OPEN

DATA DESCRIPTOR

Hourly 5-km surface total and diffuse solar radiation in China, 2007–2018

Hou Jiang^{1,2}, Ning Lu^{1,3,4}✉, Jun Qin¹ & Ling Yao^{1,3,4}

Surface solar radiation is an indispensable parameter for numerical models, and the diffuse component contributes to the carbon uptake in ecosystems. We generated a 12-year (2007–2018) hourly dataset from Multi-functional Transport Satellite (MTSAT) satellite observations, including surface total solar radiation (R_s) and diffuse radiation (R_{dif}), with 5-km spatial resolution through deep learning techniques. The used deep network tackles the integration of spatial pattern and the simulation of complex radiation transfer by combining convolutional neural network and multi-layer perceptron. Validation against ground measurements shows the correlation coefficient, mean bias error and root mean square error are 0.94, 2.48 W/m² and 89.75 W/m² for hourly R_s and 0.85, 8.63 W/m² and 66.14 W/m² for hourly R_{dif} , respectively. The correlation coefficient of R_s and R_{dif} increases to 0.94 (0.96) and 0.89 (0.92) at daily (monthly) scales, respectively. The spatially continuous hourly maps accurately reflect regional differences and restore the diurnal cycles of solar radiation at fine resolution. This dataset can be valuable for studies on regional climate changes, terrestrial ecosystem simulations and photovoltaic applications.

Background & Summary

In recent years, research on quantitative estimation of surface total solar radiation (R_s) and diffuse solar radiation (R_{dif}) has attracted growing interest in view of its great scientific value and socioeconomic benefits¹. R_s is a prerequisite for modelling terrestrial ecosystem productivity², and an essential element for estimating heat fluxes, soil moisture and evapotranspiration³. The distribution and intensity of R_s are required for site selection of solar photovoltaic power and further estimation of power production⁴. Previous studies revealed that R_{dif} contributes to the ecosystem carbon uptake by increasing the canopy light use efficiency^{5–7}. The knowledge on R_{dif} is required to assess its impacts on plant productivity and carbon dynamics of terrestrial ecosystems^{7–10}. For instance, surface downward direct and diffuse radiation are necessary inputs for Forest Biomass, Assimilation, Allocation, and Respiration (FöBAAR) model to simulate forest carbon cycle¹¹. The perturbations of R_{dif} are required when using Yale Interactive terrestrial Biosphere (YIBs) model to study the response of global carbon cycle to fire pollutions¹². Besides, the fraction of diffuse and direct solar radiation as well as their variations are essential for modelling radiation-use efficiency of wheat during its vegetative phase¹³ and the early assessment of crop (i.e., soybean, wheat and sunflower) yield on a daily or shorter basis¹⁴.

Although great efforts have been made to establish globally covered surface-radiation networks, such as the Baseline Surface Radiation Network (BSRN), World Radiation Data Centre (WRDC) and Global Energy Balance Archive (GEBA), it remains insufficient to derive high-resolution radiation estimates from measurements alone because of the sparsity and heterogeneity of stations¹⁵. Since meteorological variables are commonly available and easily accessible, empirical models such as temperature-based, sunshine duration-based, as well as relative humidity- and cloud-based models are developed to extend R_s estimates to more meteorological stations^{2,16}, but their accuracies are strongly affected by measurements under insufficient calibration schedule¹⁷. Retrieval from satellite observations is the most reliable way to gain spatially continuous estimates of R_s as digital signals on sensors carry massive information about the atmospheric state and underlying land surface¹⁸. These algorithms

¹State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. ²College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100190, China. ³Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, 511458, China. ⁴Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, 210023, China. ✉e-mail: lvn@lreis.ac.cn

Fig. 1 Locations of radiation stations used in our study. Triangles with black cross are used for independent model validation while others are for model training or fine-tuning. The background land cover types are derived from MODIS MCD12Q1 products in 2008.

include two categories: constructing empirical relationships between top of atmosphere and surface radiative fluxes^{19,20}, and driving radiative transfer models by utilizing satellite-derived atmospheric parameters^{1,21}.

Several global R_s datasets have been generated through satellite retrievals. For instance, the Global Land Surface Satellite (GLASS)²² provides global 5-km resolution, 3-h interval R_s ; Tang *et al.*²³ produced a 16-year dataset (2000–2015) of high-resolution (3 h, 10 km) global R_s . Nevertheless, none of them provide estimate of R_{dif} . In addition, large uncertainties frequently occur under broken clouds due to the neglect of adjacency effect in their pixel-based retrieval schemes^{24–26} that depend on an assumption of plane-parallel homogeneous clouds. However, this assumption does not always hold. For example, in the presence of broken clouds, multiple reflections and scattering events off the sides of clouds lead to significant photon transport^{27–29}, which makes great difference at fine scales where R_s of an individual footprint under inhomogeneous clouds is relevant to multiple adjacent satellite pixels²⁴. Therefore, area-to-point retrievals seem the optimal solutions, i.e., adjacent signals within a certain extent are involved for radiation estimation.

The notable progress of deep learning in modelling spatial context opens new perspectives³⁰. Convolutional neural networks (CNN) have been widely utilized to extract spatial features from satellite images for definition and classification of extreme situations, for instance, storms, spiral hurricanes, and atmospheric rivers³¹. Thus, it is feasible to capture the spatial distribution of clouds/aerosols through CNNs for handling spatial adjacent effects caused by photon transport. In our previous work, a deep network consisting of CNN module and multi-layer perceptron (MLP) has been developed for R_s estimation for the first time³², and achieved breakthrough of data accuracy at hourly scale. In this study, we further extend the previous network to fit the requirements of R_{dif} estimation through transfer learning, and then use the newly trained network and previous one to generate high-resolution (hourly, 5 km) R_s and R_{dif} time series data in China. The final published dataset³³ includes R_s and R_{dif} at hourly, daily and monthly scales from 2007 to 2018. This unique data source are useful for analysis of regional characteristics and temporal cycles of solar radiation at fine scales, as well as radiation-related applications or scientific researches particularly climate changes and utilization of renewable solar energy.

