

First estimation of high-resolution solar photovoltaic resource maps over China with Fengyun-4A satellite and machine learning

Hongrong Shi^{a,b}, Dazhi Yang^{c, **}, Wenting Wang^c, Disong Fu^a, Ling Gao^d, Jinqiang Zhang^a, Bo Hu^e, Yunpeng Shan^f, Yingjie Zhang^g, Yuxuan Bian^b, Hongbin Chen^a, Xiangao Xia^{a,*}

^a Key Laboratory of Middle Atmosphere and Global Environment Observation, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

^b State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

^c School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China

^d National Satellite Meteorological Center, China Meteorological Administration, Beijing, China

^e State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

^f Pacific Northwest National Laboratory, Richland, WA, USA

^g School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, 100083, China



ARTICLE INFO

Keywords:

Fengyun-4A
Solar photovoltaic
Machine learning
Surface solar radiation

ABSTRACT

Fengyun-4A (FY-4A), which is the latest-generation Chinese geostationary meteorological satellite, measures solar reflection and thermal emission with high temporal, spatial, and spectral resolutions. It is expected to be highly beneficial for solar resource assessment and forecasting in China. This study is the first to estimate, using FY-4A and a random forest model, the global horizontal irradiance (GHI) at a 4-km-15-min spatio-temporal resolution over China, as a means to arrive at a solar photovoltaic (PV) resource map. In terms of GHI estimates, the root mean square error and mean bias error between hourly measured and retrieved values are 147.02 (35.2%), -5.64 W/m^2 (-1.4%), respectively, whereas the values of daily estimates are 29.20 (18.0%), -2.97 W/m^2 (-1.3%). The retrieval accuracy is found much better for instances with solar zenith angles smaller than 60° . Relatively larger errors are found at locations in the Sichuan Basin and northeastern China, which can be attributed to bright surfaces and/or strong cloud transients. With the retrieved irradiance, PV resource is derived through a physical model chain. The annual mean PV resource map suggests that, over most of the west areas, the annual mean effective irradiance exceeds 1700 kWh/m^2 , with the highest value found in Tibet (around 2000 kWh/m^2 per annum). Eastern China has an annual effective irradiance of only $1300\text{--}1500 \text{ kWh/m}^2$. The region with poorest solar resource is the Sichuan Basin (less than 1100 kWh/m^2 per annum).

1. Introduction

China announced its grand goals of reaching its peak carbon emissions by 2030 and carbon neutrality by 2060. To meet such goals, lowering carbon emission while boosting carbon sink capacity is

unprecedentedly urgent. Clearly then, one can fully anticipate a substantial ramp up of installed capacity of wind and solar over the coming decades. Indeed, the installation of photovoltaic (PV), which is a major form of renewable and clean energy conversion technology with no operational emission of carbon dioxide or other air pollutants, has rocketed exponentially over the past decades worldwide. An enormous

Abbreviations: AGRI, the Advanced Geosynchronous Radiation Imager; BNI, beam normal irradiance; BTI, beam irradiance on the tilted surface; CAMS, Copernicus Atmosphere Monitoring Service; the CERES, the Clouds and the Earth's Radiant Energy System; CERN, the Chinese Ecosystem Research Network; CMA, the China Meteorological Administration; CNN, convolutional neural network; DHI, diffuse horizontal irradiance; DISC, Direct Insolation Simulation Code; DNN, deep neural network; DTI, diffused irradiance on the tilted surface; DTs, decision trees; FY-4A, Fengyun-4A geostationary satellite; GBDT, gradient boosting decision tree; GHI, global horizontal irradiance; GTI, global tilted irradiance; MBE, mean bias error; MLP, multi-layer perceptron; MSG, Meteosat Second Generation satellite; NSRDB, the National Solar Radiation Database; POA, plane-of-array; PV, photovoltaic; R, correlation coefficient; RF, random forest; RMSE, root mean square error; RTM, radiative transfer model; SHAP, Shapley Additive explanation; STC, standard test condition; VIF, variance inflation factor; WCRP, the World Climate Research Program.

* Corresponding author.

** Corresponding author.

E-mail addresses: yangdazhi.nus@gmail.com (D. Yang), xxa@mail.iap.ac.cn (X. Xia).

Nomenclature		
Indexes		
<i>i</i>	Index for time	
Functions		
f_1	Spectral loss coefficient	
Variables		
ρ	Albedo of the foreground	
B_n	Beam normal irradiance	
B_c	Beam radiation	
κ	Clear-sky index	k_t Clearness index
	G_{hc} Clear-sky expectation of global horizontal irradiance	G_{hc} Clear-sky index
D_h	Diffuse horizontal irradiance	k_t Clearness
D_c	Diffuse radiation	
R_d	Diffuse transposition factor	
E_0	Extraterrestrial global horizontal irradiance	
G_h	Global horizontal irradiance	
Parameters		
a_0, a_1, a_2, a_3, a_4	Coefficients from Canadian Solar CS5P-220 M	
n_{PV}	PV module's refraction index	
θ'	Refractive angle obtainable via the Snell's law	
n_T	Refraction index of the pyranometer cover	
F_1, F_2	Sky-dependent model coefficient	

expansion of PV electricity generation in China has been planned, that is, a goal of providing no less than 10% of total electricity demand by 2030. One of essential steps moving forward is to attain high-quality solar energy resource maps over China, which are to be of support to sizing, siting, and performance evaluation of PV systems, and thereby facilitating massive grid integration of solar [1].

The global horizontal irradiance (GHI), which is also known by meteorologists as the downward shortwave surface radiation, is a parameter with the foremost importance for solar resource assessment. It can be typically acquired from three complementary sources: (1) ground-based radiometry, (2) dynamical weather models, and (3) remote-sensing products [2]. Ground-based radiometry is unarguably the preferred means to obtain long-term GHI data, insofar as accuracy is of primary concern. Measurements from networks of regularly calibrated and well-maintained radiometers—such as the two operational networks in China, i.e., the China Meteorological Administration (CMA) radiation observation network and the Chinese Ecosystem Research Network (CERN), provide hourly GHI measurements at 128 and 38 stations, respectively [3–5]—are often used as the standard of reference to gauge other irradiance estimates. Historically, practices which seek to correlate GHI with proxy measurements such as sunshine duration were once popular, although the correlations were empirical and hence were associated with large uncertainties [6]. In fact, concerns in regard to sunshine-duration-based irradiance estimates have long been voiced out by expert solar energy meteorologists, formally, first in 1993 and then in 2009 [7,8]. (Most top journals have stopped accepting the Angström–Prescott type of radiation models long ago [9].) More generally, even if some gridded GHI maps can be attained from these point-location ground measurements, perhaps with some simple and some not so simple interpolation methods, the accuracy of those maps would not be sufficient to support modern resource assessment at any rate. As the interpolation of ground-based measurements, in all meteorology problems of this sort, is thought to be a perpetually suboptimal approach with respect to other means of acquiring gridded data, dynamical weather modeling and remote sensing are now, no doubt, commonplaces.

Weather modeling is able to provide GHI data with wide-area coverage and over long time periods. For example, the latest MERRA-2 (the Modern-Era Retrospective Analysis for Research and Applications, Version 2) reanalysis provides hourly GHI estimates, with a $0.625^\circ \times 0.5^\circ$ spatial resolution, in longitude and latitude, over the entire globe. However, the high uncertainty associated with reanalysis

data prevents them from being widely applied during resource assessment exercises [10,11]. In comparison, remote sensing can provide decade-long gridded GHI estimates that are far more reliable. Nonetheless, one should note that there are several well-known synoptic-scale satellite-derived irradiance datasets, such as the Clouds and the Earth's Radiant Energy System (CERES) product and the World Climate Research Program (WRCP) surface radiation budget project product, just to name a few, which provide GHI at a 100-km spatial resolution [12], and therefore are too coarse to cater for modern resource assessment purposes, which require km-range data. To that end, the kind of satellite-derived irradiance estimates concerning the present study ought to have on a near-microscale horizontal resolution (<5 km, 0.5 km on visible band and 1–4 km on others), of which the provision is supported by the latest-generation geostationary satellites.

Geostationary satellites with onboard radiometers with near-microscale spatial resolutions had been successfully launched by European and American space agencies 20 years ago, namely, Meteosat Second Generation (MSG) [13] and GOES-R [14]. Correspondingly, satellite-derived irradiance products such as the Copernicus Atmosphere Monitoring Service (CAMS) radiation service [15] and the National Solar Radiation Database (NSRDB [16]) now consist of about 20 years of high-spatial-resolution data. In Asia, the progress in remote sensing lagged that of the West by about a decade, but since the launches of the Japanese Himawari-7/-8, the gap has been substantially narrowed. For instance, 2-km products from Himawari-7/-8 are now available in NSRDB, and have a time span of 2011–2020. The Japan Meteorological Agency has also produced their own Himawari-based product.

Despite the aforementioned progress in Asian satellite-derived irradiance products, Himawari-8 does not cover the western and middle Asia in its entirety, due to its geostationary positioning (140° E). Although the western and middle Asia is covered by Meteosat-9 (replacing Meteosat-8 from 1 June 2022), which is positioned at 45.5° E, there is still a fatal problem. It is well known that, owing to the unresolved observation, all satellite-derived irradiance products lose accuracy towards the boundaries of the satellites' field-of-view disks [17, 18]. Hence, performing effective solar resource assessment over regions near the disk boundaries has hitherto been challenging. Fig. 1 depicts the coverage of geostationary weather satellites over China.

