} \quad (6)$$

Mean Absolute Percentage Error (MAPE): Expresses the average absolute error as a percentage of the observed value, providing a scale-independent measure of relative accuracy. Unlike MAE and RMSE, MAPE is not sensitive to the magnitude of the target variable, making it particularly appropriate for comparing model performance across conditions with substantially different GHI levels, such as clear-sky versus cloudy-sky conditions.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{S_i - G_i}{Gi}\right| \times 100\% \qquad (7)$$

where S_i denotes the satellite-estimated GHI for sample i; G_i denotes the corresponding ground-observed GHI; $\overline{S}$ and $\overline{G}$ are the means of all satellite estimates and ground observations, respectively; and n is the total number of samples.

Together, these metrics provide complementary perspectives on model performance: MAE and RMSE characterize overall average accuracy, while CC reflects the degree of linear agreement between estimates and observations. Low RMSE and MAE, combined with CC close to 1, indicate both accurate and consistent predictions across diverse geographic and weather conditions.

3. Results and Validation

3.1. Model General Performance

Initially, we fixed the key hyperparameters using 10-fold cross-validation. With these hyperparameters established, we proceeded to train the XGBoost model. Prior to training, the dataset was divided into training and testing sets, with an 80% (355,146 samples) to 20% (88,787 samples) ratio. All training data was utilized to build the model, while the testing data remained untouched during the training process, serving solely for model generalization. We compared the model's estimations with the true solar radiation observations, and the fitted results are presented in **Figure 2**. The similar performance on both training and testing datasets indicates that there are no issues of overfitting or underfitting. The resulting Root Mean Square Error (RMSE) values are 131.89 W/m^2 for the training set and 134.75 W/m^2 for the testing set.

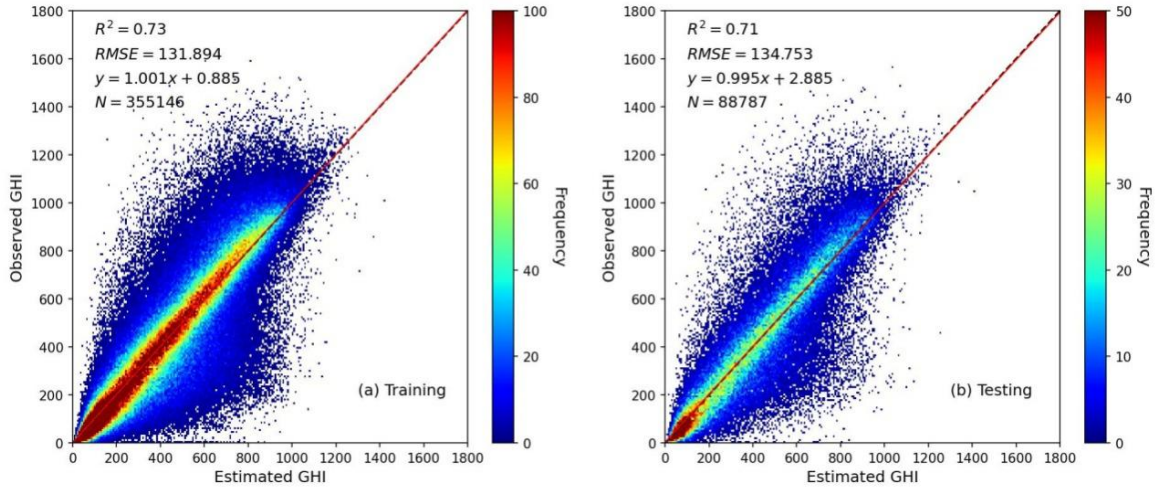

Figure 2. Density scatter plot of model fitting validation on training (left) and testing (right) dataset.

Note: The dot color represents the counts of data points. The black dashed line is the 1:1 line. The red line is the fitted line of estimation and observation.

3.2. Clear/Cloudy Validation

It can be observed that the estimation results from these three sites are close to those obtained from the training and testing datasets, indicating good generalization of the model. Even though the data from these 3 sites were not used for model training, satisfactory performance was still achieved. Among these sites, the estimates most closely matched actual observations at Guilin, followed by Beihai, and then Nanning. We also validated the results under clear and cloudy conditions separately. The results are illustrated in **Figure 3**. When evaluated using RMSE, the model appears to perform comparably or better under cloudy conditions: at Guilin, cloudy-sky RMSE (127.78 W/m^2) is slightly lower than clear-sky RMSE (134.35 W/m^2); at Nanning, the difference is similarly small (154.34 vs. 156.54 W/m^2); and at