Methods

Basic data. To train the proposed deep network, training samples should be prepared at first. The output corresponds to ground measurements of R_s or R_{dif} . The inputs include satellite image blocks and associated attributes of time (month, day, and hour) and location (latitude, longitude and altitude). Hourly R_s and R_{dif} measurements are available from China Meteorological Administration (CMA) (<http://data.cma.cn/> last accessed: 11 Jan. 2020). The used hourly records involve 98 radiation stations and cover a period from 1 Jan. 2007 to 31 Dec. 2008. The data in 2008 were used for training of deep network while that in 2007 were for independent validation. Figure 1 shows the spatial distribution of all related stations, of which 81 sites (circles) only provide R_s while the rest 17 sites (triangles) provide both R_s and R_{dif} . These stations locate in different climate zones and their background land cover types include forests, grasslands, croplands, bare lands etc., ensuring the representativeness of training samples for deep network. A simple physical threshold test³⁴ was adopted to exclude the spurious and erroneous measurements. In total, 0.49% of all records not passing the test were deleted and 441547 samples for R_s and

55096 samples for R_{dif} were retained for subsequent experiments. Besides, daily and monthly records of 98 radiation stations from 2007 to 2014 were used for validation of time-series products. Their quality was controlled based on the reconstructed daily and monthly integrated R_s data³⁵.

The used satellite images are Multi-functional Transport Satellites (MTSAT) data provided by the Japan Meteorological Agency (JMA). The MTSAT-1R, positioned at 104°E above the equator, scans the surface every 30 minutes and provides images over Asia-Pacific region (70°N–20°S, 70°E–160°E) in five channels: one visible channel (VIS, 0.55–0.80 μm), two split-window channels (IR1, 10.3–11.3 μm ; IR2, 11.5–12.5 μm), one water vapour channel (IR3, 6.5–7.0 μm) and one shortwave infrared channel (IR4, 3.5–4.0 μm). The original MTSAT-1R satellite data are resampled to so-called hourly GAME products with a resolution of 0.05°, which is freely accessible at <http://weather.is.kochi-u.ac.jp/> (last accessed: 11 Jan. 2020). We utilized the visible channel of GAME products to estimate target radiation, i.e., R_s and R_{dif} .

Finally, altitude of each pixel should be determined thus DEM data are required. DEM data are from Shuttle Radar Topography Mission that generates the most complete high-resolution digital topographic database of the Earth, covering over 80% of the Earth's land surface between 60°N and 56°S. The data can be obtained from the website <http://srtm.csi.cgiar.org/srtmdata/> (last accessed: 11 Jan. 2020). The original DEM data with data points posted approximately 30 m were resampled to grids with 0.05° resolution. DEM data provide elevation information for gridded inputs during spatially continuous estimation.

Estimation of surface solar radiation. The method we used to estimate surface solar radiation is mainly based on the CNN-based deep network developed in our previous work³². The network is demonstrated to be effective in handling spatial adjacent effects of surface radiation and simulating complicated radiative transfer processes and to be successful in retrieving accurate estimation of R_s from geostationary satellite data. In summary, the deep network consists of CNN module and MLP. CNN module takes image blocks as inputs thereby allowing identical treatment of adjacent satellite pixels, and is further stacked to construct deep residual structure to extract hierarchical features from low-level details (e.g., geometric shapes, sizes, orientations, edges and distribution) to high-level comprehensive abstract representations (e.g., intrinsic physical and optical properties of mixed clouds). Such hierarchical architecture of spatial features is a reflection of the scattering and absorption effects as well as their interactions in the atmosphere; hence, it can be considered as substitutes for input parameters in radiative transfer models to describe atmospheric state. The MLP is utilized to link extracted features of CNN and additional auxiliary information (involving the state in time and space) to target hourly R_s through implicit non-linear expressions, whose parameters are learnt from pre-prepared training samples in a supervised manner. Traditional physical algorithms retrieve surface radiation from satellite signals through various radiative transfer models or their simplified versions, where geometric/atmospheric conditions and aerosol types should be strictly defined, complex atmospheric processes need to be precisely simulated, and clear-sky and cloudy retrieval modes are independently developed. In contrast, all-sky situations are under a unified framework in our CNN-based algorithm and tedious intermediate simulations are avoided. Another advantage is that the deep network is capable of handling spatial adjacent effects of surface radiation, in other words, considering the influence of neighbouring pixels on radiation estimation of the central point. More details on the network structure and the spatial adjacent effects can refer to refs. ^{32,36}.

The graphical structure of the proposed deep network is illustrated in Fig. 2a. There are two input flows: *Input1* for satellite image blocks and *Input2* for additional attributes corresponding to the central point of *Input1*. The *Output* is target R_s associated with the central point of *Input1*. More details can refer to ref. ³². The input size for CNN is 16×16 pixels ($\sim 80 \times 80$ km on the ground) based on the recommendation that time series of satellite pixels are most correlative within an extent of approximately 60 km at hourly scale²⁵ and our previous experiments on the spatial scale effect of satellite-based R_s estimation³⁶. This setting also fits in the requirements of classical CNN structure and ensures the extraction of edge features. In addition, only visible band of satellite data is utilized for the convenience of cross-sensor applications because visible channel is available for nearly all satellite sensors. It is reasonable as visible channel provides the most proportion of information on aerosols, clouds and other atmospheric properties³⁰.

