remote sensing

MDPI

Article

Estimating Surface Downward Shortwave Radiation over China Based on the Gradient Boosting Decision Tree Method

Lu Yang [1], Xiaotong Zhang [1,*], Shunlin Liang [2], Yunjun Yao [1], Kun Jia [1] and Aolin Jia [1]

[1] State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; lyang201314@163.com (L.Y.); boyyunjun@163.com (Y.Y.); jiakun@bnu.edu.cn (K.J.); aolin@mail.bnu.edu.cn (A.J.)

[2] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA; sliang@umd.edu

* Correspondence: xtngzhang@bnu.edu.cn;

Received: 4 November 2017; Accepted: 22 January 2018; Published: 26 January 2018

Abstract: Downward shortwave radiation (DSR) is an essential parameter in the terrestrial radiation budget and a necessary input for models of land-surface processes. Although several radiation products using satellite observations have been released, coarse spatial resolution and low accuracy limited their application. It is important to develop robust and accurate retrieval methods with higher spatial resolution. Machine learning methods may be powerful candidates for estimating the DSR from remotely sensed data because of their ability to perform adaptive, nonlinear data fitting. In this study, the gradient boosting regression tree (GBRT) was employed to retrieve DSR measurements with the ground observation data in China collected from the China Meteorological Administration (CMA) Meteorological Information Center and the satellite observations from the Advanced Very High Resolution Radiometer (AVHRR) at a spatial resolution of 5 km. The validation results of the DSR estimates based on the GBRT method in China at a daily time scale for clear sky conditions show an R^2 value of 0.82 and a root mean square error (RMSE) value of 27.71 $W \cdot m^{-2}$ (38.38%). These values are 0.64 and 42.97 $W \cdot m^{-2}$ (34.57%), respectively, for cloudy sky conditions. The monthly DSR estimates were also evaluated using ground measurements. The monthly DSR estimates have an overall R^2 value of 0.92 and an RMSE of 15.40 $W \cdot m^{-2}$ (12.93%). Comparison of the DSR estimates with the reanalyzed and retrieved DSR measurements from satellite observations showed that the estimated DSR is reasonably accurate but has a higher spatial resolution. Moreover, the proposed GBRT method has good scalability and is easy to apply to other parameter inversion problems by changing the parameters and training data.

Keywords: downward shortwave radiation; machine learning; gradient boosting regression tree; AVHRR; CMA

1. Introduction

Downward shortwave radiation (DSR) is a key parameter in the land-atmosphere interaction, which largely controls human life and ecosystems due to its important role in energy cycles [1,2], the hydrological cycle [3,4], the carbon cycles [5,6], and solar energy utilizations [7–13]. Therefore, knowledge of DSR is essential for improving our understanding of the Earth's climate and potential climatic changes [14]. A number of gridded global DSR products exist from remote sensing, reanalysis, and general circulation models (GCMs). Satellite remote sensing is one of the most practical ways to derive DSR measurements with relatively higher spatial resolution and accuracy.

Currently, DSR data can be obtained in three ways. The first is through collection from ground measurements. This method is characterized by high precision and uneven geographic distribution.

The second is estimation from reanalysis data and simulations from GCMs, which have relatively low spatial resolution and accuracy [15–17]. Examples include the ERA-Interim provided by European Center for Medium-Range Weather Forecast (ECMWF), the Japanese 55-year Reanalysis (JRA-55) provided by Japan Meteorological Agency, and the Modern-Era Retrospective analysis for Research and Applications (MERRA) reanalysis dataset provided by NASA. The third way is retrieval from remote sensing data [18–20], which can provide spatio-temporal continuous DSR estimates with relatively higher precision. Commonly used remote sensing datasets of surface solar radiance include the Global Energy and Water Cycle Experiment-Surface Radiation Budget (GEWEX-SRB), the International Satellite Cloud Climatology Project-Flux Data (ISCCP-FD), the University of Maryland-Shortwave Radiation Budget (UMD-SRB), and the Earth's Radiant Energy System (CERES). Each type of DSR data from a different source has advantages and limitations: the ground measurements provide accurate but sparse spatial coverage, whereas products from the two other methods may have larger uncertainties. The ground measurements are always used to evaluate the other two types of DSR estimates. GCMs are widely believed to have an advantage in simulating global scale climate changes [21]. A reanalysis product is a combination of a model and measurements. It uses observations to constrain the dynamic model to optimize complete coverage and accuracy [22]. DSR retrievals from remote sensing data always have a relatively higher accuracy than those from reanalysis data and simulations from GCMs. These DSR products have been widely evaluated using ground measurements [23–26]. For example, Zhang et al. [26] evaluated four current representative existing remote sensing products using 1151 sites from the Global Energy Balance Archive and the China Meteorological Administration (CMA). The results implied that DSR estimates from remotely sensed data were more accurate than those acquired from reanalysis and simulations from GCMs. The maximum spatial resolution of these four products is 0.5°, and the temporal resolution is thrice-hourly. Although the current global radiation products have finer temporal resolution, they have lower spatial resolutions, which limit their application [27]. Therefore, it is still necessary to generate higher spatial resolution DSR estimates using satellite observation.

Several algorithms have been developed for retrieving DSR measurements. The first way is to estimate DSR based on statistical models [28–33]. Perez et al. [31] developed a simple solar radiation forecast model using sky cover predictions. Yang et al. [32] used a hybrid model with CMA routine data to estimate DSR, and the validation results of this proposed model against ground measurements collected in Tibetan Plateau were better than satellite estimations from existing satellite products. Wang et al. [33] used a statistical model to establish the relationship between top of atmosphere (TOA) reflectance and net surface shortwave radiation using multiple regression and revised methods, and they then compared the precision of these methods using various parameters. Empirical statistical models usually construct a regression model directly using observed data and measured DSR values. These models are easy to apply but are disadvantaged by their lack of universality; the relationship established in a particular atmospheric condition or region may not be applicable in another area. The second method to retrieve DSR measurements is to estimate them based on parametric physical modeling methods [34–39]. Li et al. [37] proposed a parameterized model in which the normalized net surface shortwave radiation flux of the top incident irradiance of the atmosphere was used to establish a parametric relationship with the planetary albedo. Qin et al. [38] used satellite atmospheric and land products—including ozone thickness, precipitable water, aerosol loading, cloud water path, clouds effective particle radius, cloud fraction, and ground surface albedo—to establish a physically based parameterization model. They then used the model to estimate surface solar irradiance with a mean RMSE of approximately 100 W/m^2 and 35 W/m^2 on an instantaneous and daily mean basis, respectively. López et al. [39] proposed a new, simple parametric physical model to estimate global solar radiance under cloudy sky conditions. These methods often construct a physical model by simulating direct interaction between solar radiation and the atmosphere. This requires many parameters (e.g., aerosol optical depth, surface albedo, and moisture). It is obvious that model accuracy depends on these parameters.