Beihai, cloudy-sky RMSE (144.25 W/m^2) is slightly higher than the clear-sky value (138.39 W/m^2). It should be noted that RMSE as an absolute metric is inherently sensitive to the magnitude of the variable: under clear-sky conditions, GHI values are intrinsically higher, which inflates absolute errors even when relative model accuracy is comparable or superior. To provide a more physically meaningful comparison, MAPE was used as well as RMSE and CC. The MAPE results reveal that clear-sky MAPE values are substantially lower than their cloudy-sky counterparts at all three stations —Guilin (31.46% vs. 76.68%), Nanning (26.38% vs. 56.41%), and Beihai (38.33% vs. 54.43%). These results confirm that the model achieves smaller relative errors under clear-sky conditions, and that RMSE alone substantially underestimates the difficulty of cloudy-sky estimation due to the lower absolute GHI values during overcast periods. The similarity in RMSE across sky conditions also needs to be interpreted alongside the substantial sample size imbalance. As shown in **Figure 3**, there is a significant disparity in the number of cloudy versus clear samples throughout the year 2018 for these 3 sites. Cloudy samples outnumber clear-sky samples by approximately 3:1 at Guilin (N = 2,477 vs. 789) and Nanning (N = 2,479 vs. 770), and by as high as 10:1 at Beihai (N = 3,002 vs. 301). Therefore, the better performance under cloudy conditions in terms of RMSE is likely a combined effect of the lower absolute GHI values during overcast periods and the larger sample sizes for cloudy conditions.

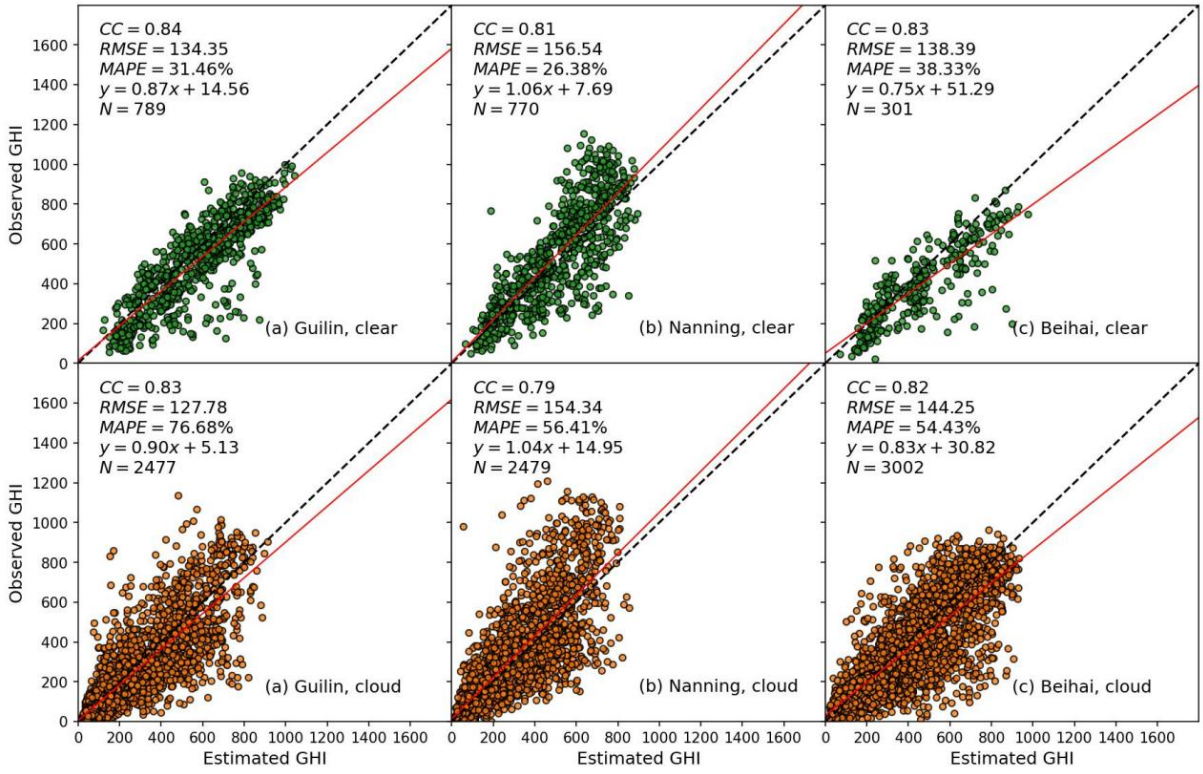

Figure 3. Scatter plot illustrating the validation results for data from 3 stations in Guangxi during 2018.
Note: The upper panel displays results under clear sky conditions, while the lower panel represents cloudy conditions. Clear and cloudy classifications are determined using the cloud mask product from FY-4A.

3.3. Seasonal Validation

Building on these initial findings, we further examined the model's performance under varying seasonal conditions. We evaluated the model's performance across the four seasons: spring, summer, autumn, and winter. As shown in **Figure 4**, the model's estimation results for the 3 stations were best in winter and least accurate in summer. The performance during spring and autumn varied; specifically, in Guilin and Nanning, the autumn results were superior to those in spring, while in Beihai, the opposite was true, with smaller estimation errors observed in spring.