Fengyun-4A (FY-4A), which was launched in 2016, is the first of the latest-generation Chinese geostationary meteorological satellites. One of major payloads is the Advanced Geosynchronous Radiation Imager (AGRI), which is able to measure spectral solar reflectance and thermal

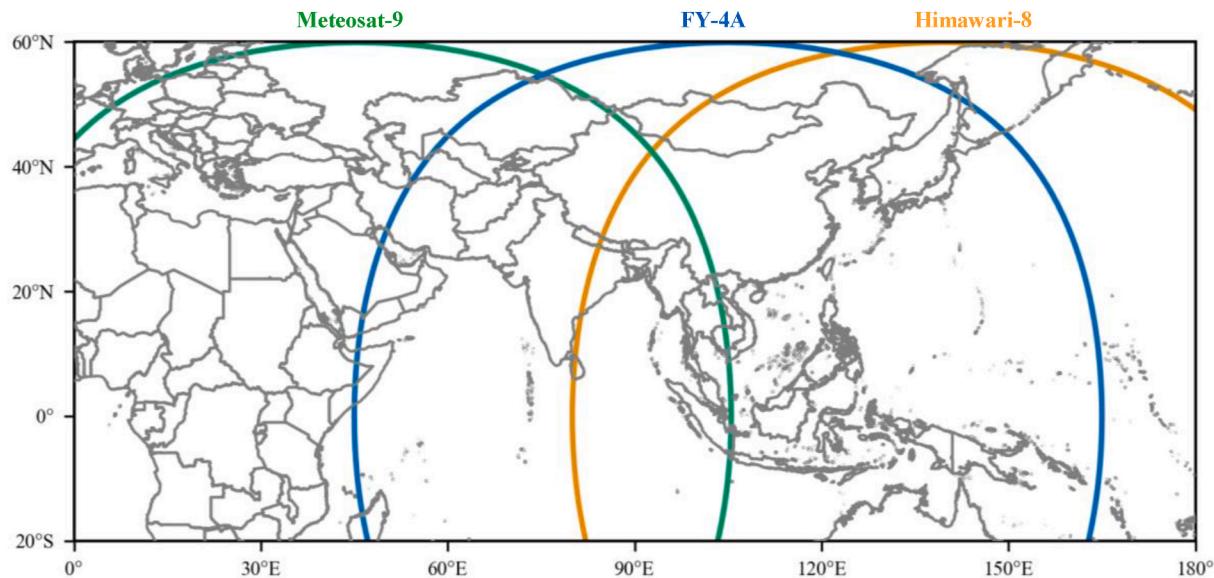


Fig. 1. Geostationary weather satellites provide images that cover China domain from the Europe (Meteosat-9), Japan (Himawari-8), and China (FY-4A).

emission with high spatio-temporal resolution. This provides an unprecedented opportunity for PV resource assessments and solar energy forecasting over China [19,20]. Notwithstanding, one still demands a high-performance retrieval algorithm to estimate GHI from FY-4A measurements. Irradiance retrieval techniques can be broadly differentiated as two types, first of statistical and second of physical [21,22], which distinguish one from the other on the basis of whether radiative transfer is considered.

Indeed, physical methods consider radiative transfer from the top of the atmosphere to the Earth's surface. However, the complex radiative transfer process is usually not rigorously carried out but simplified through look-up tables, broadband surrogates, and parameterizations. In physical methods, the accuracy of cloud and aerosol microphysical properties is crucial when it comes to making highly-accurate estimations of GHI with physical methods [16,23,24]. For instance, in cloudless pixels the retrieval method would need to determine mostly just the aerosol and water vapor contents, whereas for the cloudy pixels the cloud-top height, cloud phase, and cloud optical depth would also need to be determined. The success of all this would necessarily rely on the quality of aerosol and cloud-property products, which, even produced with the most mature techniques, are still subject to considerable uncertainty due to the complexity of real atmospheric conditions. Furthermore, the aerosol and cloud-property products from FY-4A are still at a preliminary stage, which in turn prevents one from opting the physical retrieval method based on FY-4A data. Statistical methods, also known as empirical methods, retrieve GHI by establishing some form of correspondence between satellite measurements and ground measured GHI. Clearly, the problem at hand is one of regression. Recently, machine-learning regression methods have been popularly considered, as to improve the traditional regressions, due to their ability to explore the hidden and often confounding relationships among the predictor and target variables. Several popular machine-learning methods, such as multi-layer perceptron (MLP), random forest (RF), or gradient boosting decision tree (GBDT), have already had their appearances in estimating GHI from satellite data and showed promising accuracies [25–27]. It is well known that abundance of training data is the most vital prior requirement for establishing a robust machine-learning-based GHI retrieval algorithm. However, most previous studies either used measurements at very limited stations (e.g. Ref. [25]) or daily GHI measurements (e.g. Ref. [28]), which completely limit the general validity and applicability of the products—these models can hardly be generalized for retrieval at unseen locations or at high temporal resolutions. In

this work, to improve both aspects, the hourly GHI measurements at 166 stations across China from CMA and CERN are used to establish a robust machine-learning-based GHI retrieval algorithm, which, to our best knowledge, has never been attempted before.

In view of the aforementioned “double-carbon objective” of China, the overarching aim of this work is to retrieve high-resolution GHI from FY-4A geostationary satellite imagery using a machine-learning method (namely, random forest), which is then used as input to produce the first-ever FY-4A-based solar PV resource map of China. The innovation of the latter part of this work is consisted in the fact that a refined solar radiation modeling framework called the *model chain* is used for irradiance-to-PV-resource conversion. Model chain uses several energy meteorology models in cascade, in that, the output of a preceding model is used as input for the succeeding one. Though model chain has often been used for solar resource assessment and forecasting purposes [29–31], it has yet been applied, at least not in its full form, to map making, and thus the merit of this work is immediately obvious. With that said, the remaining pages of this work are devised as follows. Section 2 introduces the data and method utilized to estimate GHI and solar PV resources. Section 3 evaluates GHI products and presents the aforementioned solar resource map. Section 4 presents the conclusion and states possible future work to further improve GHI products of China.

2. Data and method

2.1. FY-4A AGRI data

FY-4A, launched on December 11, 2016, has capabilities that are greatly enhanced, as compared to its predecessors, in terms of environmental and weather monitoring, warning, and forecasting. The AGRI aboard FY-4A has 14 spectral bands from 0.47 to 13.5 μm (6 visible/near infrared, two midwave infrared, and four longwave infrared bands) [19]. In this work, the level-1 data from FY-4A in 2018 including measurements of 14 spectral bands and the associated geographical information are used. The level-1 data are available online (<http://satellite.nsdc.org.cn/>). The original dataset displayed as the geostationary projection is re-sampled onto an equal latitude-longitude grids, with a 4-km-15-min spatio-temporal resolution.

2.2. Hourly surface solar irradiance data

The ground-based measurements of hourly GHI over China during

2018 are obtained from various radiometry stations of the CMA and CERN networks (shown in Fig. 2). Thermoelectric pyranometers, with a spectral response over 0.3–3.0 μm , are used for both networks [5,18]. CM-11 pyranometers with an uncertainty of $\pm 3\%$ (Kipp & Zonen, Netherlands) and DFY-4 pyranometers with an uncertainty of $\pm 5\%$ (HY Orient, China) are used for the CMA and CERN networks, respectively. The calibration process is traceable to the World Radiometric Reference. A two-stage quality control of GHI is performed: (i) GHI values higher than the extraterrestrial GHI are rejected, and (ii) GHI values smaller than the expected minimum values (10 W/m^2) in overcast conditions are rejected [5,32]. (Although more advanced quality control routines are available, but those are generally based on all three irradiance components, and thus are not applicable to the current data, since only GHI is available.) Data points that do not meet these criteria are rejected. The elevations of sites range from 2.9 to 4507 m.a.s.l. (meters above sea level) and the land cover types include forests, desert, semi-desert, wetlands, croplands, grasslands, and urban. Data from 166 stations are randomly split into a training set and a testing set: Hourly GHI from 106 stations are used for the model training (marked as yellow squares in Fig. 1), while GHI from the remaining 60 stations as for the model testing (marked as red triangles in Fig. 1). It should be noted that both the training and testing sets cover all land-cover types, as to ensure an all-inclusive fair validation.

2.3. Random forest model

Random forest (RF) belongs to a broader class of ensemble-learning algorithms based upon decision trees (DTs), which have been found more accurate than neural networks in many applications. Especially on tabular-style datasets, RF often outperforms standard deep learning models [33,34]. One more advantage of RF lies in its model interpretability, in that, the effective methods exist in explaining its predictions from local explanations to global understanding [34].

With respect to the current application, RF modeling consists of the following steps:

- (1) The model's inputs are ground-based measurements from the stations, FY-4A spectral retrievals at the nearest grid cell to the

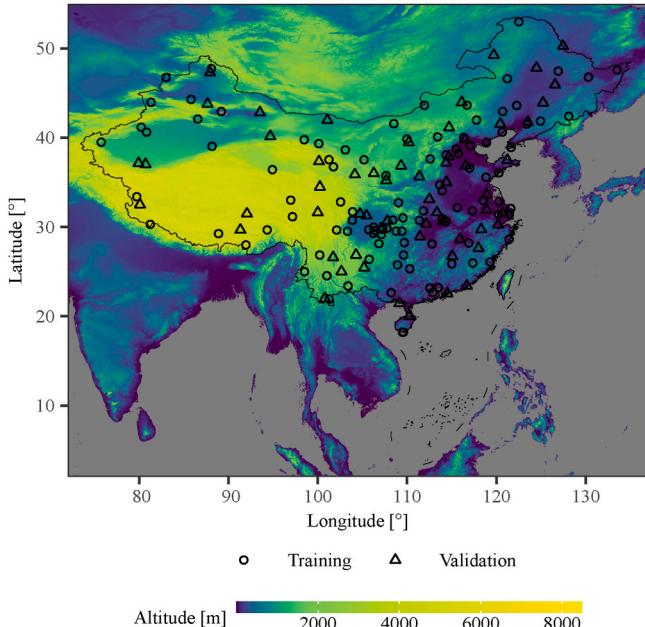


Fig. 2. Spatial distribution of pyranometer stations used in this study. Stations represented by triangles are randomly selected for the independent validation, while other stations (circle) are used for the model training.

stations, and other auxiliary data, such as geographic location, elevation, or Sun-Earth distance, with a 1-h temporal resolution. Owing to the moderately large number of features, the Pearson correlation coefficient and the variance inflation factor (VIF) between the parameters are used to inquire the presence of multicollinearity [35,36]. Figure A1 depicts the Pearson correlation coefficients among the predictors of the RF model in question, in the form of a heatmap. Table A1 lists the VIFs of all variables, before and after excluding specific variables. Finally, negative influence or highly correlated parameters are excluded and 11 variables are retained, including spectral bands of 0.55–0.75, 3.5–4, 6.9–7.3, and 11.5–12.5 μm ; time information of Sun-Earth distance, solar zenith, azimuth, and glint angle; location information of latitude, longitude, and elevation.

- (2) The RF model is determined by the best features among the randomly selected properties. In this study, the Python Scikit-Learn library implementation is used [37]. To optimize the model performance, hyperparameters need to be tuned. Grid search, which has satisfactory default implementations in many machine-learning software tools, is used for such task. The final hyperparameters used are as follows: the number of trees in the forest (i.e., n_estimators = 800), the minimum number of samples that is required to split a node (i.e., min_samples_split = 2), the maximum number of features to consider when splitting a node (i.e., max_features = 'sqrt'), the maximum tree depth (i.e., max_depth = 60), and an optional parameter indicating sampling data with or without replacement (i.e., bootstrap = 'True'). Then, with each tree producing an estimate, the final prediction is arrived at by computing the arithmetic mean of the individual-tree estimates.