Machine learning methods, which learn the relationship between inputs and outputs by fitting a flexible model directly from the data, are some of the most widely used methods to estimate DSR [40–45]. Wang [43] proposed a method try to derive DSR measurements using Moderate Resolution Imaging Spectroradiometer (MODIS) data (e.g., atmospheric profile product and surface reflectance) based on an artificial neural network (ANN) model. The validation results against ground measurements showed that the maximum root mean square error (RMSE) was less than 45 $W \cdot m^{-2}$. Qin et al. [44] used an ANN-based method to establish the relationship between the measured monthly mean of daily global solar radiation levels and available remote sensing products with the aim of estimating global solar radiation. Zhou et al. [45] suggested that the Random Forest (RF) model was another feasible way to estimate DSR using satellite observations. These machine learning methods have their own advantages and disadvantages. For example, the attractiveness of an ANN is nonlinearity and high parallelism [46], and the RF cannot extrapolate beyond the training data and may not interpret well for conditions with few samples [47]. Although machine learning may provide powerful methods for estimating DSR from remote sensing data due to their ability to perform adaptive and nonlinear data fitting [48–50], the accuracy of the results is limited and many machine learning methods are prone to the phenomenon of overfitting. This can be avoid by using the gradient boosting regression tree (GBRT) method [51]. In addition, the GBRT can efficiently provide high accuracy. However, it has not been widely used for estimating DSR.

The objective of this study is to use a machine learning method, the GBRT, to obtain high accuracy DSR estimates from remote sensing and surface observed data under both clear and cloudy sky conditions in China. Moreover, this study aims to compare the DSR estimates from the GBRT with estimated values from classical ANN and existing remote sensing and reanalysis data.

The paper is organized as follows: Section 2 provides a brief introduction to the ground measurement and remote sensing data used. Section 2 also describes the methods used. Section 3 presents the results and an analysis. The conclusions are presented in Section 4.

2. Materials and Methods

2.1. Materials

2.1.1. Ground Measurements

The measurements of daily DSR used in this study were supplied by the CMA Meteorological Information Center. DSR was first measured in 1957, and its measurement was gradually collected at a total of 122 stations. However, the measurement at some stations have stopped sometime in the past. In 1994, there were 96 stations remaining to measure DSR. Quality control of the CMA DSR data was performed before the release; this included a spatial and temporal consistency check and manual inspection and correction [52]. Previous studies showed that the systematic errors in radiation measurements due to technical failure and operation-related problems are not rare [53,54]. Hence, a critical quality control procedure was performed to the ground measurements from the CMA before they were used in this study. The procedure is as described by Zhang [26]. Figure 1 shows the geographical distributions of the sites from the CMA. For more detailed information about the radiation data, it is possible to refer to the data description at the website http://data.cma.cn/.

This study used the daily DSR data collected from 96 radiation stations in China from 2001 to 2003. The daily DSR data from 2001 and 2002 were used to train the models, and the daily DSR data from 2003 was used to validate the model.

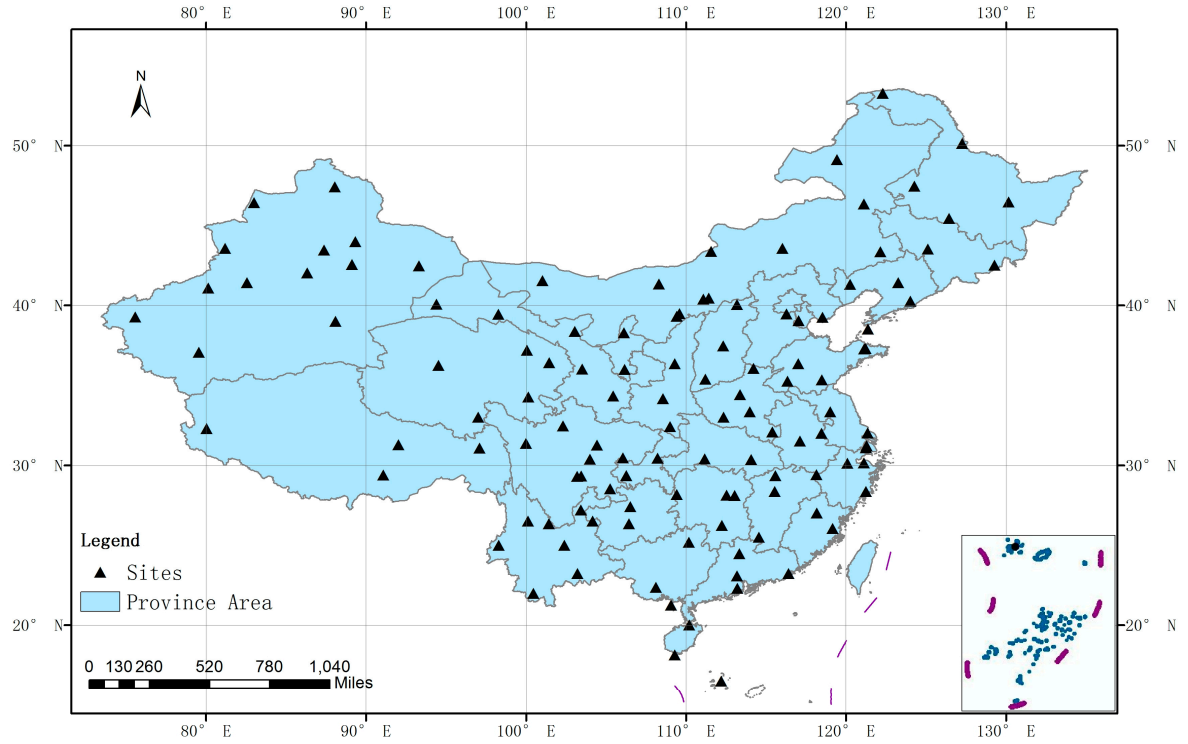

Figure 1. Spatial distribution of the radiation sites provided by the China Meteorological Administration (CMA) Meteorological Information Center.