The model's optimal performance in winter can be attributed to several concurring factors. First, winter is char-

acterized by lower absolute GHI values compared to summer, which narrows the dynamic range of the target variable and simplifies the prediction task. It should be noted that the lower RMSE observed under cloudy conditions in Section 3.2 is largely a statistical artifact driven by sample size imbalance—cloudy samples outnumber clear-sky samples by ratios of up to 10:1 at Beihai—rather than evidence of inherently superior model capability under overcast skies. At the seasonal scale, the key driver of winter superiority is the reduced frequency and intensity of convective cloud activity, which leads to more temporally stable irradiance fields that are intrinsically easier to predict. In contrast, summer is dominated by rapidly developing, optically thick convective clouds that introduce large sub-hourly fluctuations in surface irradiance, exceeding the representational capacity of the 1-hour resolution inputs used in this model.

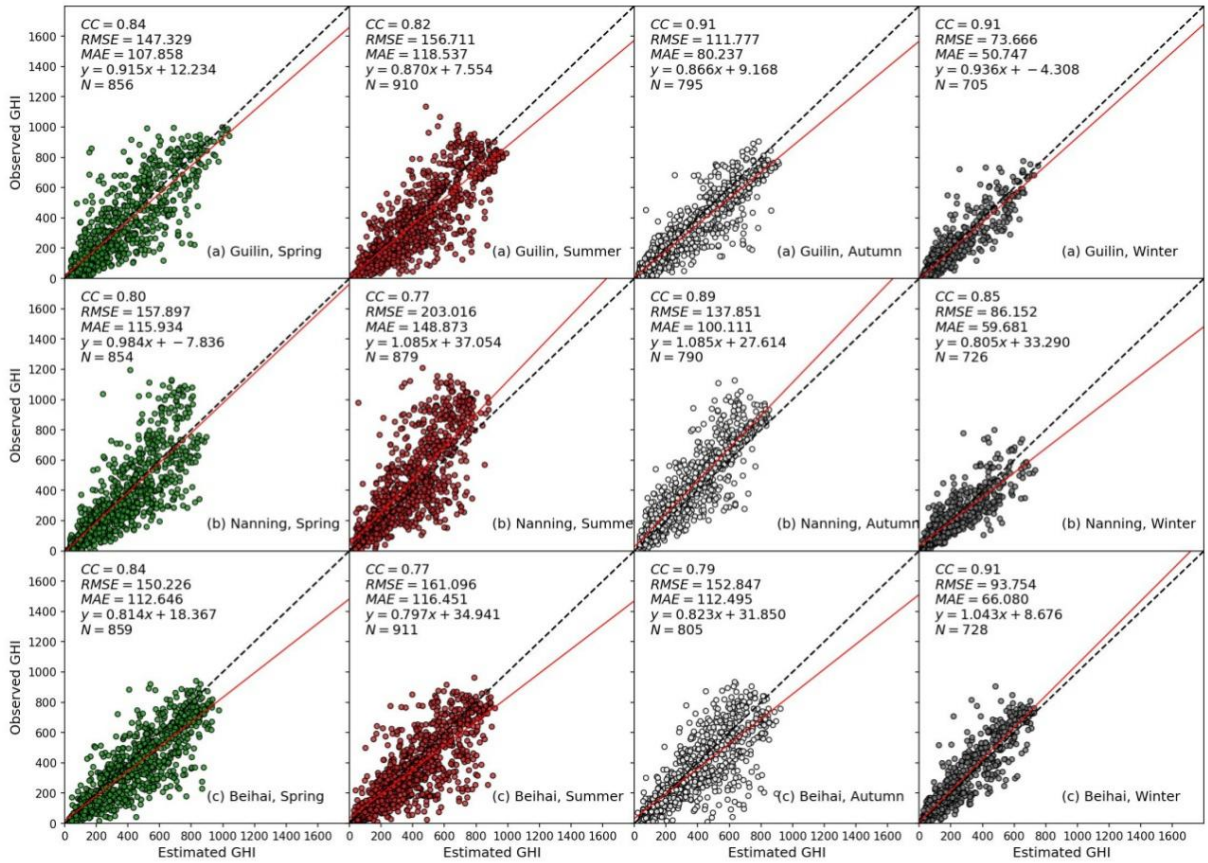

Figure 4. Scatter plot illustrating the seasonal validation results for data from 3 stations in Guangxi during 2018.

Note: The four panels display results for spring, summer, autumn, and winter, respectively.

3.4. Hourly Validation

In order to systematically assess the influence of both geographical and temporal variations on model accuracy, we evaluated the model's performance at different locations and times. As shown in **Figure 5**, the Guilin station exhibited the best performance during the 09–11 Local Time Clock (LTC) period, while its performance was weakest during the 15–16 LTC period; the model's accuracy during other times was relatively consistent. In contrast, the Nanning station performed optimally during the 07, 08, and 16 LTC periods, but its performance was least effective during the 15 and 12 LTC periods. The Beihai station demonstrated its best performance during the 12–13 LTC period, with the weakest performance again noted in the 15–16 LTC period, while its accuracy remained consistent during other times. Notably, the performance of the Guilin and Beihai stations was more closely aligned compared to the other locations. This analysis highlights how geographical and temporal variations can influence model accuracy, emphasizing the need for localized calibration in solar radiation estimation models.