In our previous experiment, an outstanding deep network for R_s estimate has been obtained after continuous trial-and-error process and iterative parameter optimization. Herein, we further fine-tune the previous network for the sake of R_{dif} estimation using new training samples consisting of ground measured R_{dif} and corresponding satellite image block. The transfer learning was adopted to overcome the problem associated with insufficient R_{dif} samples. The parameters for convolutional layers (*Conv*) were initialized from the trained R_s model while that for fully-connected layers (*FC*) were reset to zero. Therefore, R_{dif} samples were mainly responsible for MLP fitting. Training and tuning processes were the same as R_s . In this way, the best model for R_{dif} estimation can be obtained in short time as CNN module has mastered the rules to abstract spatial pattern from satellite image blocks. After model learning and optimization, the trained R_{dif} model in combination with previous R_s model was used to generate our radiation datasets.

Workflow of data generation. The schematic flowchart to generate our radiation datasets is illustrated in Fig. 2b. The entire workflow consists of two main sections: training and estimation. The codes and datasets for training and estimation process can be accessed at the figshare³⁷ (<https://doi.org/10.6084/m9.figshare.c.4891302>). The training section concentrates on learning the underlying non-linear relationships between satellite images and measured surface radiation, and outputs two deep networks for R_s and R_{dif} estimation. The estimation section predicts spatially continuous R_s and R_{dif} data using the trained networks by feeding gridding inputs. The main procedures are numbered in Fig. 2b and described as follows:

1. Prepare training sets. For each ground station, a 16×16 neighbouring block was cut out from GAME image and matched up with quality-controlled R_s and R_{dif} record in 2008 according to time attributes. These samples were separated into three groups: R_s training set (93 training sites in Fig. 1), R_{dif} training set (12 triangle training sites in Fig. 1) and validation set (5 triangle validation sites with black cross in Fig. 1).
2. Simulate the state at the top of Mt. Everest. To guarantee a reasonable extrapolation of the deep network at high altitudes, constraints from radiative transfer model simulation at the top of Mt. Everest were mixed into the R_s and R_{dif} training set. The Santa Barbara DISORT Atmospheric Radiative Transfer (SBDART) model was adopted for the simulation²⁰.
3. Initialize the deep network. The network was implemented using *keras* package³⁸. All parameters of the network were initialized through Xavier^{39,8}. The learning rate was initially 0.01 but multiplied by 0.5 across a learning plateau.
4. Train deep network for R_s estimation. The Adagrad optimizer⁴⁰ was used to iteratively find the optimal weights and biases that minimize the mean-squared error between the network's predictions and the training targets. An early-stopping mechanism was utilized to relieve overfitting by relinquishing further optimization when the performance ceased to improve sufficiently. During training process, 20% of the paired samples were randomly selected to serve as a validation set to identify whether the network was overfitting. The model with the best performance was preserved for subsequent estimates.
5. Fine-tune the preserved model in 4) for R_{dif} estimation. Similarly, the model with the best performance was preserved.

More parameter configurations of step 2–5 can refer to ref. ³².

6. Generate spatially continuous hourly estimation. Hourly gridded GAME products from 2007 to 2018 were associated with corresponding time/location attributes, and then the best models in 4) and 5) were used to simultaneously obtain R_s and R_{dif} maps by feeding gridded inputs. In addition, surface direct solar radiation (R_{dir}) was derived by subtracting R_{dif} from R_s .
7. Integrate daily and monthly estimates. The missing hourly value was filled by multiplying the corresponding hourly extraterrestrial radiation by the averaged clearness index calculated from available hourly estimates within the day. After that, daily values were sums of all hourly estimates within the day, and monthly values were the sum of all daily values within the corresponding month.
8. Validate radiation datasets. The spatial extensibility of deep network was evaluated using the validation set in 1) that was not involved at training phase. The accuracy of our datasets was further evaluated at hourly scale by comparing to ground measurements in 2007. Moreover, daily and monthly estimates were evaluated using station records from 2007 to 2014. Three indices were used to quantify data quality: correlation coefficient (R), mean bias error (MBE), and root-mean-squared error (RMSE) between estimates and ground measurements:

4

Parameter	Search space	R	RMSE (MJ/m ²)	Time (minute)
Learning rate	Constant value of 0.01	0.858	0.486	
	Initialized as 0.01 and multiplied by 0.5 after 10 epochs' plateau of validation loss	0.879	0.447	
	Initialized as 0.01 and multiplied by 0.2 after 10 epochs' plateau of validation loss	0.876	0.449	
	Initialized as 0.01 and multiplied by 0.1 after 10 epochs' plateau of validation loss	0.510	0.871	
Optimizer	Mini-batch gradient descent (SGD)	0.879	0.447	
	Adaptive gradient (Adagrad)	0.886	0.419	
	Adaptive moment estimation (Adam)	0.807	0.548	
Dropout rate	0.3	0.886	0.419	
	0.1	0.881	0.420	
	0.5	0.885	0.419	
Batch size	500	0.886	0.419	25.3
	200	0.897	0.398	31.3
	1000	0.862	0.460	20.1

Table 1. Search for the optimal hyper-parameters of deep network. All experiments are conducted under the framework of *keras*³⁸.

$$\text{MBE} = \frac{1}{n} \sum_{i=0}^n (y'_i - y_i)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=0}^n (y'_i - y_i)^2}$$

where n is the total number of data samples indexed by i , y represents the measured value whose mean value is $\bar{y}$, and y' is the predicted value with mean $\bar{y}'$. Relative values of MBE and RMSE (rMBE and rRMSE) were also used.

Sensitivity analysis. The crucial step of this algorithm is to equip the deep network with the ability to extract abstract spatial pattern from satellite images. The representativeness and balance of training samples and the input size of satellite image blocks affect the reliability of gained pattern for R_s estimation, thus the accuracy of estimated data. The 98 stations under different climates and with diverse land cover types guarantee the representativeness of R_s training samples. To overcome the imbalance of samples, image blocks corresponding to high radiation values whose proportion is usually small were first rotated by 90/180/270 degrees and flipped up and down, left and right, then several copies of these samples were mixed into the full training set. The investigation of spatial scale effects in ref. ³⁶ suggests an optimal input size of 16×16 pixels.