2.1.2. Satellite Data

The National Oceanic and Atmospheric Administration (NOAA) Climate Data Records (CDR) of Visible and Near Infrared Reflectance from the Advanced Very High Resolution Radiometer (AVHRR) and the NASA Langley Research Center (LaRC) Cloud and Clear Sky Radiation Properties dataset were used in the paper. The two satellite datasets are from the Advanced Very High Resolution Radiometer (AVHRR) Global Area Coverage (GAC) Level 1B data, which has been quality controlled. These were taken from the NOAA-16 sun-synchronous orbit satellite observations provided by the NOAA CDR program. The NASA LaRC Cloud and Clear Sky Radiation Properties dataset is generated using the CERES Cloud Mask and Cloud Property Retrieval System (CCPRS) [55]. The NOAA CDR of Visible and Near Infrared Reflectance from AVHRR was calibrated by a multiple invariant Earth target calibration approach [56,57]. The NASA LaRC Cloud and Clear Sky Radiation Properties dataset was generated using algorithms initially designed for application to the Tropical Rainfall Measurement Mission (TRMM) and Moderate Resolution Imaging Spectroradiometer (MODIS) imagery within the NASA Clouds and the Earth's Radiant Energy System (CERES) program [58]. The spatial and temporal resolution of the dataset is about 4 km at the nadir and one day, respectively. Variables of the radiation properties dataset include cloud and clear sky pixel detection, cloud optical depth, cloud particle effective radius, land and sea surface temperature retrieval, shortwave broadband albedo, etc. [58]. Two variables including the calibrated 0.63 micron channel reflectance (channel 1) and the calibrated 0.86 micron channel reflectance (channel 2) were utilized for DSR estimation in this study [59]. Table 1 lists the corresponding information extracted from the AVHRR dataset used in this study.

Table 1. Input settings of the GBRT-based downward shortwave radiation (DSR) clear and cloudy sky models.

Inputs Data	Model	Unit	Range
Solar zenith angle	Clear and cloudy sky	Degrees	0–180
Viewing zenith angle	Clear and cloudy sky	Degrees	0–90
Relative azimuth angle	Clear and cloudy sky	Degrees	0–180
Top of atmosphere shortwave broadband albedo	Clear and cloudy sky	N/A	0–1.5
Reflectance of channel 1 and 2 of AVHRR	Clear and cloudy sky	Percent	0–12.5
Brightness temperature of channel 4 and 5 of AVHRR	Clear and cloudy sky	Degrees/Kelvins	160–340
Cloud optical depth	Cloudy sky	N/A	0–150
Cloud mask	Clear and cloudy sky	N/A	0–1

2.1.3. DSR Products

The two DSR products, the MERRA and the GEWEX-SRB DSR, were used in the paper. The MERRA product is a second reanalysis project from NASA for the satellite era (i.e., from 1979 to the present) using an updated new version of the Goddard Earth Observing System Data Assimilation System Version 5 (GEOS-5) [60]. The spatial resolution of the daily MERRA DSR estimate is $0.5^\circ \times 0.667^\circ$. The GEWEX-SRB radiation product from remotely sensed data used here was from the NASA/GEWEX-SRB shortwave version 3.0. The primary inputs to produce the data include shortwave and longwave radiances derived from International Satellite Cloud Climatology Project (ISCCP) pixel-level (DX) data, cloud and surface properties derived from the same source, temperature and moisture profiles, etc. [61]. The GEWEX-SRB DSR product was provided with a temporal resolution of 1 day and a spatial resolution of 1° from July 1983 to December 2008.

2.2. Methods

2.2.1. Gradient Boosting Regression Tree

The GBRT is a powerful, advanced statistical method widely used in classification and prediction. Because it does not require making assumptions on the data, it is extensively used in certain fields, such as in the optimization of recommendation systems [62,63], visual tracking algorithms [64], and traffic systems [65–68]. The attractiveness of GBRT comes from its ability to deal with the uneven distribution of data attributes, its lack of limitation for any hypothesis of input data, its better predictive capacity than a single decision tree, its power to deal with larger data size, and its transparency in terms of model development.

The GBRT produces competitive, highly robust, and interpretable procedures for both regression and classification. This was a method first proposed by Friedeman [51]. The core idea of this model is to generate a strong classifier by constructing an M amount of different weak classifiers through multiple iterations in order to reach the final combination. Each iteration is designed to improve the previous result by reducing the residuals of the previous model and establishing a new combination model in the gradient direction of the residual reduction. To describe the accuracy of the model, a loss function defined as $L(y, F)$ is introduced. The frequently employed loss functions include squared-error and absolute error [51]. Suppose that $\{x_i, y_i\}_{i=1}^{N}$ is the training sample. The x represents explanatory variables. The y represents the response variable. N is the number of the training sample. Let the M different individual decisions trees be represented by $\{h(x; \alpha_i)\}_{i=1}^{M}$, which is the parameterized function of the explanatory variables x and is characterized by $\alpha = \{\alpha_m\}_{m=1}^{M}$. β is the weight of each classifier, and α is the classifier parameter. Each tree divides the input space into the number of independent areas numbered J, as in $R_{1m}, \ldots, R_{jm}$. Each R_{jm} has a corresponding predicted value γ_{jm}. If the x-value is in the area R_{jm}, it means $x \in R_{jm}$ and the constant I equals 1. However, the constant

$I = 0$. Hence, the function ($F(x)$), which is an approximation function of the response variable. It can be written as follows:

$$\begin{cases} F(x) = \sum_{m=1}^{M} \beta_m h\,(x;\alpha_m) \\ h(x;\alpha_m) = \sum_{j=1}^{J} \gamma_{jm} I\,(x \in R_{jm}), where\ I = 1\ if\ x \in R_{jm}; I = 0,\ \text{otherwise} \end{cases} \quad (1)$$

The general process of GBRT shown in Figure 2 and more detail of GBRT can be find in Hastie et al. [69] and Ridgeway [70].