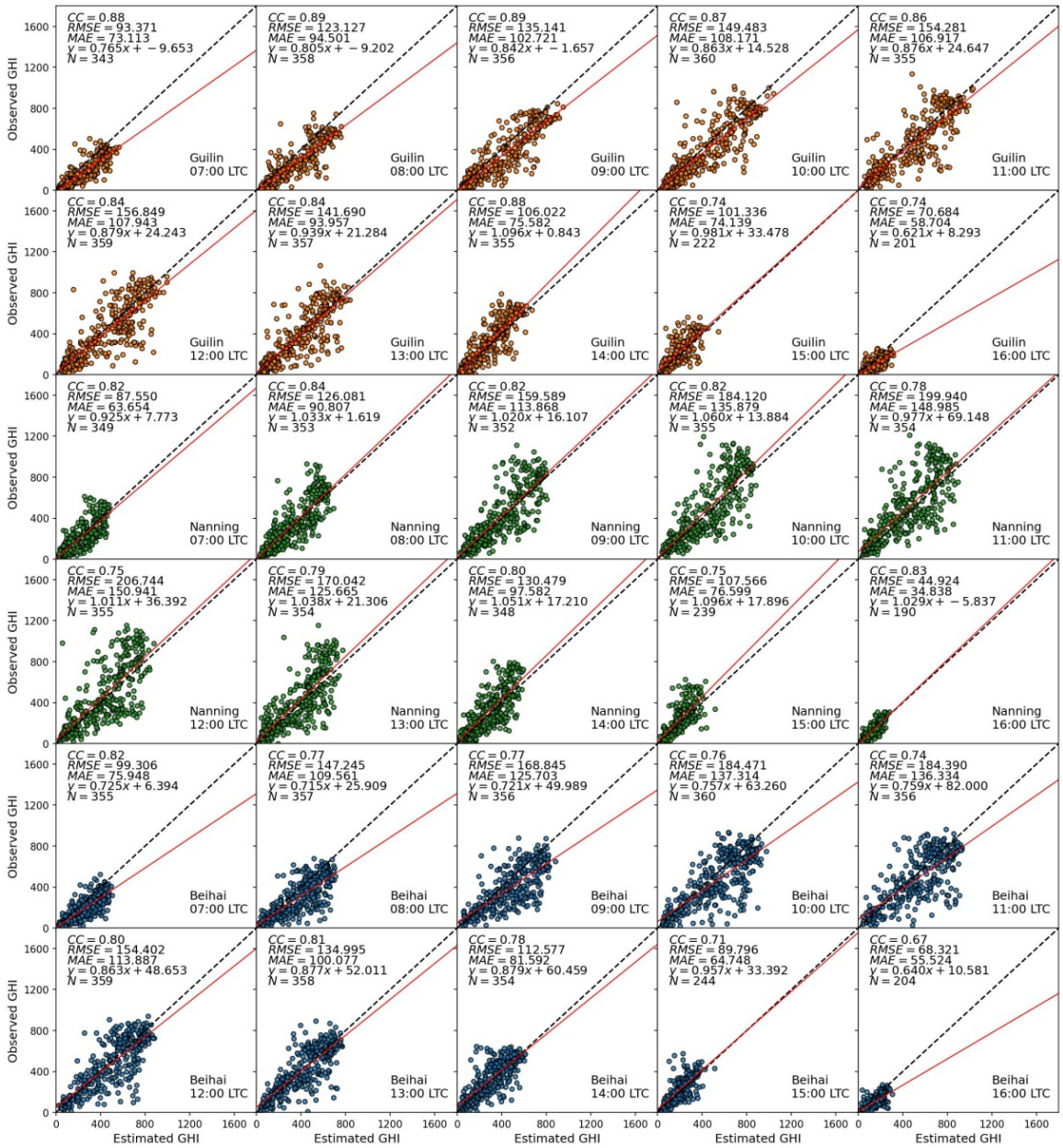

Figure 5. Scatter plot illustrating the hourly validation results for data from 3 stations in Guangxi during 2018.

Note: Each panel corresponds to a specific local time (LTC) interval.

3.5. Feature Importance

Feature importance is a critical concept in machine learning that quantifies the contribution of individual features (or variables) to the predictions made by a model. It provides a score for each feature, indicating its relative importance in influencing the target variable. By identifying which features are most influential, feature importance aids in understanding the underlying relationships within the data, enhancing model interpretability, and facilitating feature selection. This technique is particularly valuable for improving model performance and reducing complexity. By focusing on the most important features and eliminating those with minimal impact, data scientists can streamline their models, reduce overfitting, and improve computational efficiency. Additionally, feature importance plays a crucial role in explainable AI (XAI), allowing stakeholders to comprehend how decisions are made by machine learning models. Various methods exist for calculating feature importance, including permutation importance and tree-based metrics, each offering unique insights into the model's behavior.