Owing to the “black box” nature of machine learning, interpreting the output of machine-learning models has hitherto been a demanding task. On this point, a popular approach, called the Shapley Additive explanation (SHAP [34]), is often employed to explain the behavior of machine learning by calculating the contribution of each feature to the prediction. Furthermore, SHAP is computationally friendly for tree-based models that is the fundamental cause for the popularity of SHAP. This is, to our best knowledge, the first time to use SHAP to explain how input features predict GHI. The left panel of Fig. 3 displays a standard SHAP summary plot, whereas the right panel quantifies the feature importance. The plot sorts various features by the sum of SHAP magnitudes, and uses SHAP values to show the distribution of the influences of each feature on the model output. The feature value is mapped to the color bar (in a viridis palette). The higher (lower) SHAP value of a feature is, the higher (lower) the estimated GHIs of the RF model is. Indeed, the lower reflectance in the spectral band of 0.55–0.75 μm corresponds to a cloud-free situation, which explains why the higher SHAP value results in the higher estimated GHI (Fig. 3a). Meanwhile, the higher SHAP value of the elevation along with higher elevation brings the higher estimated GHI. For example, the Tibet Plateau has abundant solar energy reserve. Fig. 3b indicates that the feature with the highest importance for the present case is the solar zenith angle, followed by reflectance at 0.55–0.75 μm , the solar azimuth angle, thermal emission at 11.5–12.5 μm , and others.

- (3) After the input-output relationships are learned by RF, the effectiveness of the model is validated using unseen samples from the 60 remaining stations, which cover all climate types of China. The overall model performance is gauged through four popular measures: mean bias error (MBE), root mean square error (RMSE), relative mean bias error (rMBE), relative root means square error (rRMSE), and correlation coefficient (R) between estimated and ground measured values [18].

2.4. Clear and cloudy sky identification

To better understand the influencing factors of GHI, the collocated

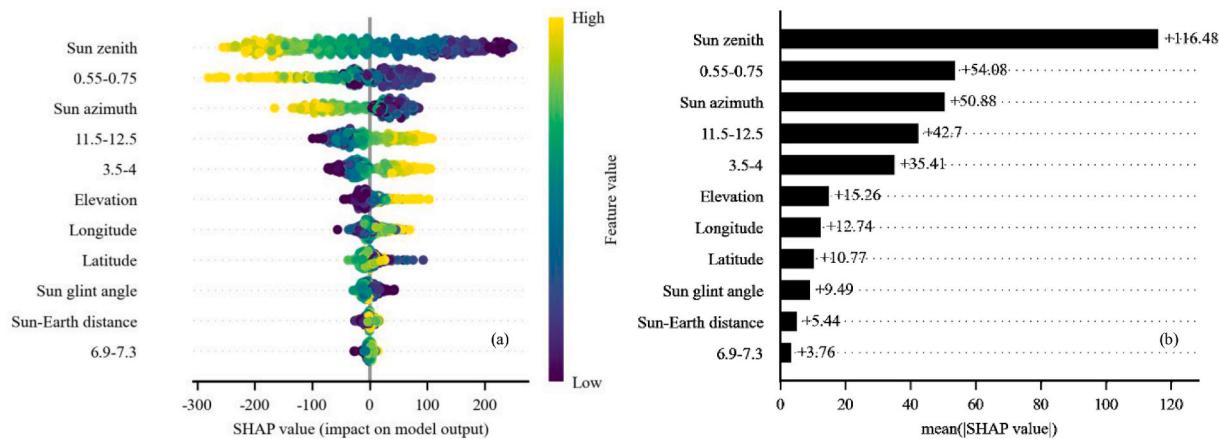


Fig. 3. (a) SHAP summary plot of the Random Forest model. 11 features (spectral bands of 0.55–0.75, 3.5–4, 6.9–7.3, and 11.5–12.5 μm ; time information of Sun–Earth distance, solar zenith, azimuth, and glint angle; location information of latitude, longitude, and elevation) are represented by a vertical line, and the horizontal position of the dots (SHAP value) shows the degree of impact the feature has on the model output. The colors of the dots indicate the values of the corresponding feature. (b) The importance of features, calculated by averaging the absolute value of the SHAP values.

data points are split into two mutually exclusive groups, first of clear-sky samples and second of cloudy-sky samples. The splitting is based on several k indexes. More specifically, an hourly GHI value is deemed to be under a clear sky if it meets four criteria. First, the clearness index k_t (see Eq. (1)), which is the ratio of GHI (G_h) to extraterrestrial GHI (E_0), does not exceed 0.82. Second, the modified clearness index k'_t (see Eq. (2)), as first appeared in Ref. [38], is within 0.65 and 1.0. Third, the variability of the k'_t within 3 h, or $\Delta k'_t$ (see Eq. (3)), is smaller than 0.05. Fourth, the clear-sky index κ (see Eq. (4)), which is the ratio between GHI to its clear-sky expectation (G_{hc}), exceeds 0.65. The clear-sky expected GHI is calculated using the REST2v5 model. The thresholds used in this study are chosen based on the combination of experience and multiple scenario testing. Whereas the detail descriptions of the clear-sky detection method and its validation can be found in Refs. [4,11,18], the aforementioned k indexes are mathematically summarized as follows:

$$k_t = G_h / E_0 \quad (1)$$

$$k'_t = k_t / \{1.031 \times \exp[-1.4 / (0.9 + 9.4 \cos Z)] + 0.1\} \quad (2)$$

$$\Delta k'_t = k'_t(i) - [k'_t(i-1) + k'_t(i) + k'_t(i+1)] / 3 \quad (3)$$

$$\kappa = G_h / G_{hc} \quad (4)$$

where Z is the solar zenith angle.

2.5. PV resource estimation

To estimate PV resource over China, the pvlib package, which is an open-source PV modeling and forecasting toolbox developed by the Sandia National Laboratories, is used [39]. More specifically, the pvlib-python package (<https://pvlib-python.readthedocs.io/en/stable/>) provides routines for simulating the performance of PV systems and related irradiance and atmospheric models; it has been found useful in a range of solar applications by researchers worldwide (e.g. Refs. [30, 40]). This study uses the package for calculating the effective irradiance incident on a PV panel placed at the rule-of-thumb orientation for a location of interest, using the FY-4A satellite-derived irradiance. The irradiance incident on PV arrays, also known as the plane-of-array (POA) irradiance or global tilted irradiance (GTI), consists of three components: beam irradiance on the tilted surface (BTI), diffused irradiance on the tilted surface (DTI), and irradiance due to ground reflections. This work assumes the use of fixed-tilt PV arrays posited at a south-facing azimuth angle and a tilt angle equal to the site's latitude.

The detailed estimation workflow as to arrive at POA irradiance is shown in Fig. 4. Fig. 4 presents a collection of component models that are used in cascade, as such, it is more widely known as a model chain.

2.5.1. Separation modeling

When the beam and diffuse irradiance components are unknown, they can be estimated from GHI, which is what separation models do. Hundreds of separation models have been proposed over the past half century. For instance, the Direct Insolation Simulation Code (DISC) model, which is choice used to produce the National Solar Radiation Data Base (NSRDB [16]), seeks a quasi-physical relationship between the beam normal clearness index and the global clearness index [41], and once the beam normal clearness index is derived, beam normal irradiance (BNI) follows. Since NSRDB has proven to be a reliable satellite-derived irradiance database, this work inherits its choice, and uses DISC as the separation model.

2.5.2. Transposition modeling

The aim of transposition model is to convert the three horizontal irradiance components into GTI. Mathematically, the general transposition equation is given as:

$$G_c = B_n \cos \theta + R_d D_h + \rho R_r G_h, \quad (5)$$

where, G_c and B_n denote GTI and BNI, respectively; D_h is the diffuse horizontal irradiance (DHI); θ is the incident angle which can be calculated via solar positioning; ρ is the foreground's albedo, which is assumed to be 0.2 here; R_d is the diffuse transposition factor, which is the only unknown in Eq. (5); R_r is the transposition factor due to the reflection of the ground, and usually, assuming a Lambertian surface (i.e., isotropy), $R_r = (1 - \cos Z) / 2$; and Z denotes the solar zenith angle.

The diffuse radiation on the surface can be obtained by the Perez model [42]:

$$R_d = (1 - F_1) \frac{1 + \cos S}{2} + F_1 \frac{a}{c} + F_2 \sin S, \quad (6)$$

where,

$$a = \max(0^\circ, \cos \theta), \quad (7)$$

$$c = \max(\cos 85^\circ, \cos Z), \quad (8)$$

in which F_1 and F_2 are sky-dependent model coefficients, and S is tilt of the inclined surface.

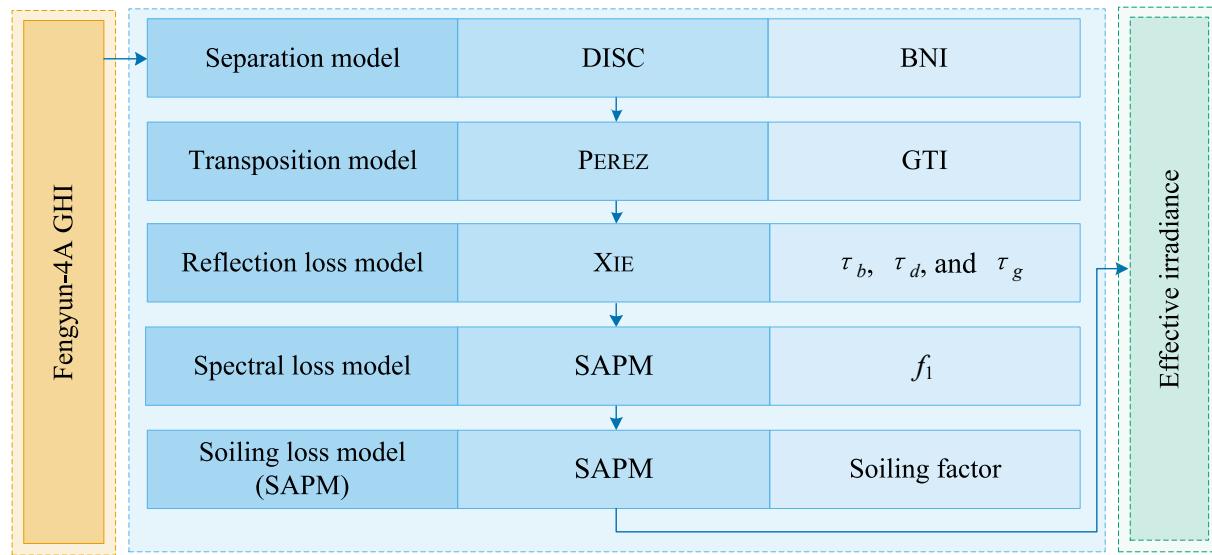


Fig. 4. Schematic of a model chain to obtain effective irradiance.