Configurations of hyper-parameters were referenced to classical classification and object detection networks in computer vision, for example, the rectified linear unit (ReLU) was used as the activation function as it is effective in alleviating vanishing gradient problems and speeding up learning process; the early-stopping was adopted to prevent overfitting thus it was not necessary to control training epochs carefully. Other sensitive hyper-parameters (listed in Table 1) were determined based on a hierarchical search. To reduce the computational cost associated with the learning procedure of deep network, our experiments were conducted using a small training dataset (twelve training sites with blue triangles in Fig. 1). We first investigated different choices of the learning rate with a fixed configuration for other parameters (the first choice in the search space). After the optimal choice of learning rate (Initial value of 0.01 and multiplied by 0.5 after 10 epochs' plateau of validation loss) was determined, we continued searching for the optimizer, then the dropout rate and batch size. For learning rate, optimizer and dropout rate, the choice (the bold one in the search space) with the best validation accuracy at the five independent stations in terms of R and RMSE was finally selected. With respect to the batch size, it seems that the smaller size, the better performance but the longer time. Therefore, we chose the intermediate size of 500 for a balance between the performance and time consumption.

Data Records

All hourly, daily and monthly radiation datasets from 2007 to 2018 are freely available from the Pangaea³³ at <https://doi.org/10.1594/PANGAEA.904136>, through which users can link to the specific data entities of each year. The dataset for one year includes twelve folders for hourly radiation (twelve months), one folder for daily total radiation, one folder for monthly total radiation as well as other supporting documents:

- Hourly radiation: twelve zipped folders named as “China_HourlyRadiation_YYYYMM.h5”. The hourly files are named as “RAD_YYYYMMDDHH.h5” and stored as int16 data type in HDF5 format in the unit of 10^{-4} MJ m⁻². “YYYY”, “MM”, “DD”, and “HH” denote year, month, day and hour (UTC time). Each file contains two variables representing R_s and R_{dif} , namely global radiation and diffuse radiation, respectively. The time coverage of hourly dataset is from 2007-01-01 0:00 to 2018-12-31 23:00 (UTC).

Fig. 3 An example from our dataset. (a–c) Reflectance of the visible channel, hourly R_s and hourly R_{dif} at UTC 6:00, 22 Jun. 2018 (BJT 14:00, 22 Jun. 2018). (d–f) Daily R_s , R_{dif} and R_{dir} on 22 Jun. 2018. (g–i) Monthly R_s , R_{dif} and R_{dir} in June 2018. The unit of radiation is MJ m^{-2} .

- Daily and monthly radiation: Daily files are named as “RAD_YYYYMMDD.h5” and monthly files are named as “RAD_YYYYMM.h5” where “YYYY”, “MM”, and “DD” denote year, month, and day. Values are stored as float-point data type in the unit of $10^{-2} \text{ MJ m}^{-2}$. Each file contains two variables representing R_s and R_{dif} , namely daily/monthly total global radiation and daily/monthly total diffuse radiation, respectively.

The datasets provide gridded radiation estimates within 71°E – 141°E and 15°N – 60°N with an increment of 0.05° (about 5 km). The hourly radiation can also be expressed in unit of W m^{-2} through the conversion: $0.01 \text{ MJ m}^{-2} \text{ hour}^{-1} = 1/0.36 \text{ W m}^{-2}$. More details and examples of data visualization can refer to the published description files in each dataset. It is stressed that all hourly data are provided in UTC time.

Technical Validation

Spatial mapping. Figure 3a shows the instantaneous atmospheric state in visible channel captured by MTSAT at UTC 6:00, 22 Jun. 2008 (BJT 14:00, 22 Jun. 2008). The estimated hourly R_s and R_{dif} are displayed in Fig. 3b,c, respectively. The influence of cloud depth, surface topography and elevation are reflected in the spatial distribution of surface radiation. Under the thick clouds (red regions in Fig. 3a), both R_s and R_{dif} are lower than surrounding areas. In contrast, with respect to regions below thin clouds (yellow regions in Fig. 3a), R_s is relatively higher as more R_{dif} is obtained on the surface. For areas under clear sky conditions (blue regions in Fig. 3a), R_s is larger in high altitude areas (e.g., the Tibetan Plateau). Figure 3d–i illustrates the spatial distribution of R_s , R_{dif} and R_{dir} at daily and monthly scales. Daily radiation on 22 Jun 2008 shares similar characteristics with hourly radiation, indicating a stable atmospheric state in the day. At monthly scale, regional differences are revealed thoroughly. The distribution of solar radiation exhibits obvious latitudinal dependency, but also affected by the surface topography, regional climate and distance to coastal line. In June, R_s is highest on the Tibetan Plateau and lowest in the Szechwan and south China due to the significant difference of R_{dif} . Conversely, R_{dif} has the minimum value on the Tibetan Plateau while the maximum value locates on the North China Plain. R_{dir} is predominant in regions with high altitudes (the Tibetan Plateau) or drought climate zones (the Mongolia Plateau) while R_{dif} occupies the main proportion for areas with abundant rainfall or frequent cloud coverage (the middle and lower reaches of the Yangtze River, the Szechwan Basin and Guizhou). Although deep networks used for estimation are trained by samples within China, they also provide reasonable estimation in surrounding areas. For example, in June R_{dif} contributes to the majority of surface radiation in India and Southeast Asia due to the coming rainy season.