1. Initialize $F_0(x)$ to be a constant, $F_0(x)=\arg\min_{\rho}\sum_{i=1}^{N} L(y_i,\rho)$.
2. For m=1 to M do:
   For i =1 to N do :
3. Compute the negative gradient

$$\tilde{y}_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x-1)}$$

4. End;
5. Fit a regression tree $h(x;\alpha_m)$ to predict the targets $\tilde{y}_{im}$ from covariates x_i for all training data.
6. The α_m can be obtained as followed:

$$\alpha_m = \arg\min_{\alpha,\ \beta} \sum_{i=1}^{N} \left[\tilde{y}_{im} - \beta h(x_i;\alpha_m)\right]^2$$

7. Compute a gradient descent step size as:

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i;\alpha_m))$$

8. Update the model as :

$$F_m(x) = F_{m-1}(x) + \rho_m h(x_i;\alpha_m)$$

End

Output the final model $F_M(x)$

Figure 2. The main procedures of the gradient boosting regression tree (GBRT) method.

The GBRT model can be constructed in three steps: (1) the preparation of the training database, (2) the architecture design and training phase, and (3) the application of the GBRT method. The next step is then to divide the data into clear sky and cloudy sky conditions according to the NOAA CDR of cloud mask data. If the pixel was marked as "cloud" by AVHRR data, it means it is under cloudy conditions. Otherwise, there is clear sky conditions. The GBRT-based DSR clear and cloudy sky model were trained using cloud mask data provided by the AVHRR data.

The performance of the DSR estimates was tested using the holdout method, which is a simple type of cross-validation. The dataset was randomly stratified into two groups, with 80% made part of the training dataset and 20% made part of the testing dataset. The main procedures are as follows.

(1) Extracting the TOA radiance from the NOAA CDR of Visible and Near Infrared Reflectance from AVHRR;
(2) Extracting the cloud properties from the NASA LaRC Cloud and Clear Sky Radiation Properties dataset;
(3) Training the clear and cloudy sky models. The inputs of the clear sky model include the solar zenith angle, viewing zenith angle, relative Azimuth angle, TOA shortwave broadband albedo, reflectance (from channel 1 and 2) of AVHRR, and the brightness temperature (from channel 4 and 5) of AVHRR. The input of the cloudy sky model used the same input variables as the clear sky model and cloud optical depth;
(4) Configuring the model coefficients. The optimal parameterization scheme was determined by looping in each parameter threshold. Table 2 shows the parameter setting details to determine the optimal parameterization for both the clear sky and cloudy sky conditions through the evaluation results (highest R^2 value and lowest bias and RMSE values) of the testing dataset for each loop;
(5) Evaluating against the ground measurements.

Figure 3 shows the flowchart of the proposed GBRT model used in this study.

Table 2. Parameters setting to determine the optimal parameters for the GBRT model.

Parameters	Threshold	Intervals
The number of iterations	50–300	50
Shrinkage	0.1–1	0.3
The depth of the tree	6–9	1
Sampling rate	0.2–1	0.2

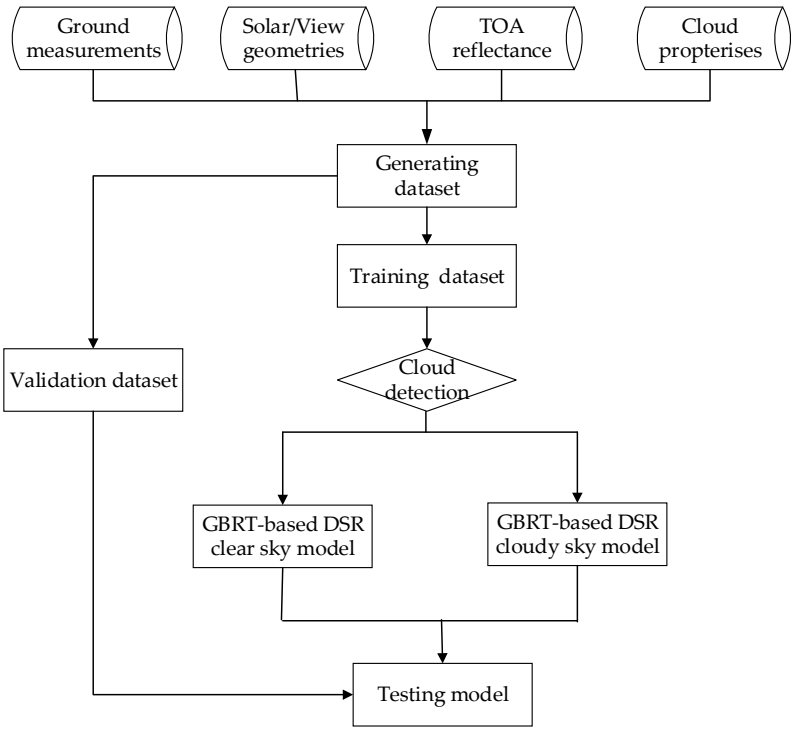

Figure 3. Flowchart of the GBRT method.

2.2.2. Artificial Neural Networks

ANNs are used as an empirical statistical method in a variety of applications such as classification, pattern recognition, forecasting, optimization, etc. [71–73]. An ANN model can be any model in which

the output variables are computed from the input variables using compositions or connections of basic functions. In this research, a feedforward backpropagation neural network consisting of several layers of neurons was used. A neuron is a simplified mathematical model of a biological neuron, and a connection is a unique information transport link from a sending to a receiving neuron. Figure 4 shows a structural diagram of the ANN used in this study. The ANN model used here consists of three layers of neurons: input layers, hidden layers, and an output layer. Input $\{x_j\}_{j=1}^{m}$ is transmitted through a connection that multiplies its strength by a weight represented by $\{w_{ij}\}_{i=1}^{k}$. This gives the value $x_i w_{ij}$, which is an argument to a transfer function f that yields an output y_i.

$$y_i = f(\sum_{j=1}^{m} w_{ij} x_j) \tag{2}$$

where i is the index of neuron in the hidden layer and j is the index of inputs to the neural network. A typical feedforward network trained with a resilient backpropagation algorithm [74,75] is employed to estimate DSR in this paper.