From **Figure 6**, it is evident that the observation of brightness temperature in Band 11 (central wavelength 7.42 μm) has the most significant impact on radiation estimation. Following this, the solar zenith angle and relative humidity also play important roles. Band 11, which typically corresponds to a specific wavelength range sensitive to mid-level water vapor, provides crucial information that influences the accuracy of solar radiation estimates. The solar zenith angle, representing the angle between the sun and the vertical direction at a given location, affects the intensity and distribution of solar radiation reaching the Earth's surface. Meanwhile, relative humidity is essential because it can alter the atmospheric absorption and scattering of solar radiation. The observation of reflectance in Band 1 is also important for estimation, as it represents the combined contribution of the atmosphere along with the surface (clear-sky) or clouds (cloudy-sky). Together, these factors highlight the complex interactions that determine solar irradiance and underscore the importance of incorporating multiple meteorological parameters for accurate radiation modeling.

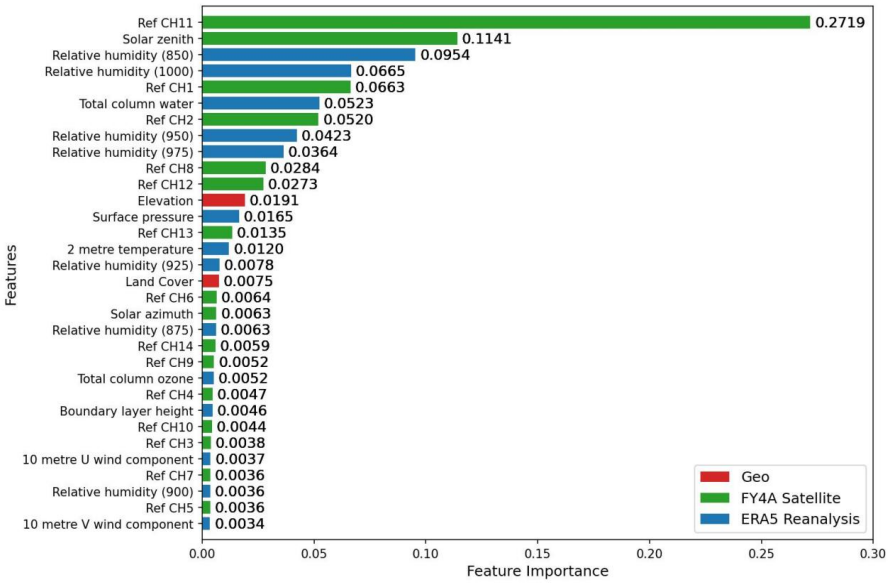

Figure 6. Feature importance provided by the XGBoost model.

4. Discussion

4.1. Physical Interpretability of Feature Importance

A critical advantage of the proposed XGBoost model is its ability to reveal the physical drivers of GHI estimation through feature importance analysis. Our results identified brightness temperature at 7.42 μm, solar zenith angle, relative humidity and TOA reflectance at 0.47 μm as the most influential predictors. From a meteorological perspective, the 7.42 μm band corresponds to the mid-level water vapor band in the FY-4A AGRI. Its high feature weight underscores the dominant role of atmospheric moisture and cloud liquid water path in modulating solar radiation attenuation. Furthermore, the strong dependence on SZA and RH aligns with the physical principles of atmospheric transmissivity, corroborating previous findings that aerosol optical properties and atmospheric stability are pivotal in determining surface irradiance [31–33]. It is also noteworthy that visible and near-infrared bands, despite carrying information on cloud optical thickness and surface albedo, ranked lower in overall importance than the water vapor band. This suggests that, in the humid subtropical climate of Guangxi, moisture-driven attenuation of incoming solar radiation is a more persistent and dominant process. The planetary boundary layer height, while ranked lower among the ERA5 variables, still contributed meaningfully to the model, reflecting its role in regulating the vertical extent of aerosol-laden air masses. Collectively, the feature importance ranking provides a physically coherent explanation for the model's predictive structure and supports the rationality of the 31-feature input design adopted in this study.

4.2. Error Analysis and Spatiotemporal Variability

While the model generalized effectively across the Guilin, Nanning, and Beihai stations, the spatiotemporal variability of the estimation errors requires careful examination. The performance degradation observed during summer and under cloudy-sky conditions can be attributed to the complex convective cloud dynamics typical of Guangxi's subtropical monsoon climate. During summer, rapid formations of highly localized, optically thick cumulus clouds introduce significant spatio-temporal heterogeneity. The ERA5 reanalysis data, despite its reliability, inherently struggles to resolve these sub-grid scale micro-physical processes due to its relatively coarse spatial resolution. Consequently, the mismatch between the spatial footprint of satellite/reanalysis grids and the point-based ground observations becomes more pronounced under fragmented cloud cover, leading to increased localized prediction errors. Furthermore, the inter-station differences in error magnitude observed across Guilin, Nanning, and Beihai can be partially attributed to differences in local topography and land cover. Guilin is surrounded by karst terrain with complex surface heterogeneity, which may amplify the representativeness error when comparing a spatially averaged satellite pixel with a point observation. Beihai, located on the coast, is subject to marine boundary layer influences and sea-breeze-driven cloud formation patterns that are difficult to fully capture with ERA5 reanalysis at its native resolution. Nanning, situated in a broad inland basin, experiences relatively more homogeneous surface conditions, yet its lower CC values under certain sky conditions suggest sensitivity to systematic biases in reanalysis humidity fields during the wet season. These findings imply that future model improvements could benefit from incorporating higher-resolution ancillary data, such as land surface temperature from geostationary sensors or high-resolution terrain indices, to reduce representativeness errors at heterogeneous sites.

