Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Data Augmentation-Based Estimation of Solar Radiation Components without Referring to Local Ground Truth in China

Changkun Shao[1], Kun Yang[1,2*], Yaozhi Jiang[1], Yanyi He[1], Wenjun Tang[2], Hui Lu[1,3], and Yong Luo[1]

[1]Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Institute for Global Change Studies, Tsinghua University, Beijing 100084, China. [2]National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China. [3]Tsinghua University (Department of Earth System Science)-Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Beijing 100084, China.

*Address correspondence to: yangk@tsinghua.edu.cn

Citation: Shao C, Yang K, Jiang Y, He Y, Tang W, Lu H, Luo Y. Data Augmentation-Based Estimation of Solar Radiation Components without Referring to Local Ground Truth in China. *J. Remote Sens.* 2024;4:Article 0111. https://doi.org/10.34133/remotesensing.0111

Submitted 1 April 2023
Accepted 10 January 2024
Published 6 February 2024

The power generation of bifacial photovoltaic modules is greatly related to the diffuse solar radiation component received by the rear side, but radiation component data are scarce in China, where bifacial solar market is large. Radiation components can be estimated from satellite data, but sufficient ground truth data are needed for calibrating empirical methods or training machine learning methods. In this work, a data-augmented machine learning method was proposed to estimate radiation components. Instead of using observed ground truth, far more abundant radiation component data derived from sunshine duration measured at 2,453 routine weather stations in China were used to augment samples for training a machine-learning-based model. The inputs of the model include solar radiation (either from ground observation or satellite remote sensing) and surface meteorological data. Independent validation of the model at Chinese stations and globally distributed stations demonstrates its effectiveness and generality. Using a state-of-the-art satellite product of solar radiation as input, the model is applied to construct a satellite-based radiation component dataset over China. The new dataset not only outperforms mainstream radiation component datasets, but also has significantly higher accuracy than satellite-based datasets derived from other machine learning methods trained with limited observations, indicating the superiority of our data-augmented method. In principle, this model can be applied on the global scale without additional training with local data.

Introduction

Solar energy is a clean and environmentally friendly energy source [1], and it is expected to account for the largest share of global renewable energy by 2040 [2]. Detailed knowledge of both solar radiation and radiation components is crucial for selecting, siting, and optimizing different types of solar energy systems [3,4]. For example, flat plate photovoltaic panels need solar radiation (Rs), while concentrating solar power (CSP) systems may focus on direct radiation (Rdir) [5,6]. Bifacial photovoltaic panel, a recently developed and fast-growing solar module [7], can take advantage of both sides to increase irradiance collection areas [8], and diffuse radiation (Rdif) is the one of source of the irradiance on the rear side. Therefore, not only Rs data but also Rdif and Rdir data are needed for the solar power industry. This demand is particularly evident in China where the solar energy industry is a prominent industry and growing rapidly [9].

In situ observations of radiation component are sparse and unevenly distributed at the global scale, which is rather typical in China. For instance, there are only 17 first-class radiation stations for Rdir and Rdif maintained by the China Meteorological Administration (CMA) [10]. Since weather stations are far more than radiation stations (approximately 2,400 CMA stations are available), many previous studies aimed to extend solar radiation component estimates to routine weather stations, using sunshine duration (SunDu) [5,11], clearness index [12–14], and cloud cover [15] at these stations. Due to easy access and good maintenance, SunDu is widely used for estimating Rdir [5,16] and Rdif [17,18]. Tang et al. [5] developed a SunDu-based physical parameterization method to estimate all-sky Rdir and constructed a dataset of daily Rdir and Rs at 2,472 CMA routine weather stations. He and Wang [19] demonstrated the homogeneity and reliability of SunDu-derived Rdir and Rdif and revealed the variability and trend of Rdir and Rdif in China using the derived data. In the case of scarce ground truth of radiation components, weather-station-based Rdir and Rdif estimations serve as ideal reference and complement observed radiation components.

However, routine weather stations are spatially unevenly distributed, particularly sparse in western China with huge solar potential [20], making constructing homogeneous gridded radiation component datasets challenging. Qin et al. [21] used a physical model (REST2_V9.1) to estimate direct normal irradiance from SunDu measured at 2,474 CMA stations, and a gridded dataset was constructed through interpolation of the station-based estimates. This may introduce large spatial uncertainties in western China with sparse weather stations. Homogeneous spatial-continuous Rdir and Rdif data instead of sparse station data are required for application in solar energy.

Satellite retrieval is reliable to obtain spatial-continuous solar radiation data [22], but most existing satellite-to-irradiance studies only retrieve Rs [23,24], whereas Rdir and Rdif need to be separated using additional methods. Satellite-based radiation component datasets are really scarce [10,25]. Technically, almost all such separation models are empirical in nature [26], including more than 150 regression models [14] and some machine learning methods [10,27]. No matter conventional empirical models or machine learning, adequate ground measurements of radiation component are generally required to adjust parameters or train the model [28]. Previous studies [12,14] have fitted and validated numerous advanced models using research-grade 1-min observations from stations distributed globally, except for China. The topography and climate of China are complex, and aerosols in particular are significantly higher than in other regions of the world [29], which has a significant and unique impact on solar radiation and radiation components. Chinese observations need to be included in the training or fitting dataset to ensure the performance of the model in China. However, the insufficient data volume of Chinese radiation observation may not support such a model. Some radiation component datasets over China, such as JiEA [10], a 12-year and 5-km hourly satellite-based Rdif dataset, were based on models trained and validated with observation data at several CMA stations, and it is difficult to perform sufficient independent validation of the robustness and generality of the estimation models.

Data augmentation can be used for machine learning to improve model performance and generality in the case of insufficient training data. It increases the amount of training data by adding continuous or discrete noises to existing data or creating new data with the same distribution as existing data [30]. To date, few studies have applied data augmentation to radiation component estimation. Ma et al. [25] trained a deep learning model for estimating solar radiation and radiation component with augmented training samples generated by a radiative transfer model. The trained model performs well in estimating Rs and photosynthetically active radiation with Himawari-8 satellite data as input. However, the accuracy of radiation component data was not evaluated in their study. Furthermore, the radiative transfer model may still have uncertainties in radiation component that are more sensitive to atmospheric properties such as clouds and aerosol [31]. It seems promising to find reliable alternatives or expansion for observed radiation components to augment training samples. SunDu-derived radiation component data [5] have been comprehensively validated in a previous study. These data may have systematic errors originating from algorithms or source data but are constrained by physical mechanism. These data are far more abundant and temporally homogeneous than observed Rdir and Rdif, well meeting the requirements of data augmentation. The SunDu-based dataset with stable quality may provide a data augmentation opportunity for improving the estimation accuracy of radiation components.

Thus, this study aims to build a robust Rdir and Rdif estimation model using augmented radiation component data derived from SunDu through a machine learning model, and ultimately constructing a high-quality satellite-based gridded Rdir and Rdif dataset serving for the solar energy industry in China. The paper is organized as follows. In the "Data" section, data used in this study are given. The machine learning method for estimating daily Rdir and Rdif is introduced in the "Method" section. Results on model evaluation and the new dataset are shown in the "Results" section, along with discussions on possible factors that may bring uncertainties and potential solutions. The "Implication of This Study for Solar Energy Systems" section presents implications of the model and dataset developed in this study for solar energy industry. Finally, a summary of this study is presented in the "Conclusion" section.

Materials and Methods

Data

Multiple types of data were used in this study. Station-based data were used for model training and evaluation. Two gridded datasets were used for constructing gridded Rdir and Rdif dataset and 3 existing radiation component datasets were used for comparison. Some basic information of the data used in this study can be seen in Table 1. The details about the selection and processing of the data are illustrated in the "Station-based data for model training" section up to the "Gridded data for construction of the Rdif and Rdir dataset" section.

Station-based data for model training

For building the model, 2 types of station data are used: one is meteorological observations, including near-surface air temperature (Temp), relatively humidity (RH), and near-surface pressure (P) at 2,453 CMA weather stations (blue circles in Fig. 1) during 1961 to 2018; the other is Rdir and Rs derived from SunDu at the same stations for the same period, and the Rdif is yielded from the difference between SunDu-derived Rs and Rdir. All the SunDu-derived data are provided by Tang et al. [5]. Spatial distribution of CMA weather stations can be seen in Fig. 1. With Rdif and Rdir derived from SunDu as the target variables, SunDu-derived Rs, Temp, RH, and P are fed into the model for training. Details about training are illustrated in the "Methods" section. The selection of input variables is based on the principle that the input variables do have impacts on or correlate with Rdif and Rdir and there are high-quality gridded datasets of the input variables so that we can generate gridded Rdif and Rdir datasets using the trained model. Furthermore, the calendar date is also used as an auxiliary input variable to represent seasonal variations of the solar radiation. The above data were included in the augmented dataset, and only used for training the model.