Temporal variations. We establish time series products to observe the temporal variations of surface solar radiation. Figure 4 shows the monthly variations of statistically averaged R_s , R_{dif} and R_{dir} for different regions in China from 2007 to 2018. R_s on the Qinghai–Tibet Plateau is the highest all the year round, benefiting from

Fig. 4 Monthly variations in different regions from 2007 to 2018. **(a)** R_s , **(b)** R_{dir} and **(c)** R_{dif} . The values are from the monthly products. Different regions are as shown on the right.

significantly higher altitudes, which in contrast leads to the lowest received R_{dif} as shown in Fig. 4c. The proportion of R_{dif} exhibits the highest in the south of China (relatively lower R_s but higher R_{dif}) compared with other regions due to the frequent cloudy and rainy weather. A slight dimming of R_s is observed in 2010, followed by the brightening from 2011 to 2015, and then by a dimming from 2016 to 2017. However the long-term trends of R_{dif} are inconsistent with the variations of R_s . For instance, neither obvious brightening nor dimming is manifested in the northwest while a decreasing tendency continues until 2015 on the Qinghai-Tibet Plateau. The fluctuation of R_{dir} is more obvious than R_{dif} accounting for the overall variations of R_s , because both absorption and scattering of the atmosphere lead to decrease of R_{dir} while changes of R_{dif} radiation result from scattering of the atmosphere alone.

Validation against ground measurements. The validation in our previous work³² has demonstrated the outstanding performance of the hybrid deep network on estimation of R_s . Herein, we evaluate the model performance for R_{dif} estimation to check the viability of transfer learning. The evaluation process includes three stages: performance over training samples (12 triangle training sites in Fig. 1), independent spatial extensibility in 2008 (5 triangle validation sites with black cross in Fig. 1), and temporal extensibility in 2007 at all 17 stations, as shown in Fig. 5a–c. Overall, it provides good estimates for R_{dif} at the site scale with an R of 0.88, MBE of 3.09 W/m² and RMSE of 58.22 W/m² over training samples. The results with an R of 0.89, MBE of 9.09 W/m² and RMSE of 58.33 W/m² at five independent validation sites, and an R of 0.85, MBE of 8.63 W/m² and RMSE of 66.14 W/m² in 2007, are comparable to the training phase, revealing the powerful spatial and temporal extensibility of deep networks in estimating R_{dif} . The positive MBE values confirm that our datasets overestimate R_{dif} at some degree, which might attribute to relative lower measured values due to instrument drifting sensitivity and urbanization effects^{41,42}. In fact, it is a challenging task to estimate R_{dif} due to much higher demands for fully consideration of aerosols, clouds, and their interactions. Yet for all that, our estimates of R_{dif} (Fig. 5c) outperform the widely-used ERA5 reanalysis data released by European Centre for Medium-Range Weather Forecasts (ECMWF) which has an R of 0.85, negative MBE of 43.08 W/m² and RMSE of 96.93 W/m² when evaluated at the same CMA diffuse radiation stations in 2007⁴².

Furthermore, our datasets are evaluated against ground measurements collected at 98 CMA radiation stations from 2007 to 2014 at daily mean and monthly mean scales as shown in Fig. 5d–i. Our daily results of R_s at the spatial resolution of 5 km exhibit an R of 0.94, MBE of 3.61 W/m² and 30.65 W/m². The intrinsic difference between point nature of ground measurements and areal average of gridded radiation products usually takes part of the responsibility for above deviations²⁴. At a finer spatial resolution of 5 km the RMSE of our daily R_s is still superior to widely-used products such as the ISCCP-FD data at 2.5° resolution with an R of 0.89 and RMSE of 68.3 W/m² (see Section 3.1 of ref. ³⁶), the GEWEX-SRB data at 1° spatial resolution with an R of 0.91 and RMSE of 36.5 W/m² (see Section 4 of ref. ¹⁸), and recent ISCCP-HXG products at 10 km resolution with an R of 0.93 and RMSE of 32.4 W/m² (see Table 3 of ref. ²³) which were also validated against observations at the CMA radiation stations. At monthly scale, the R value increases to 0.96, 0.93 and 0.92 meanwhile RMSE decreases to 17.24, 19.55 and 11.48 W/m² for R_s , R_{dir} and R_{dif} respectively, which is also remarkably better than other products (compare to Table 2 of ref. ³⁶). It should be pointed out that the excellent performance at monthly scale benefits from the mutual offset of underestimation and overestimation, for instance, daily R_{dif} shows an overestimation in the low-value part and an underestimation in the high-value part (Fig. 5f) while this does not occur for monthly R_{dif} (Fig. 5i).

Uncertainties. Figure 6a,b shows the errors of hourly estimates grouped by local hours from 8:00 to 17:00. All groups correlate well with the ground measurements with the lowest R being 0.96, 0.93 and 0.87 for R_s , R_{dir} and R_{dif} respectively, proving the good performance of deep network in hourly radiation estimation. Large

Fig. 5 Density plots of our estimates versus measured values. (a–c) Evaluation of R_{dif} at 12 training sites, at five independent validation sites in 2008 and at all 17 sites in 2007, respectively; (d–f) daily mean results of R_s , R_{dir} and R_{dif} from 2007 to 2014; (g–i) monthly mean results of R_s , R_{dir} and R_{dif} from 2007 to 2014. The black solid lines are 1:1 lines. The values in brackets indicate rMBE or rRMSE. Gaussian kernels are used for density plots and density values are normalized to 0–1 through min-max normalization.

rRMSEs are likely to appear in the morning and at night when the amounts of received surface radiation are very low. The data accuracy is acceptable with the average rRMSE lower than 20% (R_s) or 40% (R_{dir} and R_{dif}). It points out that temporal deviations might result from the fact that satellite images reflect an instantaneous state of the atmosphere whereas ground measurements represent an average state within per unit time (herein one hour). When clouds move rapidly, ground stations are likely to be covered by cloud shadows during a momentary period (less than one hour) but satellite sensor may scan a clear sky because clouds have drifted across. In this case, ground measurements would be smaller than satellite-based estimates. Therefore, large positive deviations usually occur when coming across changeable clouds. A limitation of our method is that it is unable to simulate dramatic changes in short time because our trained network just takes into consideration the spatial adjacent effects of solar radiation but ignores the lag effect and cumulative effect in time series. The recurrent neural networks^{43,44} that are able to model temporal dynamic behaviour are the promising solutions.