4.3. Comparative Performance

To evaluate the advancement of our approach, the performance of the proposed XGBoost model was systematically compared with existing state-of-the-art models, particularly the satellite-based semi-empirical approach applied to FY-4A developed by Jia et al.[1]. In their study, Jia et al. utilized FY-4A satellite data combined with the cloud index methodology (CSD-SI) and the McClear clear-sky model to estimate GHI in Northern China. While their physics-based model demonstrated robust estimation under clear-sky conditions, it inherently relies on the intermediate derivation of cloud indices and idealized clear-sky baselines. This two-step derivation can introduce cascading errors under complex, fragmented cloud cover—a limitation they explicitly observed through increased normalized root mean square errors (nRMSE) during summer and autumn.

In contrast, our ensemble learning framework bypasses the need for explicit intermediate cloud parameterizations. By directly ingesting multi-spectral TOA reflectance alongside ERA5 meteorological parameters (e.g., relative humidity and planetary boundary layer height), the XGBoost algorithm effectively captures the highly non-linear interactions between atmospheric moisture, aerosol scattering, and surface radiation. This direct data-driven approach not only mitigates the error propagation typical of semi-empirical models but also exhibits distinct advantages in adaptability under dynamic cloudy conditions. Consequently, while Jia et al.[1] provide a strong physical baseline, our results demonstrate that integrating multi-source meteorological data through ensemble machine learning offers a highly competitive, scalable, and computationally efficient alternative for regional solar resource assessment, particularly in subtropical climates with complex convective cloud formations like Guangxi. It is worth noting that the training dataset in this study covers the full range of climate zones and land cover types across China, which inherently exposes the model to diverse atmospheric conditions during training. This broad training distribution is a key reason why the model generalizes effectively to Guangxi stations despite their geographic specificity. Moreover, the computational efficiency of XGBoost relative to deep learning alternatives makes it particularly attractive for operational deployment in power dispatching centers, where near-real-time solar resource assessment is required. The model's training time was approximately 1/7 of that required by a comparable Random Forest model under identical hyperparameter settings, demonstrating a clear practical advantage for large-scale implementation.

5. Conclusion

In this study, an ensemble learning framework based on the XGBoost algorithm was developed and validated for estimating global horizontal irradiance (GHI) across Guangxi Province. By synergistically integrating FY-4A satellite TOA reflectance and brightness temperature, ERA5 meteorological reanalysis, and geographical features, the model demonstrated robust predictive capabilities and strong spatial generalization, effectively estimating GHI at validation sites not explicitly included in the training phase.

The main conclusions of this study are as follows:

1. **Robust Generalization**: The model achieved high agreement with ground-based observations across different geographical zones in Guangxi, confirming the viability of the selected feature space for regional GHI mapping.
2. **Condition-Dependent Accuracy**: Model performance is highly sensitive to seasonal and cloud-cover variations. Optimal accuracy was recorded during winter, while performance across sky conditions varied by station and was strongly influenced by sample size distribution between clear-sky and cloudy observations, whereas summer and cloudy periods presented greater estimation challenges due to complex atmospheric scattering and dynamic cloud properties.
3. **Key Atmospheric Drivers**: Feature importance analysis validated the physical consistency of the AI model, highlighting mid-level water vapor (7.42 μm observations), solar zenith angle, and relative humidity as the primary modulators of surface solar radiation.

Despite the promising results, limitations remain regarding the reliance on historical reanalysis data, which may introduce latency and spatial resolution biases. To further enhance model adaptability and support grid operations, future research will focus on developing ultra-short-term (e.g., 0–4 h) solar radiation forecasting models. This will involve transitioning to higher-resolution real-time data from new-generation satellites (such as FY-4B) and constructing hybrid architectures that embed physical radiative transfer processes into advanced machine learning algorithms. Ultimately, the insights and methodologies established in this study provide a scalable foundation for optimizing photovoltaic system deployment and advancing renewable energy integration strategies.

Author Contributions

Conceptualization, J.H., Y.Q. and Q.Y.; methodology, J.H. and Q.Y.; software, J.H., Q.Y. and K.H.; validation, H.Z., J.L. and Y.Z.; formal analysis, J.H.; investigation, Y.Q.; resources, H.Q.; data curation, H.Q.; writing—original draft preparation, J.H.; writing—review and editing, Y.Q.; visualization, J.H. and Q.Y.; supervision, J.T.; project administration, J.T.; funding acquisition, J.T. All authors have read and agreed to the published version of the manuscript.