Station-based data for evaluation

For evaluating the model performance and the final datasets, directly observed Rdif and Rdir at 17 independent CMA radiation stations (red rhombuses in Fig. 1) were used in this study. Observed data since 1994 were used, mainly because sensitivity drift caused by aging instruments before 1990 and the discontinuity caused by instruments replacement between 1990 to

Table 1. The information of the data used in this study

Groups	Data	Usage
Model training	SunDu-derived Rdif and Rdir at 2,453 CMA stations	Label data for training
	SunDu-derived Rs and meteorological observations at 2,453 CMA stations	Input data for training
Model evaluation	Rs observations and meteorological observations at 17 CMA stations	Input data for estimation
	Rdif and Rdir observations at 17 CMA stations	Ground truth for evaluation
	Rs observations and meteorological observations at 16 GSOD stations	Input data for estimation
	Rdif and Rdir observations at 16 BSRN stations	Ground truth for evaluation
Dataset construction and evaluation in China	ISCCP-ITP-CNN Rs dataset and CMFD meteorological dataset	Input data for dataset construction
	Rdif and Rdir observations at 17 CMA stations	Ground truth for evaluation
Data intercomparison	CERES-SYN	Comparison with the dataset developed in this study
	ERA5 reanalysis	
	JiEA	

BSRN, Baseline Surface Radiation Network; GSOD, Global Surface Summary of the Day.

1993 introduce huge inhomogeneities in observed radiation [32–35]. A quality control scheme developed by Tang et al. [36] and used in previous studies [10,37] was adopted to exclude the erroneous values in the observed radiation data. All data at these stations were only used for independent validation and are excluded in training.

Yang [14] indicated that the performance of the model at independent locations is what really matters. Thus, observed data from the Baseline Surface Radiation Network (BSRN) [38] and Global Surface Summary of the Day (GSOD) during 2001 to 2015, which were not included in the training data and are not collocated with CMA stations, are used for an additional evaluation of the generality and robustness of the model. Details about the test are also illustrated in the "Methods" section. Radiation observations from BSRN are regarded as the most reliable observation data, benefiting from widely recognized accuracy and well-maintenance instruments (https://bsrn.awi.de/). Since only radiation observations are available at the BSRN stations, the meteorological variables at these stations are from GSOD. The GSOD dataset is produced and archived at the National Oceanic and Atmospheric Administration Climatic Data Center. It can be obtained from the website: https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day, including daily measurements for meteorological variables used in this study [39]. Because the stations of BSRN and GSOD are not exactly in the same location, the selected GSOD station must be within a 1-km radius centered on a BSRN station and has valid observations covering the same periods with the corresponding BSRN station. Sixteen pairs of GSOD (red crosses in Fig. 2) and BSRN stations (blue circles in Fig. 2) were selected, as shown in Fig. 2.

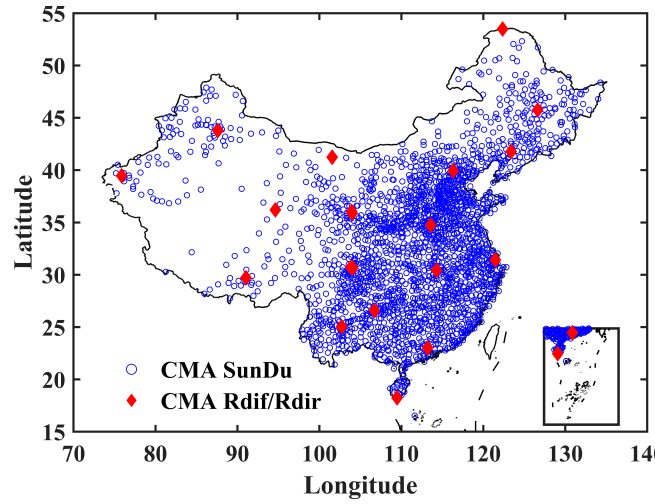

Fig. 1. Spatial distribution of CMA stations. A total of 2,453 blue circles represent routine weather stations, which have SunDu measurements. Seventeen red rhombuses represent radiation stations that have Rdir and Rdif observations.

Gridded data for construction of the Rdif and Rdir dataset

In this study, an up-to-date long-term satellite-based Rs dataset (ISCCP-ITP-CNN) and a widely used meteorological dataset in China (China Meteorological Forcing Dataset [CMFD]) were fed into the trained model to construct a gridded Rdir and Rdif dataset.

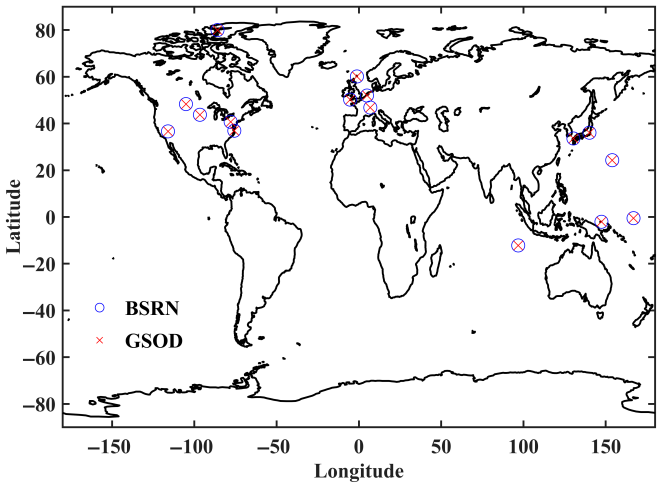

Fig. 2. Spatial distribution of the selected BSRN and GSOD stations. Blue circles represent BSRN stations and red crosses represent GSOD stations.

ISCCP-ITP-CNN is a 36-year (1983 to 2018), 10-km global Rs dataset developed by Shao et al. [40] following Tang et al. [24]. It is a satellite-based Rs dataset that has eliminated inhomogeneities in the ISCCP-ITP dataset [24] caused by different sources of satellite data. A comparison with existing global solar radiation datasets indicates that ISCCP-ITP-CNN shows the best agreement with observation, no matter in accuracy and variation in Rs [40]. In this study, ISCCP-ITP-CNN was used as the Rs input of the trained model.

CMFD is a high spatial–temporal resolution near-surface meteorological dataset, and it has been proved with high accuracy in China [41]. The spatial resolution of CMFD is 0.1° and its temporal resolution is 3 h. The dataset comes from the fusion of the in situ observation, the Princeton reanalysis, the Global Land Data Assimilation System data, and the Tropical Rainfall Measuring Mission precipitation data. The dataset is one of most widely used meteorological datasets in China. In this study, air temperature, specific humidity, and near-surface pressure from CMFD were used as the meteorological input of the trained model.

Gridded data for intercomparison

ECMWF Reanalysis v5 (ERA5) is the latest generation of global atmospheric reanalysis datasets produced by the European Centre for Medium-Range Weather Forecasts, providing a historical dataset of surface solar radiation from 1950 to the present [42]. In this study, its Rs and Rdir were used to derive Rdif, and the Rdir and Rdif components were evaluated.

CERES_SYN1deg_Ed4A (CERES-SYN) is a level 3 satellite product, designed to provide global diurnally complete surface fluxes since 2001 [43]. CERES surface fluxes are computed using 16 geostationary orbit (GEO) and Moderate-Resolution Imaging Spectroradiometer (MODIS) satellite-derived cloud properties. Its Rdir and Rdif were used in this study.

JIEA is a 12-year (2007 to 2018) hourly dataset over China derived from the Multi-functional Transport Satellite observations through a deep learning technique [10]. It provides Rs and Rdif with a spatial resolution of 5 km, and their difference yields Rdir. Its high accuracy at CMA stations and fine spatial pattern have been proved, and it is a good reference for comparison with our new product.

The 3 datasets are compared with the dataset constructed by the model.

Methods

The flowcharts of the method developed in this study are shown in Fig. 3, including training and validation (Fig. 3A), station-based independent test (Fig. 3B), and dataset construction (Fig. 3C). As Fig. 3A shows, 2 regression models based on the Light Gradient Boosting Machine (LightGBM) model estimate Rdir and Rdif, respectively; they share the same input variables but are trained separately. The augmented train set, i.e., the SunDu-derived Rs and meteorological observations at 2,453 training stations (blue circles in Fig. 1), were fed into the model and SunDu-derived Rdir (or Rdif) at these stations were set to be label data while training. The Rdir and Rdif outputted from the 2 models need to be adjusted to ensure that their sum equals to the input Rs. The validation set, namely, the observed Rs and meteorological variables at the independent 17 radiation stations (red rhombuses in Fig. 1), were fed into the trained model to estimate Rdir and Rdif, which are then evaluated with the observed Rdif and Rdir at these 17 stations. The validation aims to evaluating the overall performance of the models and the results are shown in the "Evaluation of the model with CMA and BSRN data" section.

Then, BSRN and GSOD observations were used to test the generality and robustness of the trained model (Fig. 3B). Observed Rs at BSRN stations and meteorological variables at GSOD stations were fed into the trained model to estimate Rdif and Rdir and the results were evaluated against BSRN Rdir and Rdif observations. The evaluation results are also shown in the "Evaluation of the model with CMA and BSRN data" section.