2.5.3. Reflection loss model

The nominal power rating of a PV panel is determined under the so-called standard test condition (STC), which refers to an incident irradiance of 1000 W/m^2 , a cell temperature of 25°C , an air mass 1.5 solar spectra, as well as ensuring a uniform normal incidence of irradiance. However, in field conditions, the variation of the incidence angles can result in higher reflection losses than that accounted for during STC testing. As mentioned earlier, the GTI has three radiation components, namely, BTI B_c , DTI D_c , and the ground-reflected radiation D_g , i.e., $G_c = B_c + D_c + D_g$, in that, the reflection losses under field conditions are to be accounted for each of these components, which gives rise to the absorbed irradiance, or G'_c :

$$G'_c = \tau_b B_c + \tau_d D_c + \tau_g D_g, \quad (9)$$

where τ_b , τ_d , and τ_g are the relative transmittances for beam irradiance, diffuse irradiance, and the ground-reflected irradiance, respectively. To calculate τ_b , one should use the Fresnel equation, which is the physical law according to which a beam of light is transmitted through two different media:

$$\tau_b = \frac{1 - \frac{1}{2} \left(\frac{\sin^2(\theta' - \theta)}{\sin^2(\theta + \theta)} + \frac{\tan^2(\theta' - \theta)}{\tan^2(\theta + \theta)} \right)}{1 - \left(\frac{n_{\text{PV}} - 1}{n_{\text{PV}} + 1} \right)^2}, \quad (10)$$

where θ' is the refractive angle obtainable via the well-known Snell's law, and n_{PV} is PV module's refraction index, which normally resumes a value of 1.526 [43].

On the other hand, τ_d and τ_g are far more difficult to compute. More particularly, for the past century or so, nor was there any analytical expression for them, owing to the difficulty in integration. In this regard, researchers often either use approximations [44] or perform numerical integration [45]. However, very recently, Xie and co-workers eventually found a work-around in deriving the analytical expressions of τ_d and τ_g , which have been given in Ref. [46]. Hence, to acknowledge this field-advancing formulation, the XIE reflection loss model is herein used. The relative transmittance for the diffuse radiation is:

$$\begin{aligned} \tau_d = & \frac{2w}{\pi(1 + \cos S)} \left(\frac{30}{7}\pi - \frac{160}{21}S - \frac{10}{3}\pi \cos S + \frac{160}{21}\cos S \sin S - \frac{5}{3}\pi \cos S \sin^2 S \right. \\ & \left. + \frac{20}{7}\cos S \sin^3 S - \frac{5}{16}\pi \cos S \sin^4 S + \frac{16}{105}\cos S \sin^5 S \right), \end{aligned} \quad (11)$$

where

$$\begin{aligned} w = & \frac{n_{\text{PV}}(n_T + 1)^2}{n_T(n_{\text{PV}} + 1)^2} (2.77526 \times 10^{-9} + 3.74953n_{\text{PV}} - 5.18727n_{\text{PV}}^2 \\ & + 3.41186n_{\text{PV}}^3 - 1.08794n_{\text{PV}}^4 + 0.13606n_{\text{PV}}^5). \end{aligned} \quad (12)$$

In Eq. (12), n_T is the refraction index of the pyranometer cover, and usually $n_T = 1.4585$ [46]. The relative transmittance for the ground-reflected radiance is:

$$\tau_g = \frac{40w}{21(1 - \cos S)} - \frac{1 + \cos S}{1 - \cos S}. \quad (13)$$

It should be noted that, unlike DISC and PEREZ models, the XIE model is new, and there is no standard implementation in pvlib.

2.5.4. Spectral loss model

On the one hand, the term "shortwave radiation" refers to radiation in the spectral range of 300–4000 nm. On the other hand, the spectral response of solar cells, depending on the cell technology, differs from that range, e.g., crystalline silicon has a spectral response of 300–1200 nm. What this implies is that the amount of modeled GTI must be discounted when it is used in solar resource mapping. The amount of spectral loss can be sufficiently modeled as a function of two things, first of air mass (AM_a) and second of module type. Spectral loss is modeled as a fourth-order polynomial function of AM_a [47], where the model coefficients are determined based on each module. For instance, Canadian Solar CS5P-220 M has coefficients: $a_0 = 0.928385$, $a_1 = 0.068093$, $a_2 = -0.015774$, $a_3 = 0.001661$, and $a_4 = -0.000069$. Then, the SAPM spectral loss coefficient (f_1) is:

$$f_1(AM_a) = a_0 + a_1 \times AM_a + a_2 \times AM_a^2 + a_3 \times AM_a^3 + a_4 \times AM_a^4. \quad (14)$$

2.5.5. Soiling loss model

Outdoor operation of PV is susceptible to the effect of soiling, which refers to the accumulation of dust on the panel surface. There has been a substantial amount of research conducted for soiling. However, due to the differing experimental setups, the results obtained from one study can rarely be applied to another. The amount of soiling depends primarily on three factors: (1) the accumulation of dust, which further depends on the duration of exposure (i.e., the time between successive rain events), as well as the geographical location at which the panels are installed; (2) the dust-removal effectiveness of rain events, which further depends on the amount of rain in unit time; and (3) the active cleaning

schedule. Clearly then, it would be quite difficult to come out a sensible soiling estimate in the form of gridded data. To that end, this work has no choice but to invoke some assumptions, as to account for soiling in the estimation of PV resource maps. In particular, two assumptions are made. First it is assumed that soiling increases linearly with time, which is a fairly common practice in the literature (e.g. Refs. [48,49]). The second assumption that is that the panels are cleaned after a threshold amount of soiling is reached. In recent years, the conceptions of low-cost soiling sensors became popular [50,51], and in principle, these sensors are able to notify the PV plant owners once the soiling exceeds some predefined threshold. Without loss of generality, the threshold is set to be 3%. This, alongside the linearity assumption, implies an average soiling loss of 1.5% over each cleaning cycle. Since active cleaning takes place whenever soiling exceeds the threshold, the amount of natural cleaning by rain can be relaxed during modeling. One should note however that, depending on the location and rain amount, the cost of active cleaning differs from one grid cell to another; the economic factors are nonetheless beyond the scope of this work.

3. Results

3.1. Model evaluation against ground measurements

The GHI estimates from the RF-based satellite-to-irradiance retrieval algorithm is evaluated against independent ground-based observations, namely, those data points that are not used in model training. Fig. 5 depicts the density scatterplots of hourly (solar zenith angles less than 75° excluded) and daily results for the RF model over China in 2018. The numbers of independent validation samples for hourly and daily GHI data are 195307 and 21900, respectively. For hourly GHI estimates, the values of R, RMSE (rRMSE) and MBE (rMBE) are 0.85, 147.02 (35.2%), -5.64 W/m^2 (-1.4%), respectively, showing a good agreement between the estimated and observed GHI values. The model tends to overestimate hourly GHI when ground values are low while underestimating when ground values are relatively large (Fig. 5a), which is a common trait for modeled irradiance. More particularly, if the ground station is shadowed by broken clouds or surrounding structures, the instantaneous GHI measured would be lower than that at neighboring areas within a satellite pixel, resulting in a measurement lower than the satellite-derived value, which represents just the pixel average [52]; this effect is generally known as the spatial scale mismatch [53]. As for high-irradiance situations, satellite-derived irradiance underestimates due to the lack of consideration of cloud and albedo enhancement events, which, though being instantaneous in nature, may impose some

effect in the aggregated values. Daily GHI estimates are well correlated to the surface observations with the R of 0.95, the slope of 0.90, and the intercept of 11.97 W/m^2 (Fig. 5b). The corresponding values of RMSE (rRMSE) and MBE (rMBE) are 29.20 (18.0%), -2.97 W/m^2 (-1.3%), respectively. These lower errors as compared to the case of hourly estimates can be fully anticipated, as the variability at daily time scale is far lower than that at hourly time scale, with additional possibilities of error cancellation occurring throughout the day.

The RF model based upon FY-4A measurements provides reasonably accurate GHI estimates. Especially in terms of the daily GHI estimates, the ones obtained in this study are remarkably better than those reported in previous studies that obtain satellite retrievals via machine-learning methods (Table 1). For example, the RMSE values of daily GHI are 58.1 W/m^2 from Jiang et al. (2019), 33.3 W/m^2 from Ma et al. (2020), and 35.4 W/m^2 from Hou et al. [28]. Although the RMSE value of daily GHI is slightly larger for MTSAT in 2008 from Jiang et al. [25], the hourly GHI performance behaves well due to the application of a convolutional neural network (CNN), of which the advantage can be attributed to the consideration of spatial information, such as instantaneous changeable cloud morphology. Ma et al. [26] proposed a hybrid method, combining a deep neural network (DNN) and radiative transfer model (RTM), to retrieve GHI separately under clear- and cloudy-sky conditions. Hou et al. [28] leveraged just a four-month sample for training, which may present as too short, and thus could not capture the any seasonal variation in various statistical properties in regard to irradiance. The datasets with a limited time range trained by machine learning tend not to be able to fully represent the underlying physical processes, such as radiative transfer [25]. An obvious remedy is to expand the dataset by incorporating more ground measurements into the training process, which could often improve the performance of the network in terms of stability and robustness. Besides, to estimate GHI across China from satellite data, to date, most previous studies are based on Japanese geostationary satellites [17,25,26], which should face a shortage of reliable data over western regions of China. For example, the coverage of Himawari8 reaches only up to 80°E longitudes, let alone the distortion of the images near the field-of-view edge.