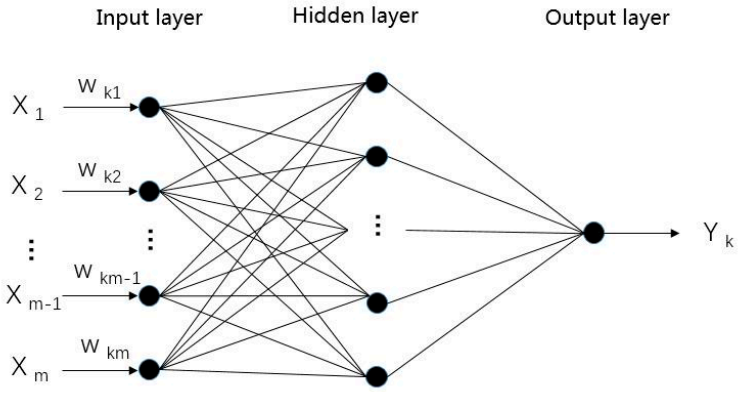

Figure 4. Artificial neural network (ANN) structure used in this study.

2.3. Constructing the Model

According to the characteristic variables in Table 1, corresponding data was extracted to establish the training dataset. Daily observed data from the CMA Meteorological Information Center from 2001 and 2002 were used as the response values (true values) of the training dataset. Information from the AVHRR cloudy and clear sky pixel detection was used to divide the training dataset into a cloudy sky training dataset and a clear sky training dataset. In addition, the missing values were removed both from the training and validation dataset.

2.3.1. Constructing the GBRT-Based DSR Model

The key step in building an efficient GBRT model is finding the optimal architecture. Building the GBRT model in a stage-wise fashion and regenerating the model minimizes the expected value of a certain loss function. After adding many trees to the model, the fitted model should have a small training error. However, it is important to remember that the generalization ability does not improve in direct proportion with the size of the fitted model; if the model is overfitted and possesses an extremely small error with the training dataset, its generalization ability will be poor. The performance of the GBRT model is influence by these four parameters as follows: the number of iterations, shrinkage, the depth of the tree, and the sampling rate [63]. As the number of iterations increases, model complexity will also increase, leading to poor prediction performance on the test dataset. Determining the appropriate number of iterations is essential to minimize future risks in prediction. Overfitting can be avoided by limiting the number of iterations and reducing the contribution of each tree. This

is also known as shrinkage (or learning rate). There is a tradeoff between the number of iterations and the learning rate. A lower learning rate value means that the model is more robust but has a slower computing speed. The size of each tree is called the depth of the tree. The depth of the tree refers to the number of nodes in a tree. This parameter depends on the number of data points and the characteristic variables of the data. In theory, if the value of this parameter is too large, the model will run at a slower rate. The sampling rate is the ratio of the subsample to the total number of training instances. When set to 0.5, it means that the model randomly collected half the data instances to grow trees. This will prevent overfitting. This procedure should be used with adjusting the learning rate and the number of iterations.

In the present case, successive performance testing showed that an architecture of 250 trees with a tree depth of 6, a sampling rate of 0.6, and a learning rate of 0.1 was optimal to estimate the DSR under clear sky conditions. These values are 250, 6, 0.8, and 0.1, respectively, under cloudy sky conditions.

Considering that cloud optical depth is related to DSR under cloudy sky conditions, the cloud optical depth was chosen as the input data for the cloudy sky model. This was different from the input data used for the clear sky model. Table 1 shows the input data of the GBRT-based DSR model under clear sky and cloudy sky conditions. The debugging procedure for key parameters such as the number of trees, the size of each tree, the learning rate, and the subsample ratio was described earlier.

2.3.2. Constructing the ANN-Based DSR Model

The ANN training databases in this study were the same as those used in the GBRT model. The architecture is mainly defined by the number of layers, the number of neurons in each layer, and the transition function associated with each neuron. As for other parameters (e.g., initial weighting), details of these will not be shown in this paper. In the present case, successive performance testing has shown that an architecture with one hidden layer is sufficient to estimate DSR. The number of nodes in the input layer was set to nine nodes, and the number of nodes in the output layer was set to one. After testing, the number of the nodes in the hidden layer was 12 under clear sky conditions and 14 under cloudy sky conditions. The transfer function of the hidden layer was a tan-sigmoid transfer function, and those of the other two layers were linear functions under both clear sky and cloudy sky conditions. Theoretically, various sets of functions such as step, linear, and no linear functions could be used as the transfer function of different layers. However, the tan-sigmoid (for the hidden layer) and linear (for the input and output layers) types were most commonly used in the literature [71].

3. Results and Analysis

The estimated daily and monthly mean DSR based on the GBRT method were not only evaluated against ground measurements but also compared with the evaluation results from those estimated from the ANN-based DSR model. Additionally, the estimated DSR values were also compared with current existing DSR products from the GEWEX-SRB and the MERRA. The validation results were shown in terms of bias, RMSE, and correlation coefficient (R^2).

3.1. Validation with Ground Measurements

3.1.1. Validation at a Daily Time Scale

The ground measurements at the selected 96 stations collected from CMA in 2003 were compared to the grid points of the estimated DSR based on the GBRT method. The performance of the GBRT-based DSR clear sky model using the training dataset and the validation dataset is shown in Figure 5. As shown in Figure 5, the daily estimated DSR correlates well with ground measurements under clear sky conditions. The daily DSR estimates under the clear sky conditions for the training dataset have an overall RMSE value of 19.05 $W \cdot m^{-2}$ (19.06%), a bias value of 0.00 $W \cdot m^{-2}$ (2.41%), and an R^2 value of 0.92. These values were 27.71 $W \cdot m^{-2}$ (38.38%), -2.53 $W \cdot m^{-2}$ (1.37%), and 0.82, respectively, for the validation dataset. The validation results at a daily time scale demonstrate that the GBRT is

a practically applicable and effective method for estimating DSR under clear sky conditions using satellite observations from AVHRR data.