The R and rRMSE of hourly R_s and R_{dif} at each site are displayed in Fig. 6c–f, where obvious geographical differentiation is observed. Overall, our estimates correlate well with ground measurements at sites with high probability of cloud-free skies, for instance, the north and northwest China. Low R and large rRMSE are likely to occur at sites located in regions with cloudy days, such as the south and southwest China, especially the Szechwan Basin perennially covered by clouds. It is known that both dust aerosol particles in the north and northeast China and dense clouds in the south and southeast China lead to non-clear skies, but model performance is inconsistent in these areas. This phenomenon indicates that our developed deep network does well in simulating radiative effects of aerosols, but slightly bad in handling clouds. Such shortcoming is attributed to the inadequate information on cloud properties⁴⁵. As only visible channel of MTSAT satellite is used, it is difficult for deep network to

No.	Description	R	RMSE (MJ/m ²)
E1	Training the network using the fully R_s training dataset	0.933	0.325
E2	Training the network using R_s measurements at the twelve R_{dif} training sites	0.886	0.419
E3	Training the network using R_s measurements at randomly selected twelve training sites	0.717	0.655
E4	Training the network using R_{dif} measurements at the twelve R_{dif} training sites	0.796	0.323
E5	Fine-tuning the trained network of E1 using R_{dif} measurements at the twelve R_{dif} training sites	0.850	0.238
E6	Fine-tuning the trained network of E1 using K-fold cross-validation strategy. The 17 R_{dif} sites were divided into 4 groups (4-4-4-5), 3 out of which were used to train the network with the rest excluded. The training was repeated four times for all the combinations, and then all simulations of the sites excluded in the four repeats were put together to calculate R and RMSE to measure the performance.	0.712	0.260
E7	A stress test where we used for validation only the five sites that are more humid or with higher elevation, or closer to cities. Other sites were used to fine-tune the trained network of E1.	0.759, 0.584, 0.790	0.333, 0.451, 0.317

Table 2. Comparative experiments to check potential sampling errors. The three values for E7 represent results of the three extreme experiments on humidity, elevation and distance to cities, respectively.

determine optical properties of clouds, thus their interactions with radiation. With the aid of additional channels, better retrieval under cloudy conditions might be possible, for instance, depending on the identification of ice clouds and liquid water clouds whose radiative effects are significantly different. Moreover, the gridded products from satellite data are inherently spatial domain-averaged while ground measurements focus on solar radiation in local areas. This discrepancy always leads to evaluation deviations in space⁴⁶. When ground station is covered by shadows of cirrus clouds, the measured surface solar radiation would be lower than satellite-derived values because the footprint of satellite grid has larger spatial coverage.

With regard to R_{dif} , the correlation between our estimates and ground measurements is worse than that of R_s . Different from R_s , estimates of R_{dif} behave well in humid areas (southern China) rather than arid areas (north-west China), against our common sense that cloudy weather conditions in the southern China strongly affect the accuracy of radiation estimation. On the premise that deep network for R_s estimation has proved its effectiveness in arid areas, the worse performance on R_{dif} estimation under the same framework might be attributed to the poor data quality. Evidence comes from the fact that measurements of R_{dif} in the western China are not in a full-automatic tracking manner but manual operations, of which the nonstandard ones often lead to measurement errors. This contradictory phenomenon also indicates that a small proportion of problematic ground measurements would not affect the performance of deep network owing to its powerful robustness.

Sampling errors. The representativeness of R_{dif} training samples is worthy of special concern as only measurements at twelve stations are involved. To reduce the influence of insufficient samples on estimated data accuracy, we adopted the transferring learning approach to reuse the rules on how CNN extracts spatial pattern from satellite blocks that have mastered during R_s estimation based on a larger dataset. We designed 7 experiments (listed in Table 2) to have an in-depth inspection of potential sampling errors associated with this approach. E1 trains the deep network using the fully R_s training dataset. E2 trains the network using R_s measurements at the twelve R_{dif} training sites. E3 trains the network using R_s measurements at randomly selected twelve training sites. E4 trains the network using R_{dif} measurements at the twelve R_{dif} training sites. E5 fine-tunes the trained network of E1 using R_{dif} measurements at the twelve R_{dif} training sites. The performance of the gained network in E1–E5 is validated at the same five independent sites in terms of R and RMSE on R_s or R_{dif} . E6 fine-tunes the trained network of E1 through K-fold cross-validation strategy, i.e., the 17 R_{dif} sites were divided into 4 groups (4-4-4-5), and then 3 out of the 4 groups were used to train the network while the rest one was excluded. The training process was repeated four times for all the combinations and the R and RMSE of all predictions of the sites excluded in the four repeats were calculated to measure the performance of E6. E7 is a stress test where we used for validation only the five sites that are more humid or with higher elevation or closer to cities.

The results show that selecting densely and evenly distributed sites is the only way to improve the generalization ability of deep network (cf. E1 and E2), but it is also beneficial to make the limited sites distributed in representative areas with diverse characteristics (cf. E2 and E3, E5 and E6). Although the comparison is conducted on R_s , we assume it bears valid information for R_{dif} as well. Regardless of the small number, diffuse radiation stations cover all typical climate zones in China (Fig. 1), maximizing their spatial representation as much as possible; hence, it is rational to believe in the reliability of the trained network for R_{dif} estimation. Compared with training a network for R_{dif} estimation from the beginning (E4), fine-tuning the trained R_s network through transferring learning (E5) makes up the limitation caused by insufficient R_{dif} samples to a certain extent. Anyhow, the comparison between E1 and E6 demonstrates the existence of sampling errors and suggests that R_{dif} estimation requires further attempts and efforts. The stress test (E7) gave us an idea of the maximum sampling error. Since R_{dif} is highly influenced by humidity (function of climate and vegetation) and probably pollution and altitude, we pertinently removed sites that are more humid or with higher elevation, or closer to cities from training samples, but used them only for validation. Due to the inevitable reduction of the representativeness of training samples, the validation accuracy was lower than that of E5. These extreme cases show that the expected maximum sampling error of our R_{dif} estimates may not exceed the worst value of E7, i.e., R of 0.584 and RMSE of 0.451 MJ/m². Anyhow, such sampling errors announce the importance to collect more representative R_{dif} measurements for improving the performance of deep network on R_{dif} estimates.