Next, we attempt to identify the driving factors of the observed bias in the hourly estimates by validating them separately under the clear-and cloudy-sky conditions. Under clear-sky conditions (Fig. 6a), the model underestimates hourly GHI with the larger negative rMBE of -5.6% but smaller rRMSE of 129.35 W/m^2 (23.8%) than those on all-sky conditions. This underestimation may originate from the parallax effect, in that, clouds may exist in the surface-to-satellite path, but not along the sun-to-surface path. This issue occurs commonly during the

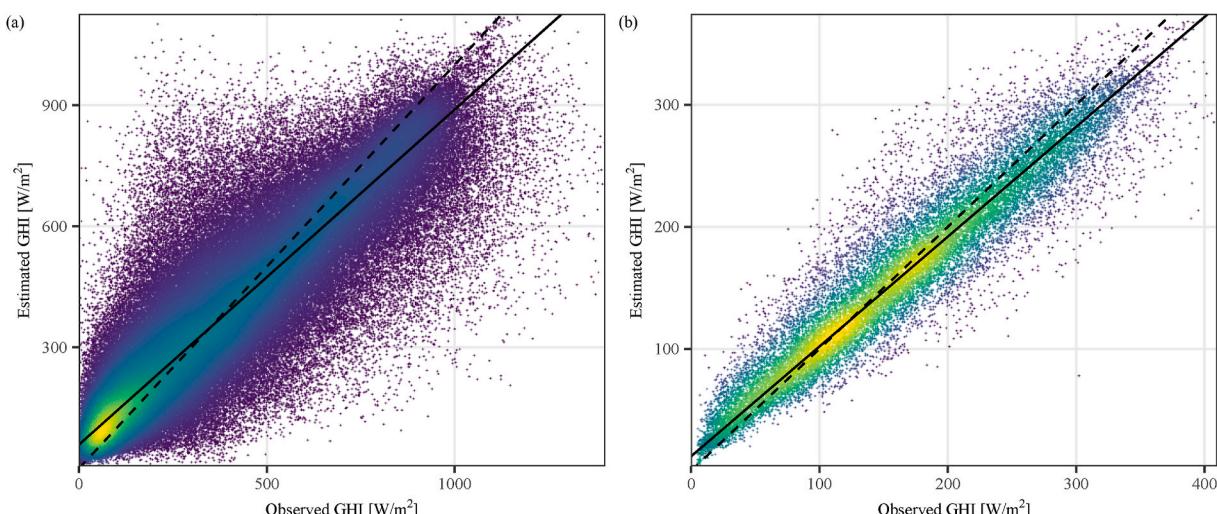
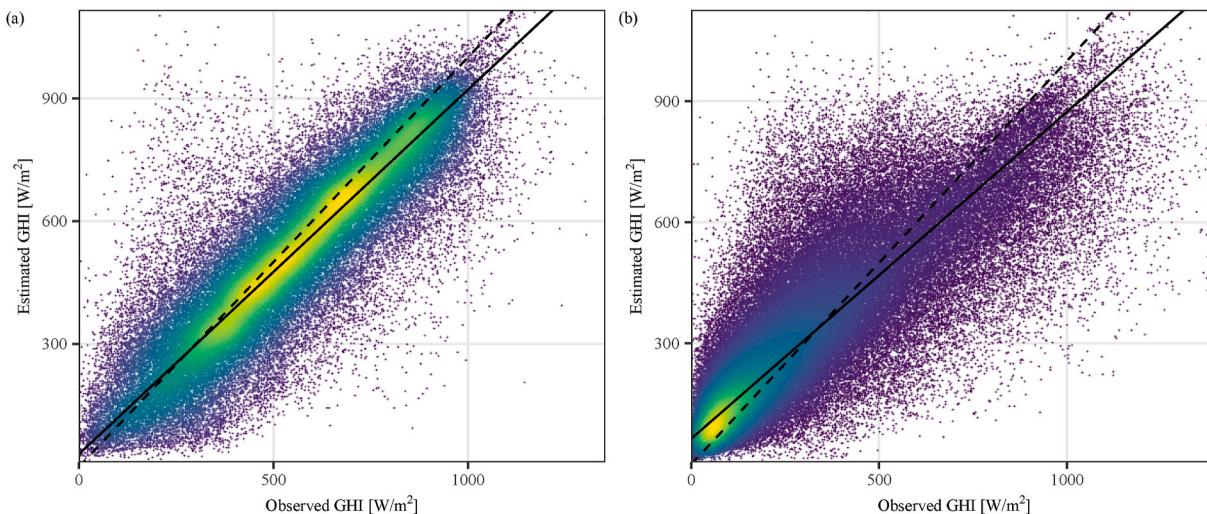


Fig. 5. Density scatterplot of hourly (a) and daily (b) GHI between FY-4A trained by RF model and validated ground measurements in 2018.

Table 1

Comparison of the estimation statistics based on satellite data across China from various references.

Datasets	Methods	Variables	Hourly Performance [W/m ²]	Daily Performance [W/m ²]	References
MTSAT (2008)	CNN + MLP	0.55–0.80 μm, Local time, Latitude, longitude, elevation	RMSE = 90.38 MBE = 5.23 R = 0.87	RMSE = 58.1 MBE = 5.0 R = 0.91	Jiang et al. (2019)
Himawari8 (2016)	Radiative Transfer, DNN	AOD, CP, COT, CER, Albedo, WV, SZA	RMSE = 106.4 MBE = 28.6 R = 0.87	RMSE = 33.3 MBE = 15.1 R = 0.89	Ma et al. (2020)
Himawari8 (Feb to May 2016)	Random Forest	16 bands, WV, Elevation, SZA, length of daytime	–	RMSE = 35.38 MBE = 0.01 R = 0.92	Hou et al. (2020)
FY-4A (2018)	Random Forest	Details in section 2.3	RMSE = 147.02 MBE = -5.64 R = 0.85	RMSE = 29.20 MBE = -2.1 R = 0.95	this study

**Fig. 6.** Density scatterplot of hourly GHI between FY-4A trained by RF model and validated ground measurements in 2018 under clear-sky (a) and cloudy (b) conditions.

validation of instantaneous satellite-based retrievals, and has been the subject of concern of several previous studies [54–56]. In contrast, the hourly GHI by RF model is overestimated with the rMBE of 2.8% under cloudy-sky conditions (Fig. 6b). This can also be explained as the aforementioned spatial scale mismatch, where ground-based observations contain more granular cloud features that are not resolved in the satellite pixels. Impacts from representativeness errors are evidently reduced at longer time scales, owing to the potential cancellation of over- and under-predictions [54]. In any case, there is still a considerable discordance between point observations and satellite products for the cloudy-sky case. In this regard, the quality of irradiance retrieval still demands further refinement, but in practice, it can be quite challenging owing to the view angle of satellite and the ubiquitous cloud inhomogeneity, which make using the FY-4A all the more important for China.

3.2. Spatial and temporal validation

Fig. 7 exemplifies the diurnal variation of hourly GHI estimated by the RF model using FY-4A on July 1, 2018. Geography suggests the sun rises in the northeast while sets in the northwest during summer. Consequently, GHI tends to form contours along the northwest–southeast direction in the morning, and along the northeast–southwest direction in the afternoon. The intensity of GHI grows gradually with decreasing solar zenith angle, and reaches the peak at local solar noon. The advection of clouds throughout this day is clearly visible, which alters accordingly the spatial distribution of GHI. The spatial distribution of

low GHI is located along the Yangtze River and has a southwest to northeast orientation, which collocates with the Meiyu frontal zone in summer China [57]. Conversely, the GHI in North China Plain is relatively high under clear-sky conditions.

Fig. 8 presents the validation of hourly estimated GHI in 2018 conditioning on solar zenith angles, which are divided into 10 bins (the first bin takes $Z < 30^\circ$ due to fewer samples and the remaining ones are divided uniformly with an interval of 5°). As the solar zenith angle increases, the rRMSE values increase from $31.3\% \pm 9.8\%$ – $54.0\% \pm 22.3\%$. The rMBE values increase from $-0.9\% \pm 12.9\%$ – $9.6\% \pm 17.8\%$ and the R declines from $78.2\% \pm 9.4\%$ – $63.2\% \pm 27.8\%$. Whereas the rRMSE and rMBE values ramp up sharply as the solar zenith angle goes beyond 60° , the R values drop rapidly. In general, the estimated hourly GHI performs well at small solar zenith angles (i.e., near solar noon), but one must be aware of the caveats, due to strongly increasing uncertainty both in measurements and radiative transfer simulations, in using the estimated GHI at the large solar zenith angles.

Fig. 9 overlays the statistical performance indicators with the map of China, revealing the spatial pattern of the prediction errors. More than 65% of the stations have low estimation errors with RMSE values < 150 W/m², respectively. Besides large RMSE values are seen in the west of Sichuan Basin, central, and northeastern China (see Fig. 9a and b). MBE shows a similar spatial distribution to the RMSE, so its interpretation is omitted. More than 80% of the stations show high R values > 0.8 (Fig. 9c). In contrast, R is relatively low at a few stations in the southwest and south of China. The distributions of clear-sky day numbers at the ground stations show that the clear-sky days over the southern,

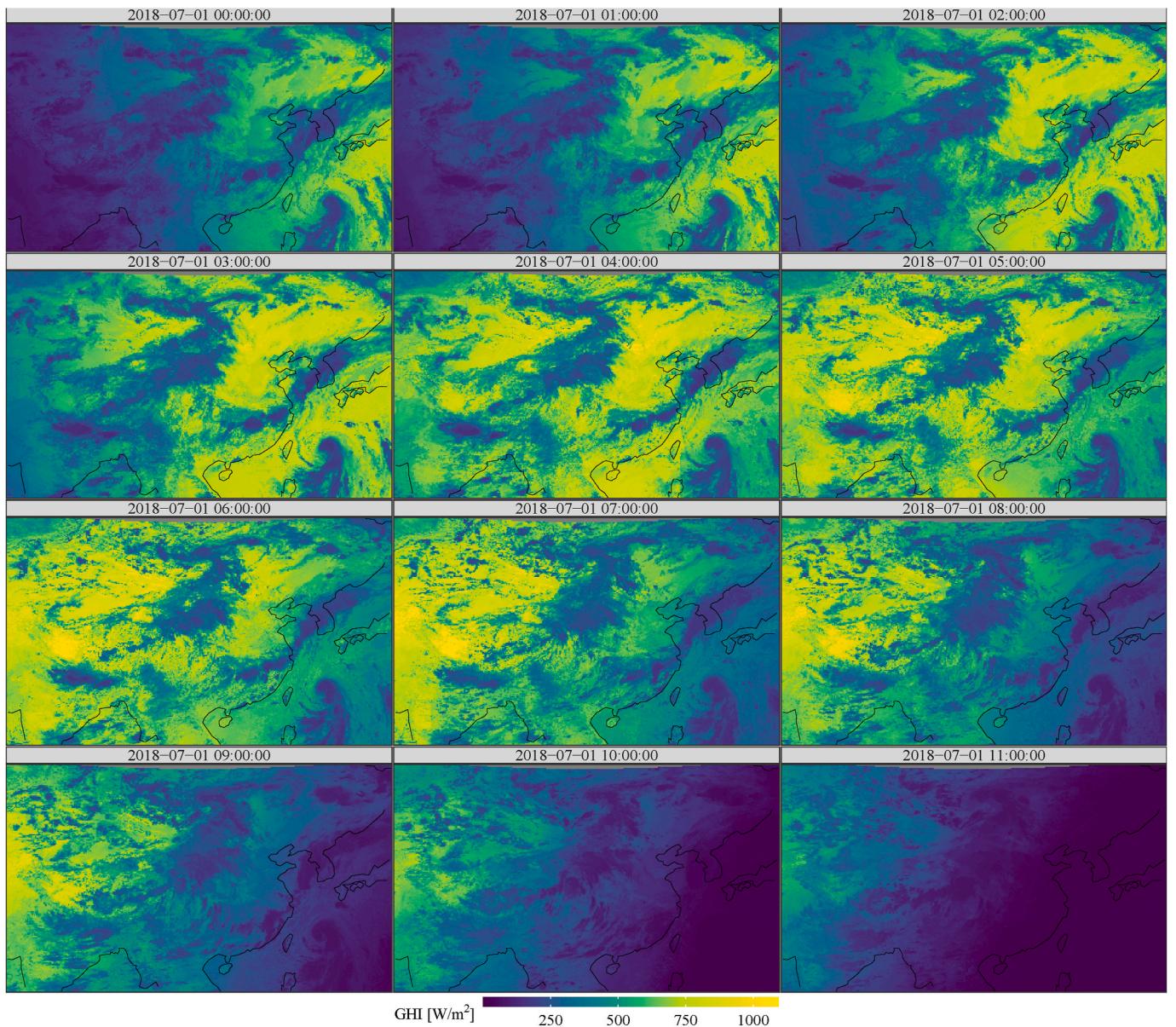


Fig. 7. The diurnal variation of the hourly GHI estimated by RF model for FY-4A from 0000 UTC to 1100 UTC on July 1, 2018, over China.