Funding

This work is supported by the Guangxi Power Grid Company's Science and Technology Project: "Research on Correction of Key Meteorological Elements for Photovoltaics Based on Satellite Monitoring Data and Solar Irradiance Forecasting Technology" (Program No. GXKJXM20230261).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study are available from the China Meteorological Administration (CMA) but are not publicly accessible due to institutional restrictions. Requests to access the data may be directed to the National Satellite Meteorological Center. FY-4A satellite data are publicly available at: http://satellite.nsmc.org.cn/portalsite/default.aspx. ERA5 reanalysis data are publicly available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analy-

ses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

AI Use Statement

During the preparation of this work, the authors used DeepSeek for language polishing. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

[1] Jia, D., Hua, J., Wang, L., et al., 2021. Estimations of Global Horizontal Irradiance and Direct Normal Irradiance by Using Fengyun-4A Satellite Data in Northern China. Remote Sensing. 13(4), 790. DOI: https://doi.org/10.3390/rs13040790

[2] Ruiz-Arias, J.A., Gueymard, C.A., 2018. Worldwide Inter-Comparison of Clear-Sky Solar Radiation Models: Consensus-Based Review of Direct and Global Irradiance Components Simulated at the Earth Surface. Solar Energy. 168, 10–29. DOI: https://doi.org/10.1016/j.solener.2018.02.008

[3] Olmo, F.J., Vida, J., Foyo, I., et al., 1999. Prediction of Global Irradiance on Inclined Surfaces from Horizontal Global Irradiance. Energy. 24(8), 689–704. DOI: https://doi.org/10.1016/S0360-5442(99)00025-0

[4] Li, D.H., Lou, S., 2018. Review of Solar Irradiance and Daylight Illuminance Modeling and Sky Classification. Renewable Energy. 126, 445–453. DOI: https://doi.org/10.1016/j.renene.2018.03.063

[5] Zhou, Y., Liu, Y., Wang, D., et al., 2021. A Review on Global Solar Radiation Prediction with Machine Learning Models in a Comprehensive Perspective. Energy Conversion and Management. 235, 113960. DOI: https://doi.org/10.1016/j.enconman.2021.113960

[6] Ramadhan, R.A.A., Heatubun, Y.R.J., Tan, S.F., et al., 2021. Comparison of Physical and Machine Learning Models for Estimating Solar Irradiance and Photovoltaic Power. Renewable Energy. 178, 1006–1019. DOI: https://doi.org/10.1016/j.renene.2021.06.079

[7] Benamrou, B., Ouardouz, M., Allaouzi, I., et al., 2020. A Proposed Model to Forecast Hourly Global Solar Irradiation Based on Satellite Derived Data, Deep Learning and Machine Learning Approaches. Journal of Ecological Engineering. 21(4), 26–38. DOI: https://doi.org/10.12911/22998993/119795

[8] Shi, H., Yang, D., Wang, W., et al., 2023. First Estimation of High-Resolution Solar Photovoltaic Resource Maps over China with Fengyun-4A Satellite and Machine Learning. Renewable and Sustainable Energy Reviews. 184, 113549. DOI: https://doi.org/10.1016/j.rser.2023.113549

[9] Journée, M., Bertrand, C., 2010. Improving the Spatio-Temporal Distribution of Surface Solar Radiation Data by Merging Ground and Satellite Measurements. Remote Sensing of Environment. 114(11), 2692–2704. DOI: https://doi.org/10.1016/j.rse.2010.06.010

[10] Letu, H., Yang, K., Nakajima, T.Y., et al., 2020. High-Resolution Retrieval of Cloud Microphysical Properties and Surface Solar Radiation Using Himawari-8/AHI Next-Generation Geostationary Satellite. Remote Sensing of Environment. 239, 111583. DOI: https://doi.org/10.1016/j.rse.2019.111583

[11] Huang, L., Kang, J., Wan, M., et al., 2021. Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events. Frontiers in Earth Science. 9, 596860. DOI: https://doi.org/10.3389/feart.2021.596860

[12] Zang, H., Liu, L., Sun, L., et al., 2020. Short-Term Global Horizontal Irradiance Forecasting Based on a Hybrid CNN-LSTM Model with Spatiotemporal Correlations. Renewable Energy. 160, 26–41. DOI: https://doi.org/10.1016/j.renene.2020.05.150

[13] Rajaprasad, S., Mukkamala, R., 2023. A Hybrid Deep Learning Framework for Modeling the Short Term Global Horizontal Irradiance Prediction of a Solar Power Plant in India. Polityka Energetyczna – Energy Policy Journal. 26(3), 101–116. DOI: https://doi.org/10.33223/epj/168115

[14] El-Shahat, D., Tolba, A., Abouhawwash, M., et al., 2024. Machine Learning and Deep Learning Models Based Grid Search Cross Validation for Short-Term Solar Irradiance Forecasting. Journal of Big Data. 11, 134. DOI: https://doi.org/10.1186/s40537-024-00991-w