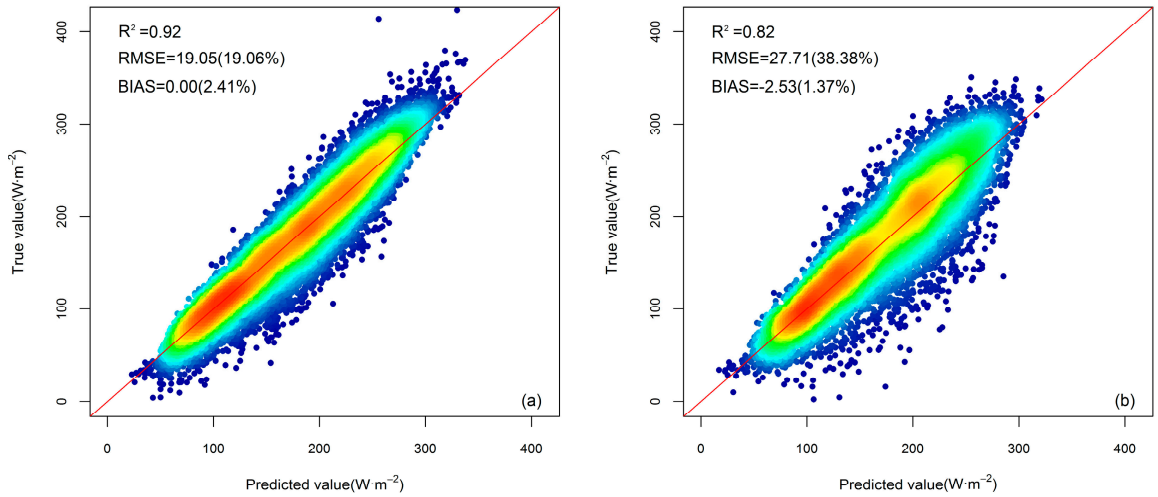

Figure 5. (**a**) Evaluation results of the training set's daily estimated DSR based on the GBRT-based clear sky model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation set's daily estimated DSR based on the GBRT-based clear sky model against ground measurements in 2003. The number in the parentheses is the percent bias or root mean square error (RMSE) value.

Figure 6 presents the evaluation results of the GBRT-based DSR cloudy sky model using the training dataset and the validation dataset. The daily DSR estimates for the training dataset under the cloudy sky conditions have an overall RMSE value of 33.37 $W \cdot m^{-2}$ (30.21%), an R^2 value of 0.79, and a bias value of 0.01 $W \cdot m^{-2}$ (4.74%). These values for the validation dataset were 42.97 $W \cdot m^{-2}$ (34.57%), 0.64, and -2.83 $W \cdot m^{-2}$ (1.45%), respectively. The accuracy was slightly lower than that of the clear sky model, which may be related to the influence of clouds [76].

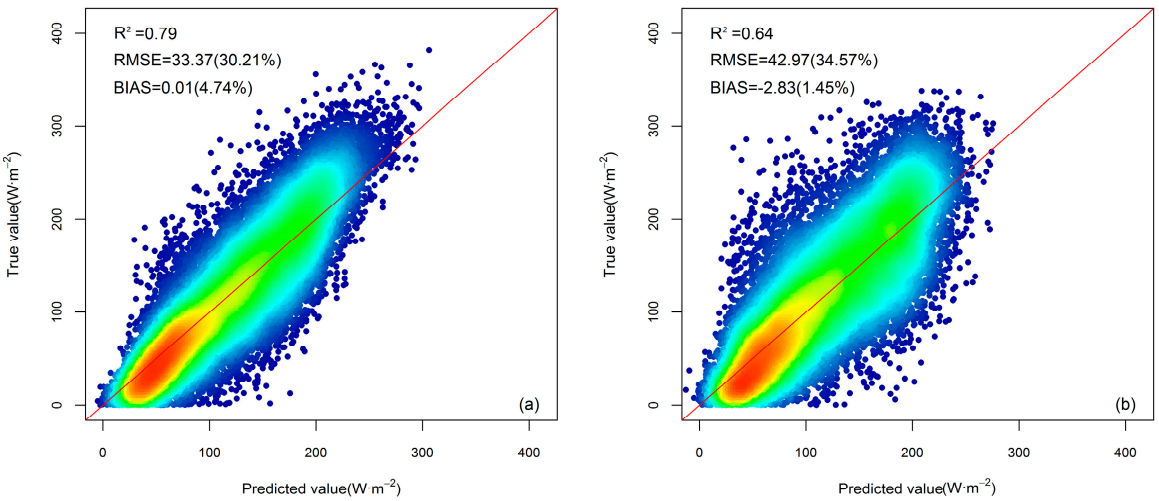

Figure 6. (**a**) Evaluation results of the training set's daily estimated DSR based on the GBRT-based cloudy sky model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation set's daily estimated DSR based on the GBRT-based cloudy sky model against ground measurements in 2003. The number in the parentheses is the percent bias or RMSE value.

When building the models for DSR estimation, we found that channel 4 and 5 influence the model accuracy. Figures 7 and 8 show a comparison of the evaluation results without considering

AVHRR channels 4 and 5 under clear and cloudy sky conditions, respectively. As shown in Figure 7, the daily DSR estimates without considering these two channels under clear sky conditions of the training dataset have an overall RMSE value of 26.52 W·m^{-2} (23.93%), a bias value of −0.26 W·m^{-2} (3.25%), and an R^2 value of 0.85. It can be concluded that the clear sky model yields higher accuracy if AVHRR channels 4 and 5 are considered. Similar results were also found under cloudy sky conditions. The daily DSR estimates without considering these two channels under cloudy sky conditions of the training dataset have an overall RMSE value of 37.52 W·m^{-2} (31.86%), a bias value of 0.16 W·m^{-2} (4.66%), and an R^2 value of 0.73. A potential reason for this may be the total atmospheric water vapor effect on DSR estimation, which may be to cause large uncertainties. Previous studies showed that AVHRR channels 4 and 5 have been widely used to retrieve the total atmospheric water vapor [77,78].