Finally, the trained model was applied to construct a 36-year (1983 to 2018), 10-km daily Rdif and Rdir dataset over China, with the satellite-based ISCCP-ITP-CNN and CMFD as input (Fig. 3C). Before ISCCP-ITP-CNN was fed into the trained model, Rs in each pixel is spatially averaged over adjacent 3×3 pixels (approximately in a 30 km × 30 km area), because Tang et al. [24] found that the accuracy of the satellite-based Rs dataset used in this study shows a significant improvement when it is upscaled to more than 30 km. The gridded Rdif and Rdir datasets were finally evaluated using observed Rdir and Rdif from 17 CMA stations (red rhombuses in Fig. 1) and the evaluation results were compared with other methods based on several gridded datasets to show the superiority of our model. The evaluation and comparison results are shown in the "Evaluation for the dataset constructed by the model" and "Comparison with other gridded Rdif and Rdir datasets" section, respectively.

Data augmentation method

To gain good performance, modern machine learning methods usually need large amounts of high-quality annotated data, which are usually obtained through observations, but it is often not feasible to obtain sufficient training data in many real-world applications [44]. Data augmentation, i.e., increasing the amount of data, is the most effective way of alleviating this problem. One of data augmentation is data synthesis [45], which creates new data samples independent from existing data [45]. Data scarcity of observed radiation component is rather typical in the world and it makes barriers for developing an effective radiation component estimation model. Augmenting the train set may be helpful to increase the model performance and robustness. Given that observed radiation components are too scarce and are needed for evaluating

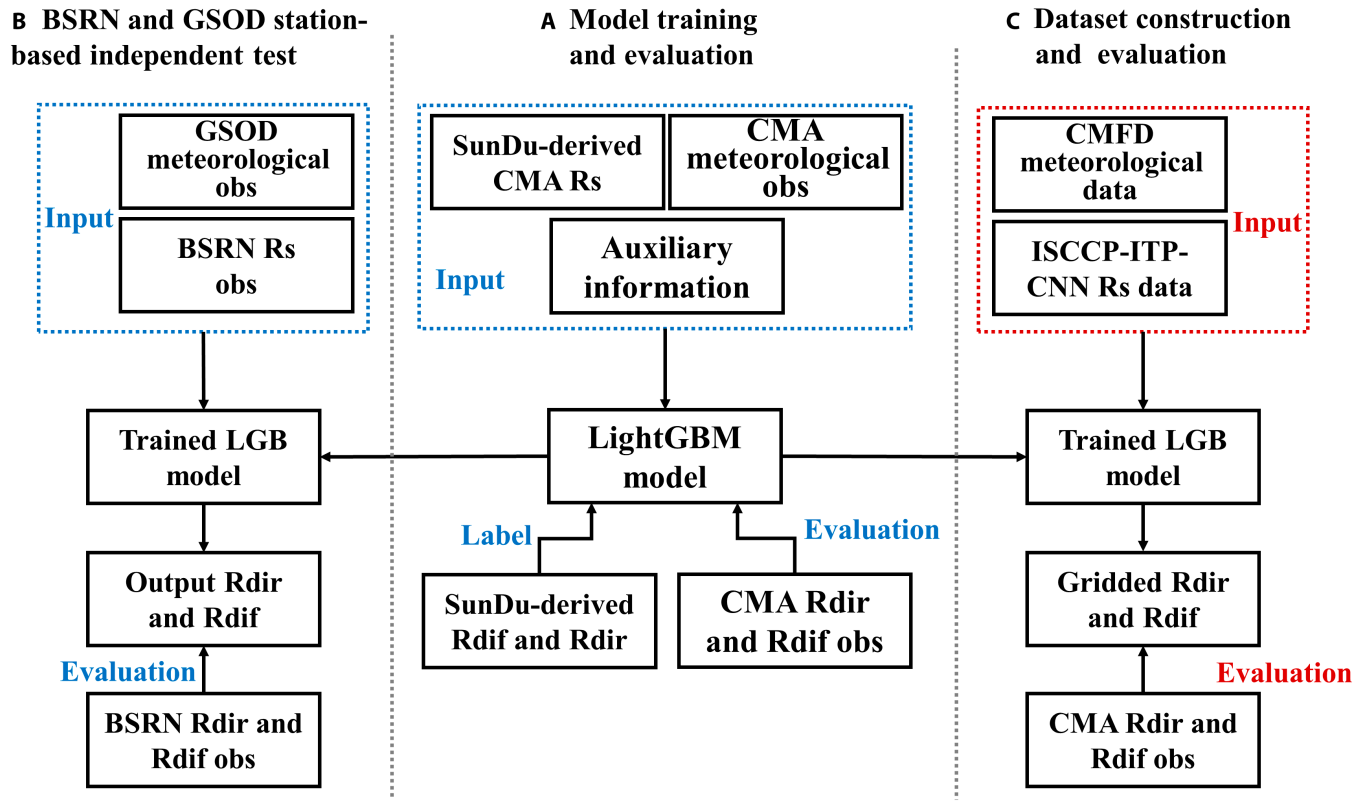

Fig. 3. Flowchart of the method developed in this study. (A) Model training and evaluation using CMA data. (B) Independent test for the model using BSRN and GSOD data. (C) Gridded Rdif and Rdir dataset construction and comparison with other gridded datasets. Rs, Rdif, and Rdir refer to global, diffuse, and direct solar radiation, respectively. Obs refers to observed data. LGB refers to LightGBM.

a constructed radiation dataset, data synthesis was used in this study. Abundant radiation component data synthesized by the SunDu-based physical parameterization scheme [24] are used to augment the training set, as mentioned above. The data augmentation is conducted prior to the entire flowchart, known as offline augmentation. The augmented dataset has in total 16,928,343 data points, about 147 times observed ground truth.

LightGBM model

The method developed in this study was based on LightGBM [46], a fast and efficient implementation of the gradient boosting decision tree (GBDT) model. GBDT adds up the results from multiple decision trees [47], which can be written as Eq. 1:

$$f(x) = \sum_{i=1}^{M} \beta_i h(x;\ \theta_i) \tag{1}$$

where x represents the input sample, β_i and θ_i represent the weight and distributed parameter of the ith decision tree, and $h(x;\theta_i)$ represents the ith decision tree, respectively. The final model is a weighted sum of M trees. To optimize the model, the trained model f tends to minimize the loss function L:

$$\min \sum_{j=1}^{N} L\left(f\left(x_j\right), y_j\right) = \min \sum_{j=1}^{N} L\left(\sum_{i=1}^{M} \beta_i h\left(x_j;\theta_i\right), y_j\right) \tag{2}$$

where y and $f(x)$ represent the observed value and output from the model, respectively. L is the loss function and $\left(\left\{\left(x_j, y_j\right)\right\}_{J=1}^{N}\right.$ is the training sample. $\sum_{i=1}^{M-1} \beta_i h\left(x;\theta_i\right)$ is assumed to be known during the whole training process. The gradient descent algorithm is used to let $\beta_M h(x;\theta_M)$, y_j fit the negative gradient g_j, which can be written as:

$$g_j = \frac{\partial L\left(f_{M-1}\left(x_j, y_j\right)\right)}{\partial f_{M-1}\left(x_j\right)} \tag{3}$$

The fitting of new samples can be represented by Eq. 4:

$$\theta_M, \beta_M = \underset{\beta,\theta}{\operatorname{argmin}} \sum_{j=1}^{N} \left\| g_j - \beta h\left(x_j;, \theta\right)\right\|^2 \tag{4}$$

The final step size that minimizes the loss function can be determined and used to determine the final model:

$$\rho_m = \underset{\rho}{\operatorname{argmin}} \sum_{j=1}^{N} L\left(f_{M-1}\left(x_j\right) + \rho h\left(x_j\right), y_j\right) \tag{5}$$

Based on the parameters determined, the final model can be written as Eq. 6:

$$f(x) = f_{M-1} + \rho_M \beta_M h\left(x;\theta_M\right) \tag{6}$$

As one of the GBDT models, the schematic diagram of the LightGBM is shown in Fig. 4. The decision tree in the LightGBM model only splits along the optimal direction. LightGBM shows a significant decrease in computational cost and works better in massive data processing compared to traditional GBDT models.

In this study, the model was constructed using Python sklearn [48] and the LightGBM package (https://github.com/microsoft/LightGBM). The number of leaves, trees, and rounds was set to be 200, 500, and 20, respectively. RMSE loss is adopted as loss function.