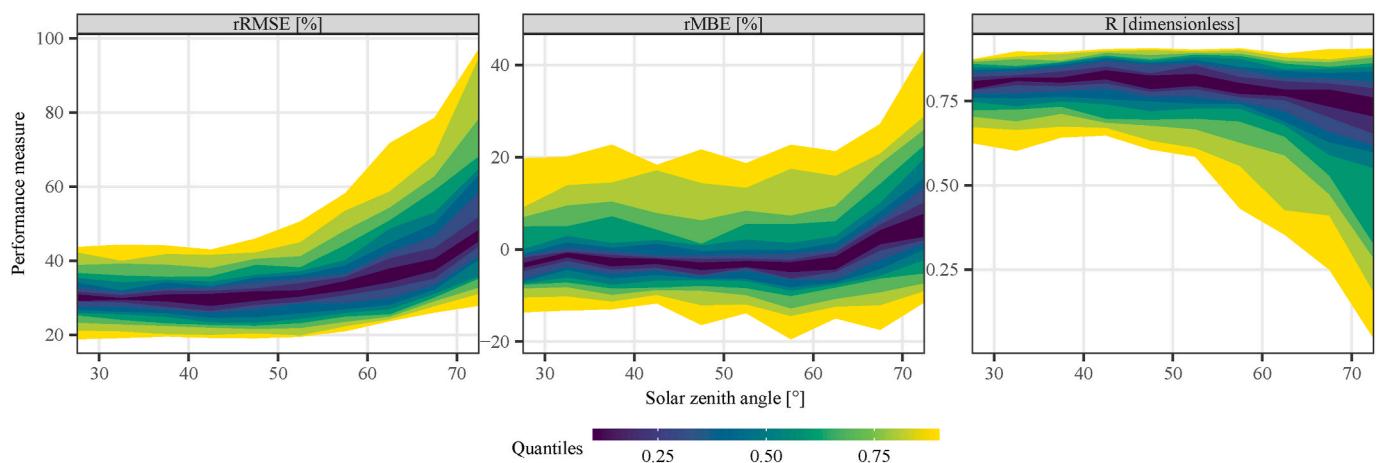


Fig. 8. The quantile performance of (a) relative root means square error, (b) relative mean bias error, and (c) correlation coefficient between hourly estimated GHI and ground observations as a function of the solar zenith angle in 2018 over 60 validation sites.

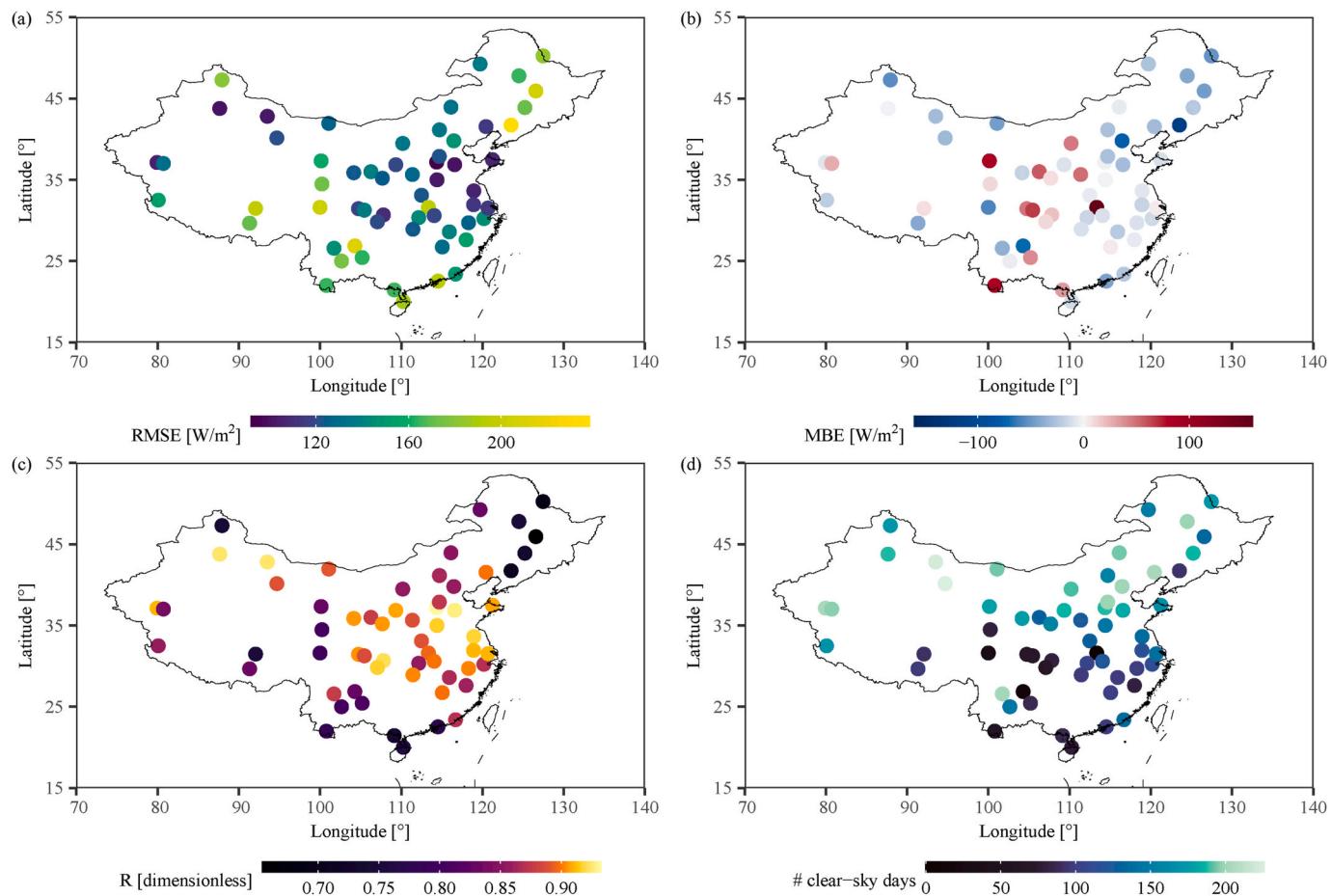


Fig. 9. Spatial distribution of the performance of hourly estimated GHI compared to the ground measurements in 2018 for (a) RMSE, (b) MBE, and (c) R. Clear-sky day numbers of the valid stations (d) is shown here.

southwestern, and central China are fewer than 50 days in 2018, as depicted in Fig. 9d, the frequently occurring clouds result in the larger estimation errors at those locations. In addition, the larger estimation error in central China could be due to the relatively fewer training samples, as shown in Fig. A2. The underestimation which results in large RMSE in northeastern China may be attributed to the snow cover and bright surfaces [16,55].

The spatial distributions of annual GHI trained by the RF model with a resolution of $4 \text{ km} \times 4 \text{ km}$ are contrasted with the GHI from the CERES on $1^\circ \times 1^\circ$ grid and MERRA-2 on $0.5^\circ \times 0.625^\circ$ in Fig. 10. The three maps resemble each other highly in spatial patterns, but the 4-km GHI map from FY-4A evidently exposes more refined spatial features. The areas with rich solar resources, in terms of annual mean GHI, are mainly found in northwestern China, Tibetan Plateau, and Inner Mongolia Plateau. Although MERRA-2 assimilates space-observed aerosols and clouds that are expected to better reflect the GHI due to between aerosols and clouds physical processes, the North-China Plain still shows overestimated GHI. The low value in the Sichuan Basin is found in the CERES products because of the reduction of the cloud cover. But, the result of the RF model reveals that the annual GHI in the Sichuan Basin is overestimated likely owing to the relatively limited number of valid training samples (discussed in Fig. 9).

3.3. Solar photovoltaic resources map

As the final goal of this work, creating a high spatio-temporal resolution map of effective irradiance is essential for the pathway towards a low-carbon future. Fig. 11 shows the annual mean effective irradiance over China, assuming south-facing PV arrays tilted at the sites'

respective latitudes. The map reveals abundant resource over northern China, which is most suitable for deployment of PV. For most of the west, although the annual effective irradiance is more than 1700 kWh/m^2 per annum, with the largest value in Tibet Plateau (around 2000 kWh/m^2 per annum), the area is not suitable for PV due to low population density and geographical challenges in logistics (e.g., project development or energy transportation). Eastern China has effective irradiance of only $1300\text{--}1500 \text{ kWh/m}^2$ per annum, which barely meets the recommended insolation level for solar project development by the China Meteorological Administration (CMA [58]), that is, an average GHI larger than 160 W/m^2 (or 1400 kWh/m^2 per annum). The poorest solar resource region locates in the Sichuan Basin, which has around 40% less insolation than western China.

It should be highlighted that the spatial distribution of solar resources agrees, in the main, with previous studies [59], but has higher estimated values, because the effective irradiance for fixed optimally oriented panels, rather than on a horizontal surface, is evaluated here. This pattern of high solar resources in the west and north, while low in the east and south, is largely driven by surface elevation and cloud cover. Besides, as just mentioned, the geographical and social factors, such as land-use type, elevation, population density, or energy consumption pattern (see Ref. [60] for a list), are also key to the practical PV design and deployment [61]. To truly develop “the one map” for solar resource assessment, further constraints are to be considered, such as local irradiance predictability (forecast accuracy and penalty), hardware strategy (e.g., oversizing and curtailment), and energy economics. These are not considered in this paper, but should be the subject for future works.