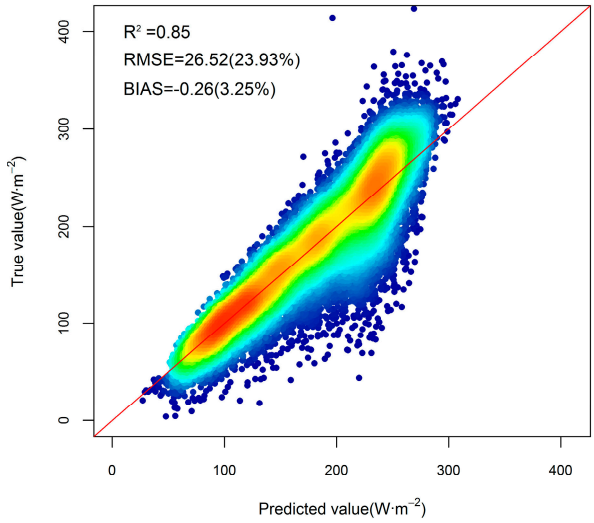

Figure 7. Validation results of the estimated daily DSR based on the GBRT model under clear sky conditions without considering Advanced Very High Resolution Radiometer (AVHRR) channels 4 and 5 as the input variables. The number in the parentheses is the percent bias or RMSE value.

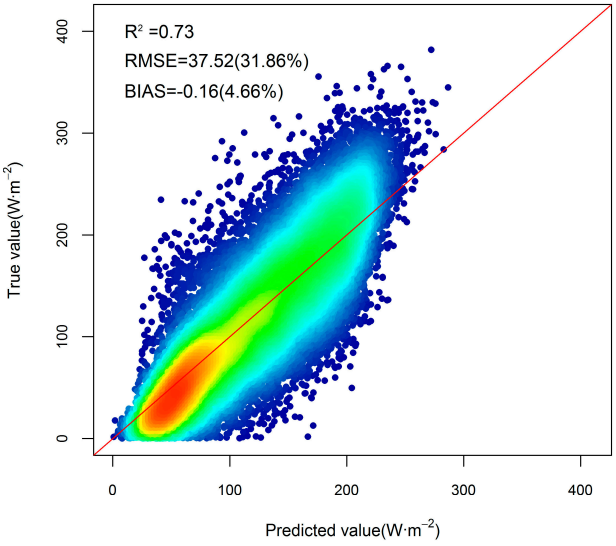

Figure 8. Validation results of the estimated daily DSR based on the GBRT model without considering AVHRR channels 4 and 5 as the input variables under cloudy sky conditions. The number in the parentheses is the percent bias or RMSE value.

3.1.2. Validation at a Monthly Time Scale

To further show the relative accuracy of the GBRT method, we also validated the estimated DSR at a monthly time scale. To perform the comparison, monthly DSR estimates were obtained by averaging the daily DSR data of each month. Figure 9 shows the evaluation results of the training dataset and validation dataset based on the GBRT model of 2003 at a monthly time scale. The monthly estimated DSR of the training dataset has an overall RMSE value of 14.22 W·m^{-2} (12.50%), a bias value of −0.30 W·m^{-2} (2.04%), and an R^2 value of 0.94. These values were 15.40 W·m^{-2} (12.93%), −2.25 W·m^{-2} (1.01%), and 0.92, respectively, for the validation dataset. Like the validation results at a daily time scale, the validation results at a monthly time scale showed that the GBRT model is reasonably accurate.

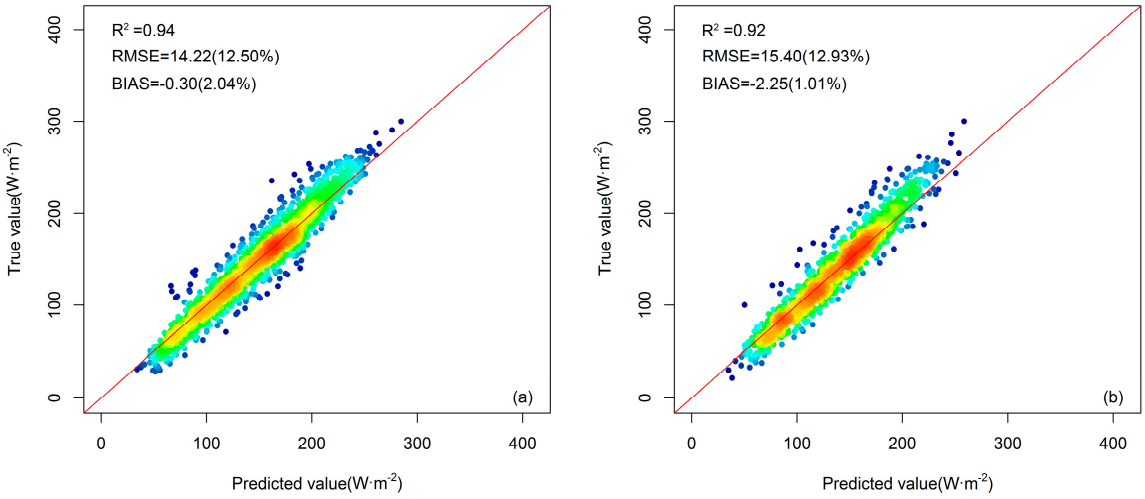

Figure 9. (**a**) Evaluation results of the training set's estimated monthly mean DSR based on the GBRT-based DSR model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation set's estimated monthly mean DSR based on the GBRT-based DSR model against ground measurements in 2003. The number in the parentheses is the percent bias or RMSE value.

3.2. Comparison with the ANN-Based Method

3.2.1. Validation at a Daily Time Scale

Figure 10a,b shows the evaluation results of the estimated daily DSR of the training and the validation dataset based on the ANN-based DSR model under clear sky and cloudy sky conditions. The daily DSR estimates based on the ANN-based clear sky model of the training dataset have an overall RMSE value of 26.53 W·m^{-2} (41.84%) and a bias value of −0.09 W·m^{-2} (0%). These values were 27.15 W·m^{-2} (46.07%) and −3.67 W·m^{-2} (1.60%), respectively, for the validation dataset. Although the RMSE of the estimated daily DSR of the validation dataset was slightly lower than that of the GBRT model, the mean absolute bias of the ANN-based model was 3.67 W·m^{-2} (1.60%), which is larger than that of the GBRT model (2.53 W·m^{-2} (1.37%)) (Table 3). The evaluation results of the ANN-based cloudy sky model are shown in Figure 11. The daily DSR estimates based on the training dataset's ANN cloudy sky DSR model have an overall RMSE value of 42.07 W·m^{-2} (33.99%) and a bias value of 0.17 W·m^{-2} (3.13%). These values were 42.39 W·m^{-2} (34.50%) and −4.35 W·m^{-2} (0.17%), respectively, for the validation dataset. According to the comparison results shown in Figures 10 and 11 and Table 3, it was clear that the predictive abilities of the GBRT model are better than the ANN model at a daily time scale.