Evaluation metrics

Three metrics were used to evaluate the results: correlation coefficient (CC), root mean square error (RMSE), and mean bias error (MBE). Relative values of MBE and RMSE (rMBE and rRMSE) were also used. They were calculated as follows:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i'-y_i\right)^2} \tag{7}$$

$$\mathrm{MBE}=\frac{1}{N}\sum_{i=1}^{N}\left(y_i'-y_i\right) \tag{8}$$

$$\mathrm{CC}=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)\left(y_i'-\overline{y'}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i'-\overline{y'}\right)^2}} \tag{9}$$

$$\mathrm{rRMSE}=\frac{\mathrm{RMSE}}{\overline{y_i}} \tag{10}$$

$$\mathrm{rMBE}=\frac{\mathrm{MBE}}{y_i} \tag{11}$$

where N is the total number of data used for evaluation; i is the ith evaluated data and observation data; y' means evaluated data and $\overline{y'}$ means the average of evaluated data; y means observation data and $\overline{y}$ means the average of observation data. All valid CMA observations since 1994 were used for evaluating the model performance and our new dataset. Because JiEA has only become available since 2007, CMA observations from 2007 to 2015 were used when comparing our new dataset in this study with other gridded datasets.

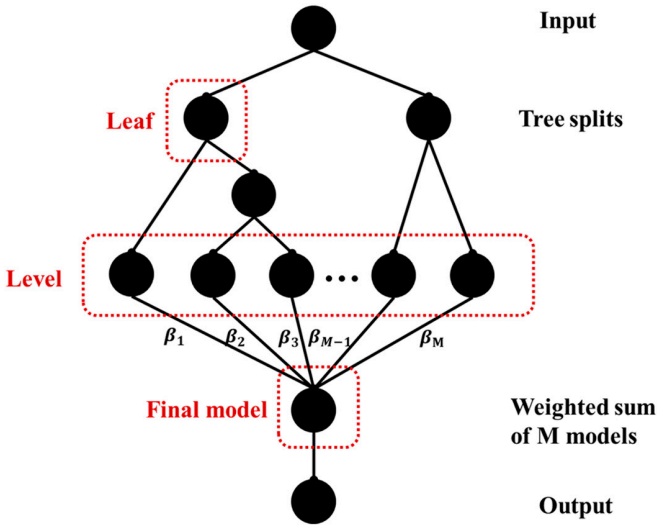

Fig. 4. The schematic diagram of the LightGBM model. Each black dot represents a leaf node and all leaf nodes at the last level are weighted sum to the final output.

Results and Discussion

In the following 3 subsections, we evaluated the trained model, and the new Rdir and Rdif dataset developed in this study and compared the new dataset with other gridded datasets. Possible factors related to uncertainties of the new dataset were also analyzed.

Evaluation of the model with CMA and BSRN data

For evaluating the model performance, observed Rs and meteorological variables at 17 CMA radiation stations were fed into the trained model and the estimated results were evaluated against observed Rdir and Rdif at these stations (i.e., the results of evaluation indicated in Fig. 3A). As a reference, SunDu-derived Rdir and Rdif, which were used as label data when training, were also evaluated at these stations. Figure 5 presents the evaluation results. The correlation coefficients of the estimated Rdif and Rdir (Fig. 5B and D) are 0.87 and 0.97, respectively, much better than those of the SunDu-based data (0.81 and 0.94, Fig. 5A and C). The RMSEs of the estimated Rdif and Rdir (Fig. 5B and D) are 20.0 W/m^2 (rRMSE = 26.3%) and 19.8 W/m^2 (22.2%), respectively, much better than those of the SunDu-based data (23.7 W/m^2 [31.3%] and 27.6 W/m^2 [30.9%], Fig. 5A and C), too. Though trained by augmented noisy radiation component data, the trained model shows a great performance in separating radiation components from observed Rs, demonstrating the effectiveness of our data augmentation strategy. Nevertheless, an "upper boundary" around 140 W/m^2 can be found in Fig. 5A and B. It may be originated from the error in SunDu-derived Rdif. SunDu records the time duration during a day when the direct solar beam is greater than 120 W/m^2. Under an overcast sky, the Rdir is always below 120 W/m^2, the SunDu can be equal to 0, and the increase/decrease of the Rdif due to stronger/weaker atmospheric scattering effects can no longer be reflected by SunDu values. Therefore, there is a threshold value above which Rdif cannot be estimated by the SunDu-based method, as demonstrated by the "upper boundary" in Fig. 5A and B.

For evaluating the generality of the model, observed Rs at BSRN stations and meteorological variables at GSOD stations were fed into the model and estimated results were evaluated against BSRN-observed Rdir and Rdif. Figure 6 presents the evaluation results (i.e., the results of evaluation indicated in Fig. 3B). The model performs reasonably well at BSRN stations, with an RMSE of 23.0 W/m^2 (30.2%, Fig. 6A) for Rdif and 23.2 W/m^2 (27.1%, Fig. 6B) for Rdir, demonstrating the generality and robustness of the trained model.

Evaluation for the dataset constructed by the model

The satellite-based ISCCP-ITP-CNN Rs and CMFD meteorological datasets rather than station-based data were fed into the trained model to construct a 10-km and 36-year (1983 to 2018) daily Rdif and Rdir dataset. Figure 7 presents the evaluation for the daily and monthly datasets (new Rdif and Rdir data) against observations during 1994 to 2015 at 17 CMA stations (i.e., the results of evaluation indicated in Fig. 3C). The correlation coefficient, RMSE, and MBE of the new Rdif (Fig. 7A) are

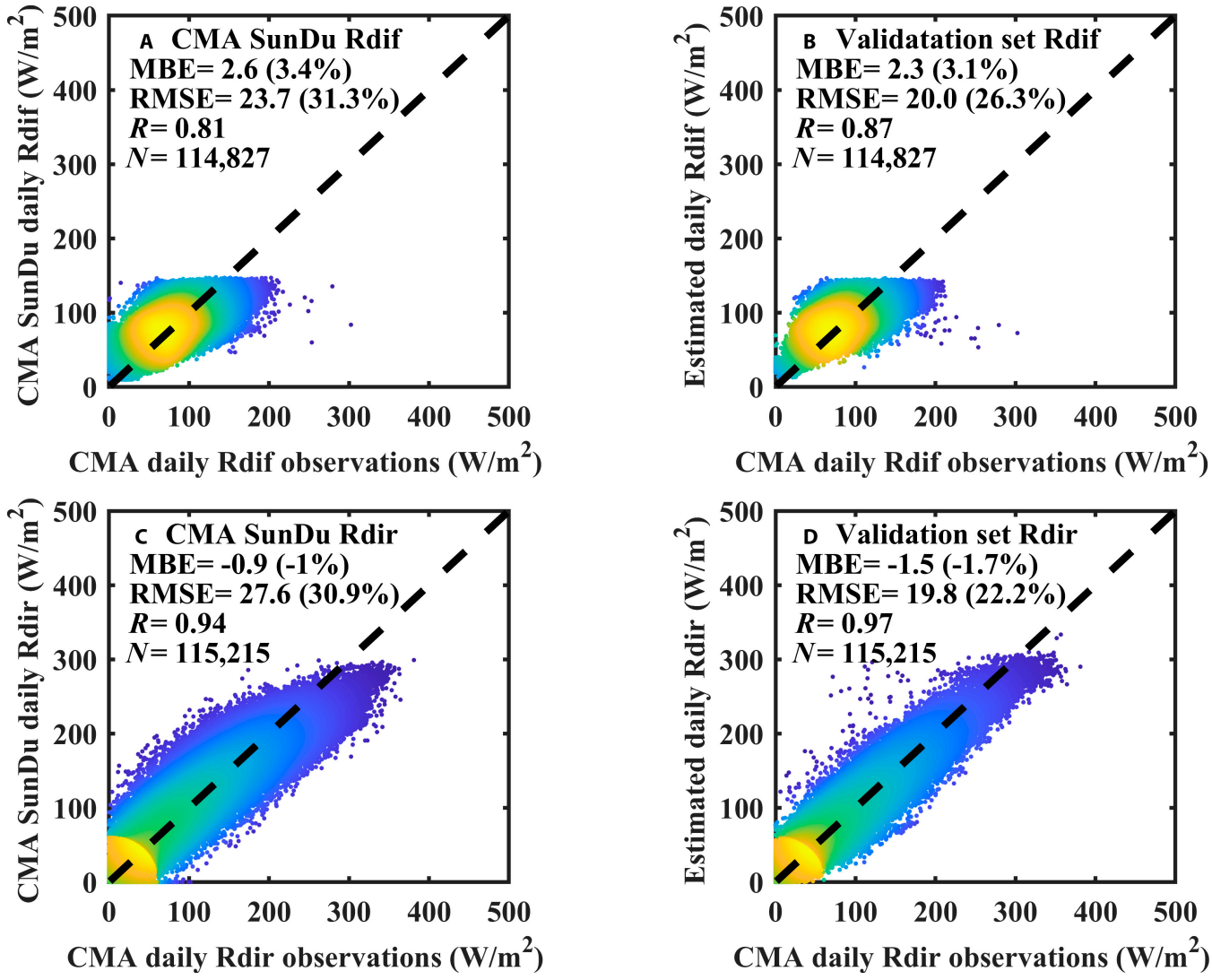

Fig. 5. Evaluation for daily Rdif (A) and Rdir (C) from SunDu-derived data and the estimated daily Rdif (B) and Rdir (D) from the trained model during 1994 to 2015 at 17 CMA validation stations.