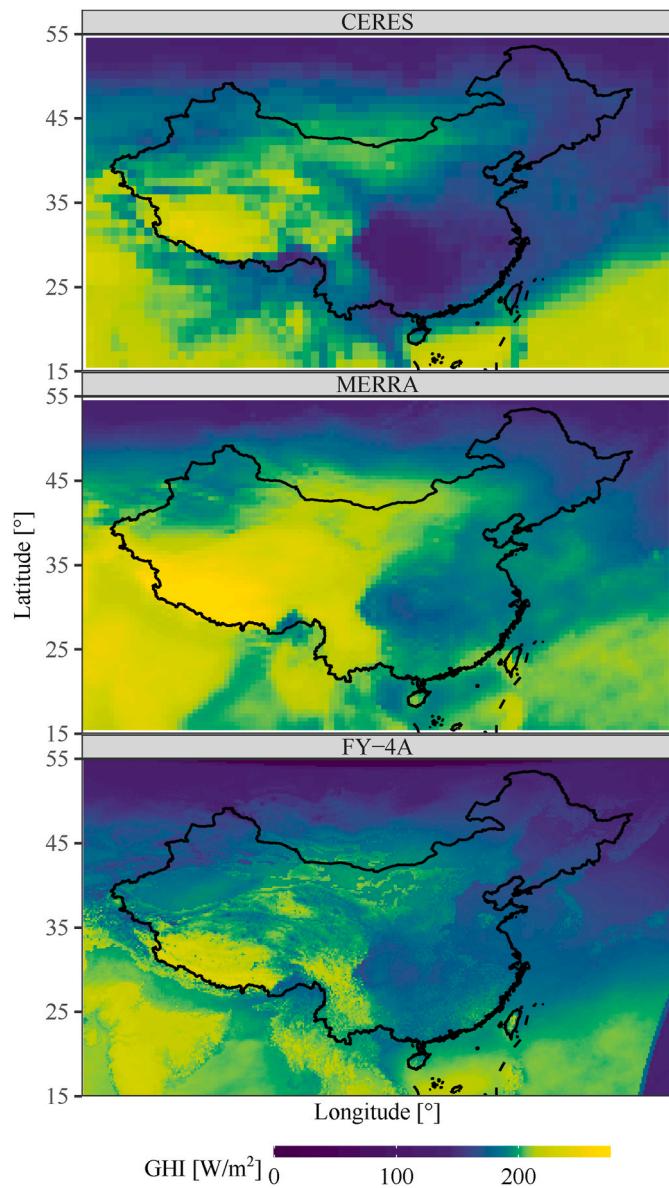


Fig. 10. Spatial distribution of annual averaged GHI from (a) CERES, (b) MERRA-2, and (c) FY-4A trained by RF model in 2018.

4. Conclusion

A satellite-derived irradiance product over China, with a high spatio-temporal resolution of 15 min and 4 km is presented, based on data from FY-4A geostationary satellite and a random forest model. The training and validation of the product use 166 ground-based radiometry stations, which is by far the most comprehensive for works of this sort conducted for China. Owing to the field-of-view of FY-4A satellite, the current product has better reliability over China than those based on Himawari or Meteosat, of which the accuracy drops substantially towards the edge of the satellite disk. More importantly, the investigation does not stop at GHI, like most other works do, but further extends to effective irradiance, which is absolutely vital for solar resource assessment for PV. To do so, a highly attractive method called the physical model chain, which leverages a sequence of energy meteorology models in cascade to arrive at a precise estimate of in-plane irradiance, is used for the first time. The novelty of this work is thus immediately evident.

The proposed FY-4A-based product is validated at both hourly and daily scales. Generally, it provides considerable accurate GHI estimates.

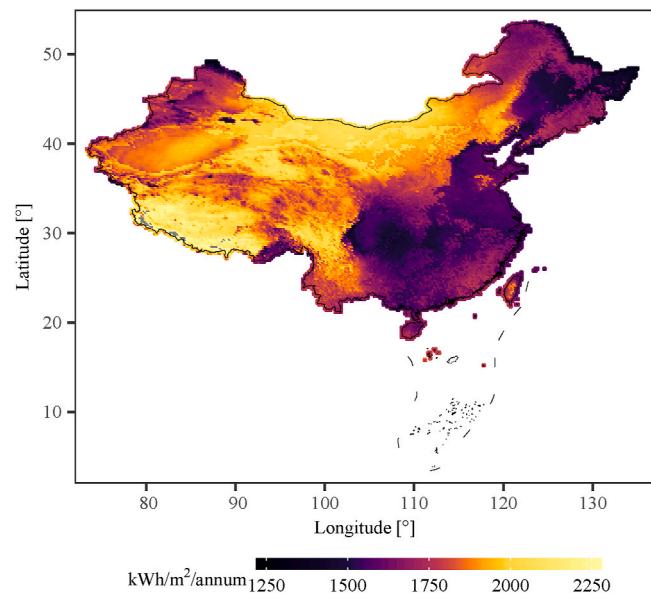


Fig. 11. Annual mean effective irradiance ($\text{kWh}/\text{m}^2/\text{annum}$) over China assuming the PV array with south facing and latitude tilt.

The daily estimates provide more reasonably accurate than the hourly results. The values of R, RMSE (rRMSE) and MBE (rMBE) for the hourly GHI estimates are 0.85, 147.02 (35.2%), $-5.64 \text{ W}/\text{m}^2$ (-1.4%), respectively. The corresponding values for the daily GHI estimates are 0.95, 29.20 (18.0%), $-2.97 \text{ W}/\text{m}^2$ (-1.3%). Besides, the estimated hourly GHI performs well at small solar zenith angles smaller than 60°, for large zenith angles, the error is high but should not raise concern, as those low-sun instances are only of remote relevance to solar energy applications. There are larger estimation errors over the Sichuan Basin, central and northeastern China, which are due either to high occurrences of clouds, or to snow and bright surfaces. The spatial distribution of the 4-km GHI estimates from FY-4A trained by RF model agrees well with CERES, but the former is able to expose more detailed features as compared to the latter. The annual mean map of effective irradiance over China features abundant insulation over northern China, which is most amenable for PV development. In contrast, the poorest solar resource region locates in the Sichuan Basin.

Despite the several aforementioned advantages to the GHI product, the current irradiance-retrieval algorithm still has room for improvement. For instance, the chosen inputs of FY-4A channels for the RF model are limited, considering the portability of the historical FY-2 series satellites. Furthermore, those aerosol-related input variables should also be taken into account, as aerosols present as the most influential factor other than clouds in absorbing and scattering incoming radiation. Besides, the limitation of hourly estimates can be attributed to the spatial factors of small-scale clouds or spatial representativeness. Thus, taking spatial features into model learning, such as using a convolutional neural network, might be able to eliminate the effect of inhomogeneous small broken clouds on the discordance between point observations and satellite products. The last drawback of the current work is that the product covers just a year, which does not allow one to study the inter-annual variability of solar resources. However, as the method is general, it is possible to expand the product once the multi-year FY-4A L1 data is acquired.

In any case, this work can offer valuable information to those involved in the design, planning, and operation of solar energy systems. Additionally, high-resolution solar resource maps can be used by researchers and academics to understand the solar energy potential of different regions and study the impact of various factors such as climate change on solar energy production [62,63]. Last but not least, solar resourcing constitutes the foremost step in grid integration of solar

energy, and the current dataset can facilitate subsequent works on siting and sizing multi-energy systems; such a workflow has been exemplified in references [64, 65].

Credit author statement

Hongrong Shi: Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing. **Dazhi Yang:** Conceptualization; Investigation; Supervision; Visualization; Roles/Writing – original draft; Writing – review & editing. **Wenting Wang:** Methodology, Software, Data curation, Writing – original draft. **Disong Fu:** Data curation; Investigation; Methodology; Software. **Ling Gao:** Data curation; Investigation; Methodology; Resources. **Jinqiang Zhang:** Data curation; Investigation; Methodology; Writing – review & editing. **Bo Hu:** Data curation; Resources; Writing – review & editing. **Yunpeng Shan:** Investigation; Methodology; Writing – review & editing. **Yingjie Zhang:** Investigation; Software; Supervision; Writing – review & editing. **Yuxuan Bian:** Data curation; Resources; Writing – review & editing. **Hongbin Chen:** Conceptualization; Investigation; Methodology; Supervision. **Xiangao Xia:** Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Roles/Writing – original draft; Writing – review & editing.

Funding

This work is supported by the National Natural Science Foundation of China [Nos., 42030608] and the Open Grants of the State Key Laboratory of Severe Weather [Nos., 2022LASW-B17].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the National Satellite Meteorological Center (NSMC) for providing access to and the guidance on the Fengyun-4 AGRI dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.rser.2023.113549>.