[15] Tan, Y., Wang, Q., Zhang, Z., 2023. Near-Real-Time Estimation of Global Horizontal Irradiance from Himawari-8 Satellite Data. Renewable Energy. 215, 118994. DOI: https://doi.org/10.1016/j.renene.2023.118994

[16] Suwanwimolkul, S., Tongamrak, N., Thungka, N., et al., 2025. Deep-Learning-Based and Near Real-Time Solar Irradiance Map Using Himawari-8 Satellite Imageries. Solar Energy. 288, 113262. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0038092X25000258

[17] Nadeem, A., Hanif, M.F., Naveed, M.S., et al., 2024. AI-Driven Precision in Solar Forecasting: Breakthroughs in Machine Learning and Deep Learning. AIMS Geosciences. 10(4), 684–734. DOI: https://doi.org/10.3934/geosci.2024035

[18] Zhu, S., Ma, Z., 2021. Does AGRI of FY4A Have the Ability to Capture the Motions of Precipitation? IEEE Geoscience and Remote Sensing Letters. 19, 1–5. DOI: https://doi.org/10.1109/LGRS.2021.3130646

[19] Gao, Y., Mao, D., Wang, X., et al., 2022. Evaluation of FY-4A Temperature Profile Products and Application to Winter Precipitation Type Diagnosis in

Southern China. Remote Sensing. 14(10), 2363. DOI: https://doi.org/10.3390/rs14102363

[20] Xu, F., Song, B., Chen, J., et al., 2024. Deep-Learning-Based Daytime COT Retrieval and Prediction Method Using FY4A AGRI Data. Remote Sensing. 16(12), 2136. DOI: https://doi.org/10.3390/rs16122136

[21] Yang, J., Zhang, Z., Wei, C., et al., 2017. Introducing the New Generation of Chinese Geostationary Weather Satellites, Fengyun-4. Bulletin of the American Meteorological Society. 98, 1637–1658. DOI: https://doi.org/10.1175/BAMS-D-16-0065.1

[22] Du, S., Zhang, Y., Wei, W., et al., 2021. Analysis of Influence of Solar Zenith Angle on Reconstruction of Hyperspectral Surface Reflectance. Acta Optica Sinica. 41(2), 0229001. DOI: https://doi.org/10.3788/AOS202141.0229001 (in Chinese)

[23] Liu, H., Hu, B., Zhang, L., et al., 2017. Ultraviolet Radiation over China: Spatial Distribution and Trends. Renewable and Sustainable Energy Reviews. 76, 1371–1383. DOI: https://doi.org/10.1016/j.rser.2017.03.102

[24] Gao, X.-Y., Liu, J.-M., Yuan, Y., et al., 2023. Global Horizontal Irradiance Prediction Model Considering the Effect of Aerosol Optical Depth Based on the Informer Model. SSRN. DOI: http://dx.doi.org/10.2139/ssrn.4522731

[25] Shrestha, A.K., Thapa, A., Gautam, H., 2019. Solar Radiation, Air Temperature, Relative Humidity, and Dew Point Study: Damak, Jhapa, Nepal. International Journal of Photoenergy. 2019(1), 8369231. DOI: https://doi.org/10.1155/2019/8369231

[26] Bilgili, M., Ozgoren, M., 2011. Daily Total Global Solar Radiation Modeling from Several Meteorological Data. Meteorology and Atmospheric Physics. 112, 125–138. DOI: https://doi.org/10.1007/s00703-011-0137-9

[27] Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[28] Taylor, K.E., 2001. Summarizing Multiple Aspects of Model Performance in a Single Diagram. Journal of Geophysical Research: Atmospheres. 106(D7), 7183–7192. DOI: https://doi.org/10.1029/2000JD900719

[29] Willmott, C.J., Matsuura, K., 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. Climate Research. 30(1), 79–82. DOI: https://doi.org/10.3354/cr030079

[30] Chai, T., Draxler, R.R., 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments against Avoiding RMSE in the Literature. Geoscientific Model Development. 7(3), 1247–1250. DOI: https://doi.org/10.5194/gmd-7-1247-2014

[31] Cronin, T.W., 2014. On the Choice of Average Solar Zenith Angle. Journal of the Atmospheric Sciences. 71(8), 2994–3003. DOI: https://doi.org/10.1175/JAS-D-13-0392.1

[32] Qiao, C., Liu, S., Huo, J., et al., 2023. Retrievals of Precipitable Water Vapor and Aerosol Optical Depth from Direct Sun Measurements with EKO MS711 and MS712 Spectroradiometers. Atmospheric Measurement Techniques. 16(6), 1539–1549. DOI: https://doi.org/10.5194/amt-16-1539-2023

[33] Teillet, P.M., Fedosejevs, G., Ahern, F.J., et al., 1994. Sensitivity of Surface Reflectance Retrieval to Uncertainties in Aerosol Optical Properties. Applied Optics. 33(18), 3933–3940. DOI: https://doi.org/10.1364/AO.33.003933