Fig. 6 Uncertainties of hourly estimates in time and space. **(a,b)** The coefficient R and $rRMSE$ in 2007 for different hours from 8:00 to 17:00; **(c,d)** Spatial distribution of the coefficient R for R_s and R_{diff} respectively; **(e,f)** Spatial distribution of $rRMSE$ for R_s and R_{diff} respectively.

Usage Notes

Datasets can be reused as stand-alone for analysis of regional characteristics and temporal trend of solar radiation, yet richer studies and applications can be done by linking to other data resources. A simple direction is comparing this dataset to other products (e.g., ERA5⁴², BESS²⁶, GLASS²² etc.) to account for merits and demerits of different approaches for radiation estimation, or gain new understanding in typical regions (e.g., the Tibetan Plateau). We also suggest the open-source Global Solar Energy Estimator (GSEE) model⁴⁷ ([www.github.com/renewablesninja/gsee](https://github.com/renewablesninja/gsee)) for accurate estimation of solar energy in China to help policy-making of energy sector⁴⁸. If data on residential rooftop locations, electricity consumption and price, capital investment etc. are available, a comprehensive assessment of resource, technical, economic and market potential of rooftop solar photovoltaics⁴⁹ can be conducted based on our high-resolution (5 km) radiation dataset. Besides, there exists the possibility to drive plant models (e.g., JULES⁷, YIB⁵⁰, SWAP⁵¹ etc.) for crop yield estimation¹³.

Code availability

The MATLAB codes for spatial visualization of files in HDF format are published along with our datasets in PANGAEA. The codes and datasets for training and estimation process can be accessed at the figshare³⁷ (<https://doi.org/10.6084/m9.figshare.c.4891302>).

Received: 20 March 2020; Accepted: 27 August 2020;

Published online: 23 September 2020

References

- Greuell, W., Meirink, J. F. & Wang, P. Retrieval and validation of global, direct, and diffuse irradiance derived from SEVIRI satellite observations. *J. Geophys. Res.-Atmos.* **118**, 2340–2361 (2013).
- Jacovides, C. P., Tymvios, F., Assimakopoulos, V. D. & Kaltsounides, N. A. The dependence of global and diffuse PAR radiation components on sky conditions at Athens, Greece. *Agr. Forest Meteorol.* **143**, 277–287 (2007).
- Zhang, Y., Rossow, W., Lacis, A. & Oinas, V. Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data. *J. Geophys. Res.* **109**, D19105 (2004).
- Prävalie, R., Patriche, C. & Bandoc, G. Spatial assessment of solar energy potential at global scale: A geographical approach. *J. Clean. Prod.* **209**, 692–721 (2019).
- Alton, P., North, P. R. J. & Los, S. The impact of diffuse sunlight on canopy light-use efficiency, gross photosynthetic product and net ecosystem exchange in three forest biomes. *Global Change Biol.* **13**, 776–787 (2007).
- Kanniah, K., Beringer, J., North, P. R. J. & Hutley, L. Control of atmospheric particles on diffuse radiation and terrestrial plant productivity: A review. *Prog. Phys. Geog.* **36**, 210–238 (2012).
- Mercado, L. *et al.* Impact of changes in diffuse radiation on the global land carbon sink. *Nature* **458**, 1014–1017 (2009).
- Gu, L. *et al.* Advantages of diffuse radiation for terrestrial ecosystem productivity. *J. Geophys. Res.-Atmos.* **107**(ACL 2-1-ACL), 2–23 (2002).
- Zhang, M. *et al.* Effects of cloudiness change on net ecosystem exchange, light use efficiency, and water use efficiency in typical ecosystems of China. *Agr. Forest Meteorol.* **151**, 803–816 (2011).
- Zhang, Q. *et al.* Improving the ability of the photochemical reflectance index to track canopy light use efficiency through differentiating sunlit and shaded leaves. *Remote Sens. Environ.* **194**, 1–15 (2017).