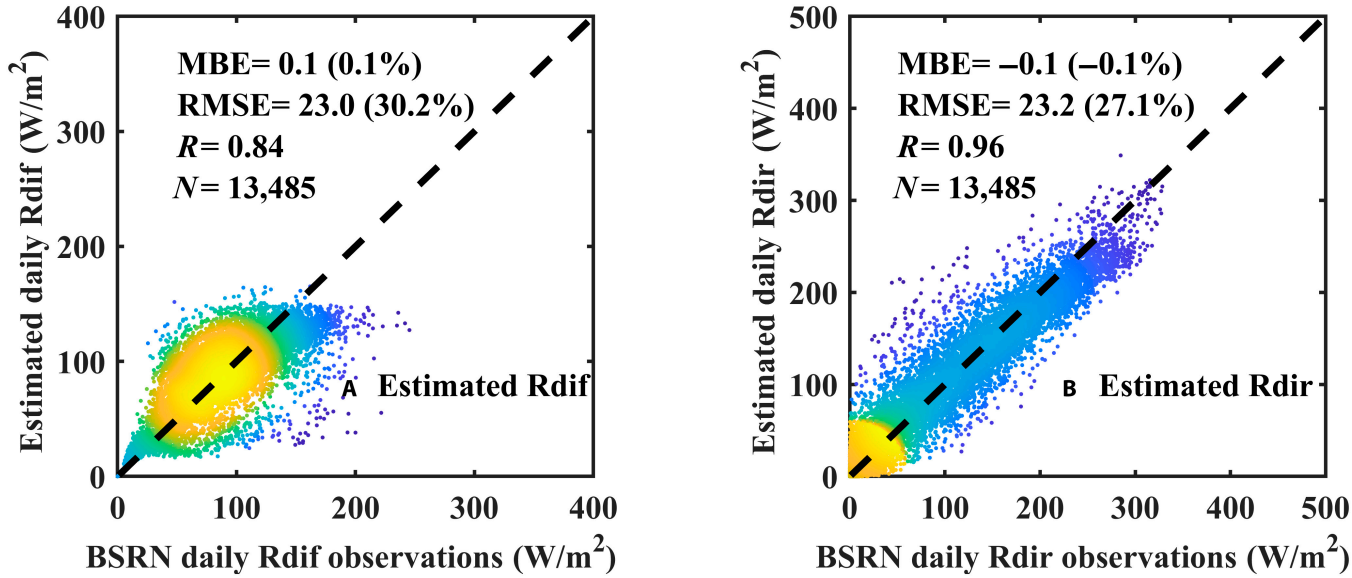

Fig. 6. Evaluation for estimated daily Rdif (A) and Rdir (B) during 2001 to 2015 from BSRN Rs and GSOD meteorological observations at 16 test stations.

0.86, 21.3 W/m^2 (28.0%), and 5.1 W/m^2 (6.7%), respectively, better than those of SunDu-derived Rdif (Fig. 5A). In contrast, with a correlation coefficient of 0.91 and an RMSE of 32.8 W/m^2 (36.8%), the new Rdir (Fig. 7B) is slightly poorer than SunDu-derived Rdir (Fig. 5C). We develop a better Rdif dataset than SunDu-derived data, possibly due to the spatial average of Rs that takes the spatial effects of clouds into account, as indicated by Jiang et al. [49]. At the monthly scale, the RMSE of new Rdif and Rdir are 10.5 W/m^2 (14.1%, Fig. 7C) and 18.6 W/m^2 (20.7%, Fig. 7D), indicating high accuracy at the monthly scale. Even if the input is replaced with satellite-retrieval Rs, the model still performs stably in separating radiation components.

Table 2 presents the evaluation results for daily new Rdir and Rdif datasets at each station during 1994 to 2015. The Rdif dataset shows high accuracy at all 17 stations, with rRMSE < 35%. The Rdir dataset shows acceptable accuracy at 14 stations (rRMSE < 40%) but has significant uncertainties at Chengdu, Guiyang, and Guangzhou stations (rRMSE > 50%). Stations with high rRMSE are mainly located at South China, where there are more clouds and lower Rs and Rdir. We consider 2 possible factors related to the relatively poor performance of Rdir: quality of input data and weakness of machine learning models in estimating extreme values.

To avoid performance degradation caused by overfitting, we usually prevent the model from completely fitting the train set. The model tends to give estimations within a narrower range of values than the train set. In other words, the model would overestimate at low values but underestimate at high values. In a rare case, the machine learning model will even map all inputs to the same near-average value, which is mathematically valid but inconsistent with the real world [50]. The more discrete the training data, the greater the errors caused by this problem. It can be seen in scatter plots (such as Figs. 5 to 7) that Rdir data are more discrete than Rdif; thus, there may be more difficulties in Rdir estimation. At cloudy and rainy regions, such as southern and southwestern China, daily Rdir appears to be low more often and can be even equal to zero, implying that estimation of Rdir has huge uncertainties at these regions. For Rdif, the insignificant spatial difference and narrower value range of Rdif may contribute to generally high accuracy at all stations. Furthermore, uncertainties may also come from the input data. Tang et al. [24] revealed the poor performance of ISCCP-ITP Rs at 9 CMA stations (RMSE > 35 W/m^2) in cloudy southern China. Inaccurate Rs input may have a negative impact on the estimation. Thus, the Rdir dataset developed in this study may have higher uncertainties in these regions.

To further improve the accuracy of the Rdir dataset in future, we need more accurate input data in the model. Alternatively, merging SunDu-derived Rdir at stations with the Rdir dataset developed in this study might somewhat improve the accuracy.

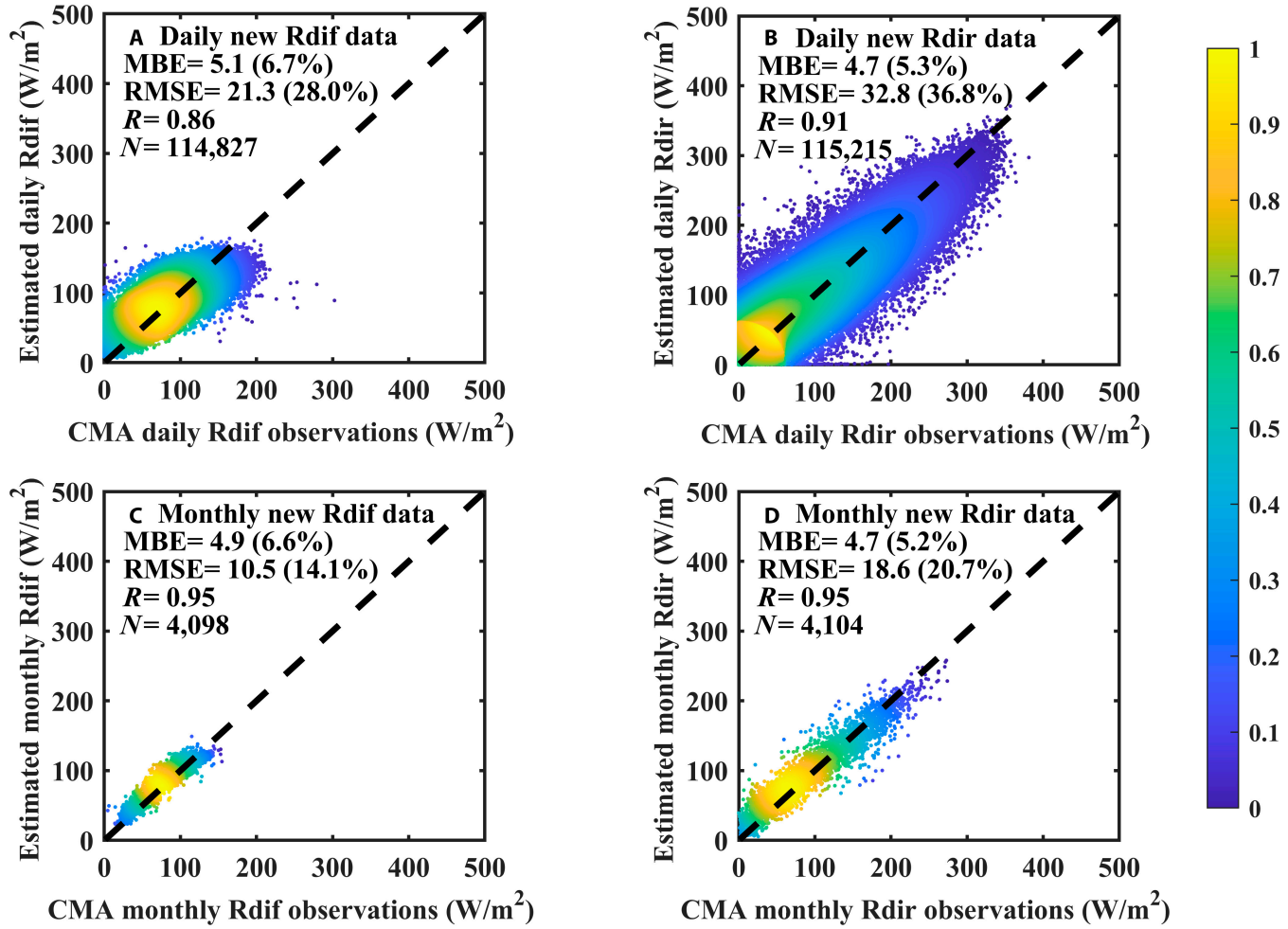

Fig. 7. Evaluation for daily Rdif (A) and Rdir (B) and monthly Rdif (C) and Rdir (D) from the new dataset developed in this study during 1994 to 2015 against observations at 17 CMA stations.