References

- [1] Li X, Wagner F, Peng W, Yang J, Mauzerall DL. Reduction of solar photovoltaic resources due to air pollution in China. *Proc Natl Acad Sci USA* 2017;114: 11867–72.
- [2] Yang D, Guemard CA. Producing high-quality solar resource maps by integrating high- and low-accuracy measurements using Gaussian processes, vol. 113. *Renewable & Sustainable Energy Reviews*; 2019.
- [3] Yang S, Wang XL, Wild M. Homogenization and trend analysis of the 1958–2016 in situ surface solar radiation records in China. *J Clim* 2018;31:4529–41.
- [4] Shi H, Zhang J, Zhao B, Xia X, Hu B, Chen H, Wei J, Liu M, Bian Y, Fu D, Gu Y. Surface brightening in eastern and Central China since the implementation of the clean air action in 2013: causes and implications. *Geophys Res Lett* 2021;48.
- [5] Liu H, Hu B, Zhang L, Zhao XJ, Shang KZ, Wang YS, Wang J. Ultraviolet radiation over China: spatial distribution and trends. *Renewable Sustainable Energy Rev* 2017;76:1371–83.
- [6] Feng F, Wang K. Merging ground-based sunshine duration observations with satellite cloud and aerosol retrievals to produce high-resolution long-term surface solar radiation over China. *Earth Syst Sci Data* 2021;13:907–22.
- [7] Kasten F, Duffie JA. Editorial. *Solar Energy*. 1993;50:383.
- [8] Guemard CA, Renne D, Vignola FE, Goswami Y. Editorial: journal's performance and publication criteria. *Sol Energy* 2009;83:2.
- [9] Yang D. Estimating 1-min beam and diffuse irradiance from the global irradiance: a review and an extensive worldwide comparison of latest separation models at 126 stations. *Renewable Sustainable Energy Rev* 2022;159:112195.
- [10] Yang D, Bright JM. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: a preliminary evaluation and overall metrics for hourly data over 27 years. *Sol Energy* 2020;210:3–19.
- [11] Du Y, Shi H, Zhang J, Xia X, Yao Z, Fu D, Hu B, Huang C. Evaluation of MERRA-2 hourly surface solar radiation across China. *Sol Energy* 2022;234:103–10.
- [12] Kato S, Loeb NG, Rose FG, Doelling DR, Rutan DA, Caldwell TE, Yu L, Weller RA. Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances. *J Clim* 2013;26:2719–40.
- [13] Schmetz J, Pili P, Tjemkes S, Just D, Kerkmann J, Rota S, Ratier A. An introduction to Meteosat second generation (MSG). *Bull Am Meteorol Soc* 2002;83:977–92.
- [14] Schmit TJ, Gunshor MM, Menzel WP, Gurka JJ, Li J, Bachmeier AS. Introducing the next-generation advanced baseline imager on GOES-R. *Bull Am Meteorol Soc* 2005; 86:1079–96.
- [15] Qu Z, Oumbe A, Blanc P, Espinar B, Gesell G, Gschwind B, Kluiser L, Lefevre M, Saboret L, Schroeder-Homscheidt M, Wald L. Fast radiative transfer parameterisation for assessing the surface solar irradiance: the Heliosat-4 method. *Meteorol Z* 2017;26:33–57.
- [16] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The national solar radiation data base (NSRDB). *Renew Sustain Energy Rev* 2018;89:51–60.
- [17] Letu H, Nakajima TY, Wang T, Shang H, Ma R, Yang K, Baran AJ, Riedi J, Ishimoto H, Yoshida M, Shi C. A new benchmark for surface radiation products over the east asia-pacific region retrieved from the himawari-8/AHI next-generation geostationary satellite. *Bull Am Meteorol Soc* 2022;103:E873–88.
- [18] Shi H, Li W, Fan X, Zhang J, Hu B, Husi L, Shang H, Han X, Song Z, Zhang Y, Wang S. First assessment of surface solar irradiance derived from Himawari-8 across China. *Sol Energy* 2018;174:164–70.
- [19] Yang J, Zhang Z, Wei C, Lu F, Guo Q. Introducing the new generation of Chinese geostationary weather satellites, fengyun-4. *Bull Am Meteorol Soc* 2017;98: 1637–58.
- [20] Xian D, Zhang P, Gao L, Sun R, Zhang H, Jia X. Fengyun meteorological satellite products for Earth system science applications. *Adv Atmos Sci* 2021;38:1267–84.
- [21] Letu H, Shi J, Li M, Wang T, Shang H, Lei Y, Ji D, Wen J, Yang K, Chen L. A review of the estimation of downward surface shortwave radiation based on satellite data: methods, progress and problems. *Sci China Earth Sci* 2020;63:774–89.
- [22] Tang W, Yang K, Qin J, Li X, Niu X. A 16-year dataset (2000–2015) of high-resolution (3 h, 10 km) global surface solar radiation. *Earth Syst Sci Data* 2019;11: 1905–15.
- [23] Xie Y, Sengupta M, Dudhia J. A fast all-sky radiation model for solar applications (FARMS): algorithm and performance evaluation. *Sol Energy* 2016;135:435–45.
- [24] Buster G, Bannister M, Habte A, Hettinger D, Maclaurin G, Rossol M, Sengupta M, Xie Y. Physics-guided machine learning for improved accuracy of the national solar radiation database. *Sol Energy* 2022;232:483–92.
- [25] Jiang H, Lu N, Qin J, Tang W, Yao L. A deep learning algorithm to estimate hourly global solar radiation from geostationary satellite data. *Renew Sustain Energy Rev* 2019;114:109327.
- [26] Ma R, Letu H, Yang K, Wang T, Shi C, Xu J, Shi J, Shi C, Chen L. Estimation of surface shortwave radiation from himawari-8 satellite data based on a combination of radiative transfer and deep neural network. *IEEE Trans Geosci Rem Sens* 2020; 58:5304–16.
- [27] Feng L, Qin W, Wang L, Lin A, Zhang M. Comparison of artificial intelligence and physical models for forecasting photosynthetically-active radiation. *Rem Sens* 2018;10:1855.
- [28] Hou N, Zhang X, Zhang W, Wei Y, Jia K, Yao Y, Jiang B, Cheng J. Estimation of surface downward shortwave radiation over China from himawari-8 AHI data based on random forest. *Rem Sens* 2020;12:181.
- [29] Lorenz E, Hurka J, Heinemann D, Beyer HG. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE J Sel Top Appl Earth Obs Rem Sens* 2009;2:2–10.
- [30] Mayer MJ, Grof G. Extensive comparison of physical models for photovoltaic power forecasting. *Appl Energy* 2021;283.
- [31] Mayer MJ, Grof G. Techno-economic optimization of grid-connected, ground-mounted photovoltaic power plants by genetic algorithm based on a comprehensive mathematical model. *Sol Energy* 2020;202:210–26.
- [32] Shan Y, Shi H, Fan J, Lin L, Gao L, He C, Gao M, Miao L, Zhang L, Xia X, Chen H. Revealing bias of cloud radiative effect in WRF simulation: bias quantification and source attribution. *J Geophys Res Atmos* 2022;127:11.
- [33] Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- [34] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;2:56–67.
- [35] Wei J, Huang W, Li Z, Xue W, Peng Y, Sun L, Cribb M. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Rem Sens Environ* 2019;231:111221.
- [36] Ghahremanloo M, Lops Y, Choi Y, Yeganeh B. Deep learning estimation of daily ground-level NO2 concentrations from remote sensing data. *J Geophys Res Atmos* 2021;126:21.
- [37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- [38] Perez R, Ineichen P, Seals R, Zelenka A. Making full use of the clearness index for parameterizing hourly insolation conditions. *Sol Energy* 1990;45:111–4.

- [39] Holmgren WF, Hansen CW, Mikofski MA. Pvlib python: a python package for modeling solar energy systems. *J Open Source Softw* 2018;3:884.
- [40] Wang W, Yang D, Huang N, Lyu C, Zhang G, Han X. Irradiance-to-power conversion based on physical model chain: an application on the optimal configuration of multi-energy microgrid in cold climate. *Renew Sustain Energy Rev* 2022;161:112356.
- [41] Maxwell EL. A quasi-physical model for converting hourly global horizontal to direct normal insolation. Golden, CO (USA): Solar Energy Research Inst.; 1987.
- [42] Perez R, Ineichen P, Seals R, Michalsky J, Stewart R. Modeling daylight availability and irradiance components from direct and global irradiance. *Sol Energy* 1990;44:271–89.
- [43] Carbonell D, Cadafalch J, Consul R. Dynamic modelling of flat plate solar collectors. Analysis and validation under thermosyphon conditions. *Sol Energy* 2013;89:100–12.
- [44] Martin N, Ruiz JM. Calculation of the PV modules angular losses under field conditions by means of an analytical model. *Sol Energy Mater Sol Cell* 2001;70:25–38.
- [45] Marion B. Numerical method for angle-of-incidence correction factors for diffuse radiation incident photovoltaic modules. *Sol Energy* 2017;147:344–8.
- [46] Xie Y, Sengupta M, Habte A, Andreas A. The "Fresnel equations" for diffuse radiation on inclined photovoltaic surfaces (FEDIS). *Renew Sustain Energy Rev* 2022;161:112362.
- [47] Kratochvil JA, Boyson WE, King DL. In: Photovoltaic array performance model. United States: p. Medium; 2004. p. 45.
- [48] Javed W, Guo B, Figgis B, Aissa B. Dust potency in the context of solar photovoltaic (PV) soiling loss. *Sol Energy* 2021;220:1040–52.
- [49] Mejia FA, Kleissl J. Soiling losses for solar photovoltaic systems in California. *Sol Energy* 2013;95:357–63.
- [50] Hussain N, Shahzad N, Yousaf T, Waqas A, Javed AH, Khan S, Ali M, Liaquat R. Designing of homemade soiling station to explore soiling loss effects on PV modules. *Sol Energy* 2021;225:624–33.
- [51] Valerino M, Bergin M, Ghoroi C, Ratnaparkhi A, Smestad GP. Low-cost solar PV soiling sensor validation and size resolved soiling impacts: a comprehensive field study in Western India. *Sol Energy* 2020;204:307–15.
- [52] Tang W, Qin J, Yang K, Liu S, Lu N, Niu X. Retrieving high-resolution surface solar radiation with cloud parameters derived by combining MODIS and MTSAT data. *Atmos Chem Phys* 2016;16:2543–57.
- [53] Yang D. Quantifying the spatial scale mismatch between satellite-derived solar irradiance and in situ measurements: a case study using CERES synoptic surface shortwave flux and the Oklahoma Mesonet. *J Renew Sustain Energy* 2020;12:056104.
- [54] Huang G, Li X, Huang C, Liu S, Ma Y, Chen H. Representativeness errors of point-scale ground-based solar radiation measurements in the validation of remote sensing products. *Rem Sens Environ* 2016;181:198–206.
- [55] Huang G, Li Z, Li X, Liang S, Yang K, Wang D, Zhang Y. Estimating surface solar irradiance from satellites: past, present, and future perspectives. *Rem Sens Environ* 2019;233:111371.
- [56] Wang D, Li R, Liang S, Jia A, Wang Z. Estimating global downward shortwave radiation from VIIRS data using a transfer-learning neural network. *Rem Sens Environ* 2022;274:112999.
- [57] Cui W, Dong X, Xi B, Liu M. Cloud and precipitation properties of MCSs along the Meiyu frontal zone in central and southern China and their associated large-scale environments. *J Geophys Res Atmos* 2020;125:6.
- [58] CMA. Assessment method for solar energy. 2008.
- [59] UNEP. Solar and wind energy resource assessment. 2006.
- [60] NREL. Renewable. Electricity futures study. 2012.
- [61] Qiu T, Wang L, Lu Y, Zhang M, Qin W, Wang S, Wang L. Potential assessment of photovoltaic power generation in China. *Renew Sustain Energy Rev* 2022;154:111900.
- [62] Kleissl J. Solar energy forecasting and resource assessment. Academic Press; 2013.
- [63] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The national solar radiation data base (NSRDB). *Renew Sustain Energy Rev* 2018;89:51–60.
- [64] Yang G, Yang D, Lyu C, Wang W, Huang N, Kleissl J, Perez MJ, Perez R, Srinivasan D. Implications of future price trends and interannual resource uncertainty on firm solar power delivery with photovoltaic overbuilding and battery storage. In: IEEE transactions on sustainable energy. Press; 2023.
- [65] Yang G, Zhang H, Wang W, Liu B, Lyu C, Yang D. Capacity optimization and economic analysis of PV-hydrogen hybrid systems with physical solar power curve modeling. *Energy Convers Manag* 2023;288:117128.