11. Lee, M. *et al.* Model-based analysis of the impact of diffuse radiation on CO₂ exchange in a temperate deciduous forest. *Agr. Forest Meteorol.* **249**, 377–389 (2017).
12. Yue, X. & Unger, N. Fire air pollution reduces global terrestrial productivity. *Nat. Commun.* **9**, 5414 (2018).
13. Choudhury, B. A sensitivity analysis of the radiation use efficiency for gross photosynthesis and net carbon accumulation by wheat. *Agr. Forest Meteorol.* **101**, 217–234 (2000).
14. Holzman, M. E., Carmona, F., Rivas, R. & Niclòs, R. Early assessment of crop yield from remotely sensed water stress and solar radiation data. *ISPRS J. Photogramm.* **145**, 297–308 (2018).
15. Liang, S. *et al.* Estimation of incident photosynthetically active radiation from Moderate Resolution Imaging Spectrometer data. *J. Geophys. Res.* **111**, D15208 (2006).
16. Besharat, F., Dehghan, A. A. & Faghieh Khorasani, A. Empirical models for estimating global solar radiation: A review and case study. *Renew. Sust. Energ. Rev.* **21**, 798–821 (2013).
17. Dumas, A. *et al.* A new correlation between global solar energy radiation and daily temperature variations. *Sol. Energy* **116**, 117–124 (2015).
18. Qin, J. *et al.* An efficient physically based parameterization to derive surface solar irradiance based on satellite atmospheric products. *J. Geophys. Res.-Atmos.* **120**, 4975–4988 (2015).
19. Linares-Rodriguez, A., Ruiz-Arias, J., Pozo-Vazquez, D. & Tovar-Pescador, J. An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images. *Energy* **61**, 636–645 (2013).
20. Lu, N., Qin, J., Yang, K. & Sun, J. A simple and efficient algorithm to estimate daily global solar radiation from geostationary satellite data. *Energy* **36**, 3179–3188 (2011).
21. Huang, G., Mingguo, M., Liang, S., Shaomin, L. & Li, X. A LUT-based approach to estimate surface solar irradiance by combining MODIS and MTSAT data. *J. Geophys. Res.* **116**, D22201 (2011).
22. 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.* **152**, 318–332 (2014).
23. Tang, W., Yang, K., Qin, J., Li, X. & Niu, X. A 16-year dataset (2000–2015) of high-resolution (3 hour, 10 km) global surface solar radiation. *Earth Syst. Sci. Data* **11**, 1905–1915 (2019).
24. Huang, G. *et al.* Estimating surface solar irradiance from satellites: Past, present, and future perspectives. *Remote Sens. Environ.* **233**, 111371 (2019).
25. Deneke, H., Knap, W. & Simmer, C. Multiresolution analysis of the temporal variance and correlation of transmittance and reflectance of an atmospheric column. *J. Geophys. Res.* **114**, D17206 (2009).
26. Ryu, Y., Jiang, C., Kobayashi, H. & Detto, M. MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5 km resolution from 2000. *Remote Sens. Environ.* **204**, 812–825 (2017).
27. Madhavan, B. L., Deneke, H., Witthuhn, J. & Macke, A. Multiresolution analysis of the spatiotemporal variability in global radiation observed by a dense network of 99 pyranometers. *Atmos. Chem. Phys.* **17**, 3317–3338 (2017).
28. Oreopoulos, L., Marshak, A., Cahalan, R. & Wen, G. Cloud three-dimensional effects evidenced in Landsat spatial power spectra and autocorrelation functions. *J. Geophys. Res.-Atmos.* **105**, 14777–14788 (2000).
29. Schewski, M. & Macke, A. Correlation between domain averaged cloud properties, and solar radiative fluxes for three-dimensional inhomogeneous mixed phase clouds. *Meteorol. Z.* **12**, 293–299 (2003).
30. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
31. Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195 (2019).
32. 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. Sust. Energy Rev.* **114**, 109327 (2019).
33. Jiang, H. & Lu, N. High-resolution surface global solar radiation and the diffuse component dataset over China. PANGAEA <https://doi.org/10.1594/PANGAEA.904136> (2019).
34. Roebeling, R., Putten, E., Genovese, G. & Rosema, A. Application of Meteosat derived meteorological information for crop yield predictions in Europe. *Int. J. Remote Sens.* **25**, 5389–5401 (2004).
35. Zhang, X., Liang, S., Wild, M. & Jiang, B. Analysis of surface incident shortwave radiation from four satellite products. *Remote Sens. Environ.* **165**, 186–202 (2015).
36. 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. Energ.* **270**, 115178 (2020).
37. Jiang, H., Lu, N., Qin, J. & Yao, L. Hourly 5-km surface total and diffuse solar radiation in China, 2007–2018. *figshare* <https://doi.org/10.6084/m9.figshare.c.4891302> (2020).
38. Chollet, F. Keras, <https://github.com/fchollet/keras> (2015).
39. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **9**, 249–256 (2010).
40. Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011).
41. Wang, K., Ma, Q., Wang, X. & Wild, M. Urban impacts on mean and trend of surface incident solar radiation. *Geophys. Res. Lett.* **41**, 4664–4668 (2014).
42. Jiang, H., Yang, Y., Bai, Y. & Wang, H. Evaluation of the total, direct, and diffuse solar radiations from the ERA5 reanalysis data in China. *IEEE Geosci. Remote S.* **17**, 47–51 (2020).
43. Heck, J. & Salem, F. Simplified minimal gated unit variations for recurrent neural networks. in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, pp. 1593–1596 (2017).
44. Hochreiter, S. & Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
45. Tang, W. *et al.* Retrieving high-resolution surface solar radiation with cloud parameters derived by combining MODIS and MTSAT data. *Atmos. Chem. Phys.* **16**, 2543–2557 (2016).
46. Greuell, W. & Roebeling, R. Toward a standard procedure for validation of satellite-derived cloud liquid water path: A study with SEVIRI data. *J. Appl. Meteorol. Climatol.* **48**, 1575–1590 (2009).
47. Pfenninger, S. & Staffell, I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy* **114**, 1251–1265 (2016).
48. Sweerts, B. *et al.* Estimation of losses in solar energy production from air pollution in China since 1960 using surface radiation data. *Nat. Energy* **4**, 657–663 (2019).
49. Bódis, K., Kougias, I., Jäger-Waldau, A., Taylor, N. & Szabó, S. A high-resolution geospatial assessment of the rooftop solar photovoltaic potential in the European Union. *Renew. Sust. Energ. Rev.* **114**, 109309 (2019).
50. Yue, X. & Unger, N. The Yale Interactive terrestrial Biosphere model version 1.0: description, evaluation and implementation into NASA GISS Model E2. *Geosci. Model Dev.* **8**, 2399–2417 (2015).
51. Dam, J. C. *et al.* *Theory of SWAP, Version 2.0.* (Wageningen Agricultural University and DLO Winand Staring Center, 1997).

Acknowledgements

We are very grateful to the China Meteorological Administration for providing ground measurements of surface radiation data. The MTSAT satellite data were obtained from the Kochi University and SRTM DEM data were available from the U.S. Geological Survey. This work was supported by the National Natural Science Foundation of China (No.41971312 and 41771380), and the Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (GML2019ZD0301).

Author contributions

H.J. and N.L. developed the deep network and generated published datasets, J.Q. collected MTSAT data and CMA radiation measurements, L.Y. performed data pre-processing, H.J. wrote the manuscript. N.L. and J.Q. provided assistance in the organization of this article.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2020