Table 2. Evaluation metrics for daily Rdir and Rdif datasets developed in this study at every CMA radiation station

Station	Longitude	Latitude	Rdir (W/m^2)			Rdif (W/m^2)		
			MBE (rMBE)	RMSE (rRMSE)	*R*	MBE (rMBE)	RMSE (rRMSE)	*R*
Mohe	122.3	53.5	6.8 (9.4%)	29.7 (39.0%)	0.92	5.9 (8.7%)	20.8 (31.0%)	0.86
Harbin	126.6	45.8	3.5 (4.2%)	27.1 (31.8%)	0.92	5.6 (7.6%)	16.8 (22.7%)	0.91
Urumchi	87.6	43.8	7.2 (6.4%)	30.0 (26.7%)	0.95	11.8 (16.2%)	23.7 (32.6%)	0.80
Kashgar	75.9	39.5	18.7 (14.7%)	36.0 (28.2%)	0.93	3.3 (4%)	22.9 (28.1%)	0.82
Ejinaqi	101.6	41.2	−8.6 (−6.8%)	33.4 (26.5%)	0.92	3.8 (5.0%)	23.4 (30.9%)	0.83
Geermu	94.6	36.2	−0.6 (−0.4%)	31.2 (21.6%)	0.93	5.4 (6.8%)	23.5 (29.6%)	0.87
Lanzhou	103.9	36.0	−2.2 (−2.2%)	25.4 (25.3%)	0.95	2.4 (2.9%)	17.3 (21.4%)	0.85
Shenyang	123.4	41.8	6.2 (7.3%)	27.2 (31.9%)	0.91	2.5 (3.3%)	17.5 (22.4%)	0.91
Beijing	116.3	39.9	7.1 (7.9%)	26.8 (30.1%)	0.93	4.9 (6.2%)	19.4 (24.3%)	0.90
Lhasa	91.0	29.7	−14.3 (−8.9%)	38.9 (24.3%)	0.88	16.0 (17.6%)	27.2 (30.0%)	0.83
Chengdu	104.1	30.7	15.9 (44.6%)	28.2 (79.1%)	0.84	5.9 (7.5%)	22.8 (29.2%)	0.87
Kunming	102.7	25.0	7.2 (7.1%)	28.2 (28.0%)	0.93	5.2 (6%)	20.9 (24.2%)	0.84
Zhengzhou	113.6	34.8	9.1 (13.4%)	24.5 (36.1%)	0.92	−2.0 (−2.3%)	18.0 (21.0%)	0.92
Wuhan	114.3	30.4	4.8 (8.4%)	24.6 (37.8%)	0.92	2.1 (2.7%)	19.3 (24.4%)	0.90
Guiyang	106.7	26.6	8.8 (19.4%)	26.8 (59.2%)	0.90	−5.8 (−7.9%)	23.0 (31.2%)	0.85
Guangzhou	113.2	23.0	8.9 (16.1%)	27.2 (48.9%)	0.87	6.5 (7.5%)	21.2 (24.3%)	0.80
Sanya	109.5	18.2	−3.6 (−3.7%)	37.5 (39.2%)	0.85	8.6 (8.9%)	23.9 (24.7%)	0.65
Overall	–	–	4.7 (5.3%)	32.8 (36.8%)	0.91	5.1 (6.7%)	21.3 (28.0%)	0.86

Table 3. Evaluation metrics of new daily radiation component data and other gridded datasets against daily observations during 2007 to 2015 at 17 CMA stations

Data		New data	JiEA	CERES-SYN	ERA5
Rdif (W/m^2)	RMSE	20.6 (26.1%)	24.6 (31.2%)	31.8 (40.3%)	31.9 (40.4%)
	MBE	2.5 (3.1%)	−1.4 (−1.8%)	18.8 (23.8%)	−14.3 (−18.0%)
	R	0.86	0.84	0.84	0.72
Rdir (W/m^2)	RMSE	31.8 (36.2%)	36.4 (41.5%)	38.9 (44.3%)	56.5 (64.3%)
	MBE	4.8 (5.5%)	6.4 (7.2%)	−12.9 (−14.7%)	32.7 (37.2%)
	R	0.92	0.90	0.89	0.82

Comparison with other gridded Rdif and Rdir datasets

To demonstrate the superiority of our method, we compared the new dataset with JiEA, the other dataset based on machine learning, and CERES-SYN and ERA5, the representatives for global satellite-retrieval and reanalysis, using CMA observations during 2007 to 2015 (Tables 3 and 4).

As Table 3 shows, with a correlation coefficient of 0.86 and an RMSE of 20.6 W/m^2 (rRMSE of 26.1%), the new Rdif is more accurate than JiEA (24.6 W/m^2, 31.2%), CERES-SYN (31.8 W/m^2, 40.3%), and ERA5 (31.9 W/m^2, 40.4%). MBE results indicate that ERA5 and CERES-SYN obviously misrepresent Rdif. Jiang et al. [49] point out that the presence of nonhomogeneous clouds and their induced radiation interactions make estimation of Rdif scale-dependent. Thus, the other three higher-resolution datasets were upscaled to the same spatial resolution of 1° as CERES-SYN and re-evaluated, as is shown in Table 4. At a spatial resolution of 1°, all these 3 Rdif datasets show degradation of the accuracy, but new data still outperform the other datasets.

New Rdir outperforms the other 3 datasets (Table 3), too, with a correlation coefficient of 0.92 and an RMSE of 31.8 W/m^2 (36.2%), followed by JiEA (36.4 W/m^2, 41.5%), CERES-SYN (38.9 W/m^2, 44.3%), and ERA5 (56.5 W/m^2, 64.3%). Similarly, ERA5 and CERES-SYN also obviously misrepresent Rdir. JiEA shows a poorer performance than the new dataset for both Rdir and Rdif, though these 2 datasets are both based on machine learning methods. The model for constructing JiEA is trained by a very limited hourly CMA observed radiation component. A

Table 4. Evaluation metrics of new daily radiation component data and other gridded datasets at 1° spatial resolution against daily observations during 2007 to 2015 at 17 CMA stations

Data		New data	JiEA	CERES-SYN	ERA5
Rdif (W/m^2)	RMSE	22.1 (27.6%)	27.6 (34.5%)	31.8 (40.3%)	33.5 (41.9%)
	MBE	1.8 (2.3%)	−3.7 (−4.6%)	18.8 (23.8%)	−14.7 (−18.4%)
	R	0.86	0.76	0.84	0.71
Rdir (W/m^2)	RMSE	32.0 (36.5%)	35.6 (40.7%)	38.9 (44.3%)	56.6 (64.6%)
	MBE	3.8 (4.3%)	6.5 (7.4%)	−12.9 (−14.7%)	32.9 (37.6%)
	R	0.92	0.90	0.89	0.82

model based on limited training data may not be strong enough to estimate radiation components. Conversely, with an augmented training dataset included massive SunDu-derived data, the model developed in this study has a better and more stable performance. As Table 4 shows, new Rdir data derived from our model still have the highest accuracy when being upscaled.

Implication of This Study for Solar Energy Systems

With the help of high-resolution spatial-continuous input satellite data, the fine spatial distribution of radiation components is revealed in this study. Figure 8 presents the multi-year average Rdif and Rdir from the new dataset during 1984 to 2018. As shown in Fig. 8A, Rdif ranges from 65.4 to 99.7 W/m^2. Low latitude and cloudy weather may contribute to higher Rdif. The southeastern Tibetan Plateau, eastern China, and the southern coastal region receive higher Rdif. The lower value of Rdif appears in northern China and the Sichuan Basin. Conversely, the average Rdir value is high at low latitudes and high altitudes, as shown in Fig. 8B, owing to the high solar radiation intensity and weak atmospheric scattering effect. In areas with high frequent cloud cover, solar radiation absorption and scattering increase, causing considerably low Rdir. Therefore, Rdir has a wide range from 35.4 to 194.7 W/m^2. The maximum Rdir is on the southwest of the Tibetan Plateau, while the minimum is on southwestern China, especially on the Sichuan Basin. Due to the high altitude and cloudy weather, the Hengduan Mountains (22 to 32°N, 97 to 103°E) at the southeastern Tibetan Plateau receive both Rdir (average values over 140 W/m^2) and Rdif (average values over 90 W/m^2) at a relatively high level. In Fig. 8A, a spatial pattern with discontinuities, although not obvious, can be seen in North China around Beijing. This artifact is caused by the spatial discontinuities of Rs in the ISCCP-ITP-CNN dataset and has a minor impact on the accuracy of the dataset.

Figure 9 presents the multi-year average Rdir/Rs from the new dataset during 1984 to 2018. Rdif is higher than Rdir (Rdir/Rs < 50%) in eastern and southern China while Rdir is dominant in northern China and the Tibetan Plateau. As we mentioned above, the intensity and spatial distribution of radiation components are crucial for the deployment of different solar energy modules. CSP modules may be suitable in areas where Rdir is dominant and high, such as the southwestern Tibetan Plateau. Bifacial panels can take advantages of both sides to increase their collection area and Rdif is one of the sources of the irradiance on the rear side of the module. It is effective to deploy bifacial photovoltaic panels in areas with comparably high Rdir and Rdif, such as the eastern Tibetan Plateau and northwestern China, so that the solar power systems can take full advantage of solar energy. In contrast, in areas with low Rdif, the additional benefits of deploying bifacial photovoltaic panels may be reduced.

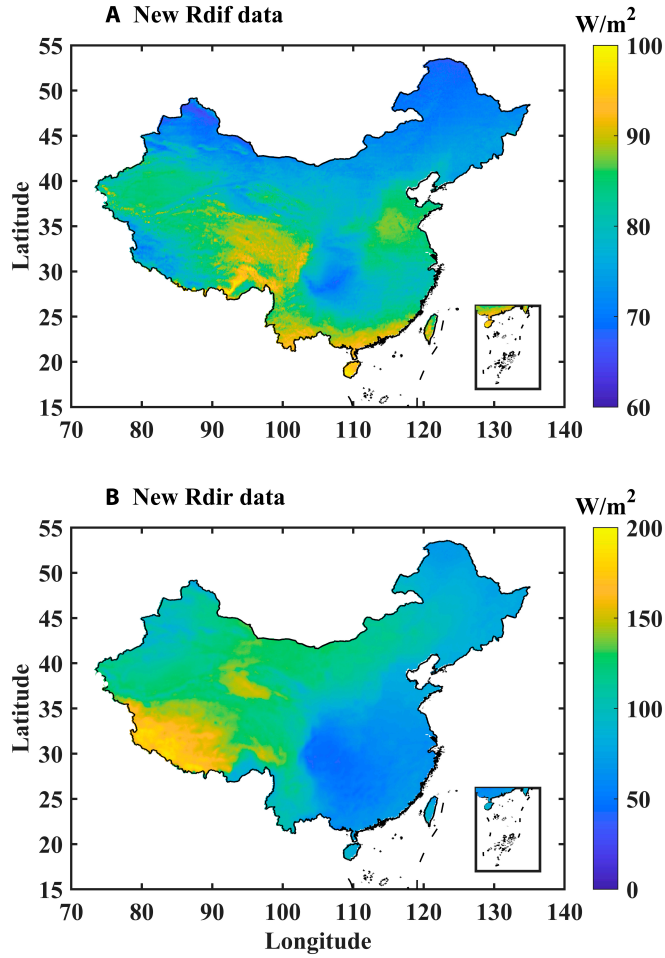

Fig. 8. Spatial pattern of multi-year average Rdif (A) and Rdir (B) from the new data during 1984 to 2018.

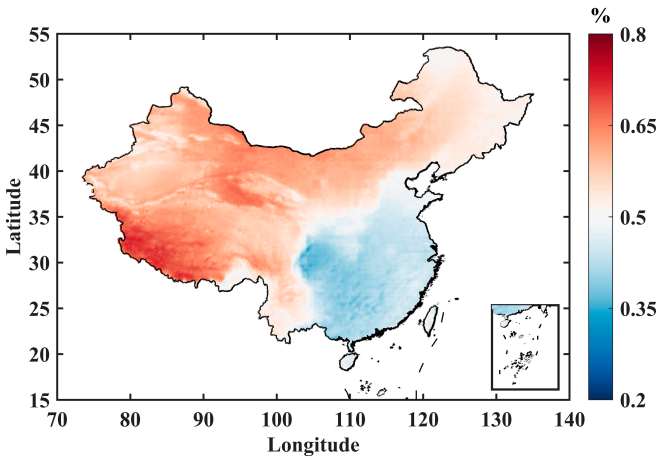

Fig. 9. Spatial pattern of multi-year average Rdir/Rs from the new data during 1984 to 2018.

In general, the intensity and spatial distribution of Rdif and Rdir can guide researchers on how to fully use solar energy in China. Furthermore, the evaluations at globally distributed BSRN stations also reveal the potential of the model to work on a global scale. With support from proper high-quality global meteorological input data, such as ERA5-Land [42] and Global Land Data Assimilation System [51], it is possible to apply this model to construct high-quality spatial-continuous global Rdif and Rdir data vital for global solar energy applications.

Conclusion

Radiation component data can be estimated from satellite data, but existing estimation methods usually need observed ground truth to fit or train the model, which cannot be met by the scarce observations in China. Given that SunDu-derived radiation component data are far more abundant and temporally homogeneous than observed data, they are used to train a data-augmented machine learning model for Rdir and Rdif estimation, without reference to observed ground truth.

The data-augmented model was effective at independent CMA radiation stations, with an RMSE of 20.0 W/m^2 (26.3%) and 19.8 W/m^2 (22.2%) for Rdif and Rdir, respectively. As an additional test, the trained model was validated at globally distributed independent BSRN stations, and the estimation shows high agreement with observations, with an RMSE of 23.0 W/m^2 (30.2%) and 23.2 W/m^2 (27.1%) for Rdif and Rdir, respectively. The additional test not only demonstrates the robustness and generality of the model, but also reveals the potential of the model to work on a global coverage. It is possible to extend the application of this model on a global scale with proper input data.

The trained model was applied to construct a 10-km gridded Rdif and Rdir dataset during 1984 to 2018 over China, with a state-of-the-art satellite-based Rs dataset (ISCCP-ITP-CNN) and CMFD meteorological dataset as input, and the dataset is much more accurate than other machine-learning-based, reanalysis, or satellite retrieved products, with an RMSE of 20.6W/m^2 (26.1%) and 31.8 W/m^2 (36.2%) for Rdif and Rdir, respectively, demonstrating the superiority of our data-augmented methods. The intensity and spatial distribution are important for the selection and deployment of different solar energy modules. CSP modules may be suitable for areas such as the southwestern Tibetan Plateau, where Rdir is dominant and high. Bifacial photovoltaic panels can be deployed in areas with comparably high Rdir and Rdif, such as the eastern Tibetan Plateau and northwestern China. This study can provide guidance on how to fully use solar energy in China.

Acknowledgments

The authors would like to thank the CMA for providing the surface solar radiation and meteorological observations data and the BSRN observation teams for their maintenance work.
Funding: This work was supported by the Sustainable Development International Cooperation Program of National Science Foundation of China (Grant No. 42361144875) and the National Natural Science Foundation of China (Grant No. 42171360).
Author contributions: C.S.: Methodology, software, investigation, and writing original draft. K.Y.: Conceptualization, supervision, review, and editing. Y.J.: Software, validation, and review. Y.H.: Software, data curation, and review. W.T.: Software, resources, data curation, and review. H.L.: Supervision and review. Y.L.: Supervision and review.
Competing interests: The authors declare that they have no competing interests.

Data Availability

The data used in this study are publicly available. The ERA5 reanalysis data were downloaded from https://cds.climate.copernicus.eu/ (last access: 2022 June 5). The CERES-SYN data were downloaded from https://ceres.larc.nasa.gov/data/ (last access: 2022 June 4).

References

1. Sweerts B, Pfenninger S, Yang S, Folini D, Van Der Zwaan B, Wild M. Estimation of losses in solar energy production from air pollution in China since 1960 using surface radiation data. *Nat Energy*. 2019;4(8):657–663.
2. Heusinger J, Broadbent AM, Sailor DJ, Georgescu M. Introduction, evaluation and application of an energy balance model for photovoltaic modules. *Sol Energy*. 2020;195:382–395.
3. Karakoti I, Pande B, Pandey K. Evaluation of different diffuse radiation models for Indian stations and predicting the best fit model. *Renew Sust Energ Rev*. 2011;15(5):2378–2384.
4. Mellit A, Eleuch H, Benghanem M, Elaoun C, Pavan AM. An adaptive model for predicting of global, direct and diffuse hourly solar irradiance. *Energy Convers Manag*. 2010;51(4):771–782.
5. Tang W, Yang K, Qin J, Min M, Niu X. First effort for constructing a direct solar radiation data set in China for solar energy applications. *J Geophys Res Atmos*. 2018;123(3):1724–1734.
6. Boland J, Huang J, Ridley B. Decomposing global solar radiation into its direct and diffuse components. *Renew Sust Energ Rev*. 2013;28:749–756.
7. Rodríguez-Gallegos CD, Bieri M, Gandhi O, Singh JP, Reindl T, Panda SK. Monofacial vs bifacial Si-based PV modules: Which one is more cost-effective? *Sol Energy*. 2018;176:412–438.
8. Pelaez SA, Deline C, Macalpine SM, Marion B, Stein JS, Kostuk RK. Comparison of bifacial solar irradiance

model predictions with field validation. *IEEE J Photovolt*. 2019;9(1):82–88.
9. Han J, Chang H. Development and opportunities of clean energy in China. *Appl Sci*. 2022;12(9):4783.
10. Jiang H, Lu N, Qin J, Yao L. Hourly 5-km surface total and diffuse solar radiation in China, 2007–2018. *Sci Data*. 2020;7(1):311.
11. Feng L, Lin A, Wang L, Qin W, Gong W. Evaluation of sunshine-based models for predicting diffuse solar radiation in China. *Renew Sust Energ Rev*. 2018;94:168–182.
12. Gueymard CA, Ruiz-Arias JA. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. *Sol Energy*. 2016;128:1–30.
13. Tapakis R, Michaelides S, Charalambides AG. Computations of diffuse fraction of global irradiance: Part 1—Analytical modelling. *Sol Energy*. 2016;139:711–722.
14. 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. *Renew Sust Energ Rev*. 2022;159:112195.
15. Furlan C, de Oliveira AP, Soares J, Codato G, Escobedo JF. The role of clouds in improving the regression model for hourly values of diffuse solar radiation. *Appl Energy*. 2012;92:240–254.
16. Gueymard CA. Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. *Sol Energy*. 2012;86(8): 2145–2169.
17. Wang L, Lu Y, Zou L, Feng L, Wei J, Qin W, Niu Z. Prediction of diffuse solar radiation based on multiple variables in China. *Renew Sust Energ Rev*. 2019;103:151–216.
18. Zhu T, Li J, He L, Wu D, Tong X, Mu Q, Yu Q. The improvement and comparison of diffuse radiation models in different climatic zones of China. *Atmos Res*. 2021;254:105505.
19. He Y, Wang K. Variability in direct and diffuse solar radiation across China from 1958 to 2017. *Geophys Res Lett*. 2020;47(1):e84570.
20. Qiu T, Wang L, Lu Y, Zhang M, Qin W, Wang S, Wang L. Potential assessment of photovoltaic power generation in China. *Renew Sust Energ Rev*. 2022;154:111900.
21. Qin W, Wang L, Gueymard CA, Bilal M, Lin A, Wei J, Zhang M, Yang X. Constructing a gridded direct normal irradiance dataset in China during 1981–2014. *Renew Sust Energ Rev*. 2020;131:110004.
22. Qin J, Tang W, Yang K, Lu N, Niu X, Liang S. An efficient physically based parameterization to derive surface solar irradiance based on satellite atmospheric products. *J Geophys Res Atmos*. 2015;120(10):4975–4988.
23. Stengel M, Stapelberg S, Sus O, Finkensieper S, Würzler B, Philipp D, Hollmann R, Poulsen C, Christensen M, McGarragh G. Cloud_cci advanced very high resolution radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties. *Earth Syst Sci Data*. 2020;12(1):41–60.
24. 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(4):1905–1915.
25. 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 Remote Sens*. 2020;58(8):5304–5316.
26. Yang D, Wang W, Xia X. A concise overview on solar resource assessment and forecasting. *Adv Atmos Sci*. 2022;39(8):1239–1251.
27. Wu J, Fang H, Qin W, Wang L, Song Y, Su X, Zhang Y. Constructing high-resolution (10 km) daily diffuse solar radiation dataset across China during 1982–2020 through ensemble model. *Remote Sens*. 2022;14(15):3695.
28. Laguarda A, Giacosa G, Alonso-Suárez R, Abal G. Performance of the site-adapted CAMS database and locally adjusted cloud index models for estimating global solar horizontal irradiation over the Pampa Húmeda. *Sol Energy*. 2020;199:295–307.
29. Li Z, Li C, Chen H, Tsay SC, Holben B, Huang J, Li B, Maring H, Qian Y, Shi G, et al. East Asian studies of tropospheric aerosols and their impact on regional climate (EAST-AIRC): An overview. *J Geophys Res*. 2011;116(D7).
30. Li B, Hou Y, Che W. Data augmentation approaches in natural language processing: A survey. *AI Open*. 2022;3:71–90.
31. Janjai S, Prathumsit J, Buntoung S, Wattan R, Pattarapanitchai S, Masiri I. Modeling the luminous efficacy of direct and diffuse solar radiation using information on cloud, aerosol and water vapor in the tropics. *Renew Energy*. 2014;66:111–117.
32. Shi G-Y, Hayasaka T, Ohmura A, Chen Z-H, Wang B, Zhao J-Q, Che HZ, Xu L. Data quality assessment and the long-term trend of ground solar radiation in China. *J Appl Meteorol Climatol*. 2008;47(4):1006–1016.
33. Tang WJ, Yang K, Qin J, Cheng CCK, He J. Solar radiation trend across China in recent decades: A revisit with quality-controlled data. *Atmos Chem Phys*. 2011;11(1):393–406.
34. Wang K. Measurement biases explain discrepancies between the observed and simulated decadal variability of surface incident solar radiation. *Sci Rep*. 2014;4:6144.
35. Wang K, Ma Q, Li Z, Wang J. Decadal variability of surface incident solar radiation over China: Observations, satellite retrievals, and reanalyses. *J Geophys Res Atmos*. 2015;120(13):6500–6514.
36. Tang W, Yang K, He J, Qin J. Quality control and estimation of global solar radiation in China. *Sol Energy*. 2010;84(3): 466–475.
37. Zhang X, Liang S, Zhou G, Wu H, Zhao X. Generating global land surface satellite incident shortwave radiation and photosynthetically active radiation products from multiple satellite data. *Remote Sens Environ*. 2014;152:318–332.
38. Driemel A, Augustine J, Behrens K, Colle S, Cox C, Cuevas-Agulló E, Denn FM, Duprat T, Fukuda M, Grobe H, et al. Baseline surface radiation network (BSRN): Structure and data description (1992–2017). *Earth Syst Sci Data*. 2018;10(3): 1491–1501.
39. Kilibarda M, Tadić MP, Hengl T, Luković J, Bajat B. Global geographic and feature space coverage of temperature data in the context of spatio-temporal interpolation. *Spat Stat*. 2015;14:22–38.
40. Shao C, Yang K, Tang W, He Y, Jiang Y, Lu H, Fu H, Zheng J. Convolutional neural network-based homogenization for constructing a long-term global surface solar radiation dataset. *Renew Sust Energ Rev*. 2022;169:112952.
41. He J, Yang K, Tang W, Lu H, Qin J, Chen Y, Li X. The first high-resolution meteorological forcing dataset for land process studies over China. *Sci Data*. 2020;7(1):25.
42. Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers

D, et al. The ERA5 global reanalysis. *Q J R Meteorol Soc*. 2020;146(730):1999–2049.
43. Kato S, Rose FG, Rutan DA, Thorsen TJ, Loeb NG, Doelling DR, Huang X, Smith WL, Su W, Ham SH. Surface irradiances of edition 4.0 clouds and the Earth's radiant energy system (CERES) energy balanced and filled (EBAF) data product. *J Clim*. 2018;31(11):4501–4527.
44. Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. *Glob Transit Proc*. 2022;3(1):91–99.
45. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. *Array*. 2022;16:100258.
46. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach (CA): Curran Associates Inc.; 2017. p. 3149–3157.
47. Duan S, Huang S, Bu W, Ge X, Chen H, Liu J, Luo J. LightGBM low-temperature prediction model based on LassoCV feature selection. *Math Probl Eng*. 2021;2021:1776805.
48. Pedregosa F, Gl V, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12(85):2825–2830.
49. Jiang H, Lu N, Huang G, Yao L, Qin J, Liu H. Spatial scale effects on retrieval accuracy of surface solar radiation using satellite data. *Appl Energy*. 2020;270:115178.
50. Pan B, Anderson GJ, Goncalves A, Lucas DD, Bonfils CJW, Lee J, Tian Y, Ma HY. Learning to correct climate projection biases. *J Adv Model Earth Syst*. 2021;13(10):e2021MS002509.
51. Rodell M, Houser PR, Jambor U, Gottschalck J, Mitchell K, Meng C-J, Arsenault K, Cosgrove B, Radakovich J, Bosilovich M, et al. The global land data assimilation system. *Bull Am Meteorol Soc*. 2004;85(3):381–394.