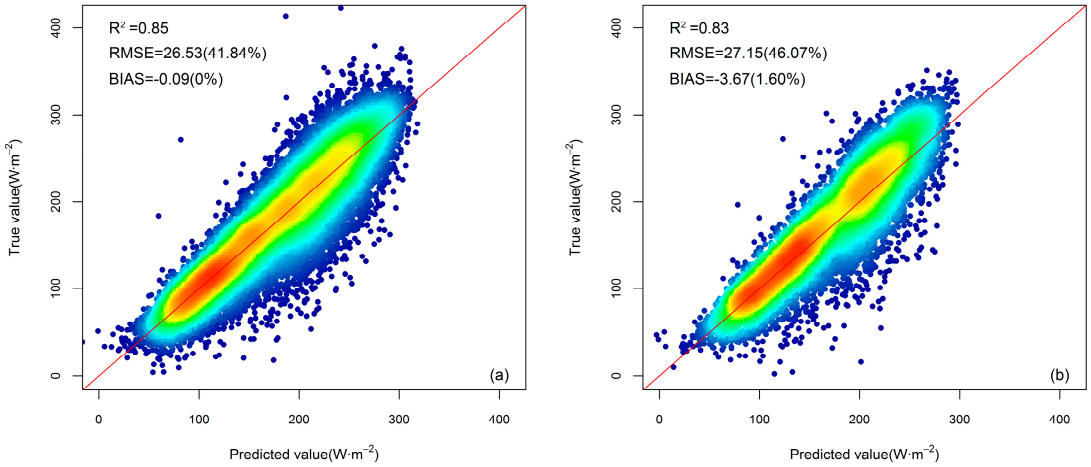

Figure 10. (**a**) Evaluation results of the training dataset's daily estimated DSR based on the ANN-based clear sky model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation dataset's daily estimated DSR based on the ANN-based clear sky model against ground measurements in 2003. The number in the parentheses is the RMSE value.

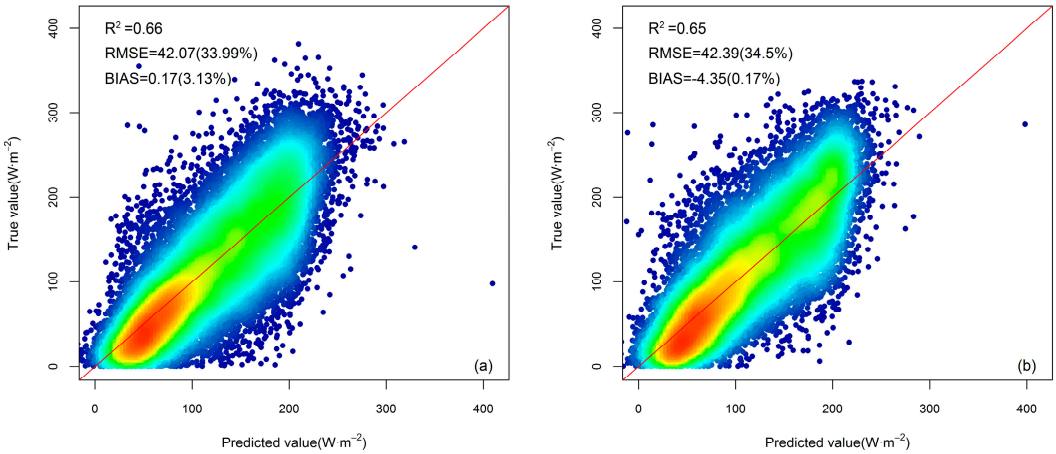

Figure 11. (**a**) Evaluation results of the training dataset's daily estimated DSR based on the ANN-based cloudy sky model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation dataset's daily estimated DSR based on the ANN-based cloudy sky model against ground measurements in 2003. The number in the parentheses is the percent bias or RMSE value.

Table 3. Comparison of the results of the ANN and GBRT models at a daily time scale (using measurements from 2001 and 2002 as the training dataset and measurements from 2003 as the validation dataset). The number in the parentheses is the percent bias or RMSE value.

Sky Condition	Dataset	Method	R^2	RMSE ($W \cdot m^{-2}$)	Bias ($W \cdot m^{-2}$)
Clear sky	Training set	GBRT	0.92	19.05 (19.06%)	0 (2.41%)
		ANN	0.85	26.53 (41.84%)	−0.09 (0%)
	Validation set	GBRT	0.82	27.71 (38.38%)	−2.53 (1.37%)
		ANN	0.83	27.15 (46.07%)	−3.67 (1.60%)
Cloudy sky	Training set	GBRT	0.79	33.37 (30.21%)	0.01 (4.74%)
		ANN	0.66	42.07 (33.99%)	0.17 (3.13%)
	Validation set	GBRT	0.64	42.97 (34.57%)	−2.83 (1.45%)
		ANN	0.65	42.39 (34.50%)	−4.35 (0.17%)

3.2.2. Validation at a Monthly Time Scale

Similar to what we did with the GBRT model, we also validated the estimated DSR at a monthly time scale to further show the accuracy of the ANN method. To perform the comparison, monthly DSR estimates were calculated by averaging the daily DSR of each month. Figure 12a shows the evaluation results of the training dataset based on the ANN-based DSR model in 2003 at a monthly time scale. The R^2 was 0.88, which was lower than that of GBRT model. The RMSE was 18.95 W·m^{-2} (15.81%) larger than that of GBRT model. The evaluation results of the validation dataset's monthly estimated DSR based on the ANN-based DSR model is shown in Figure 12b. The R^2 was 0.87, and the RMSE was 20.05 W·m^{-2} (16.20%). As in the evaluation results at a daily time scale, it is obvious that the GBRT model performs better than the ANN model at a monthly time scale.

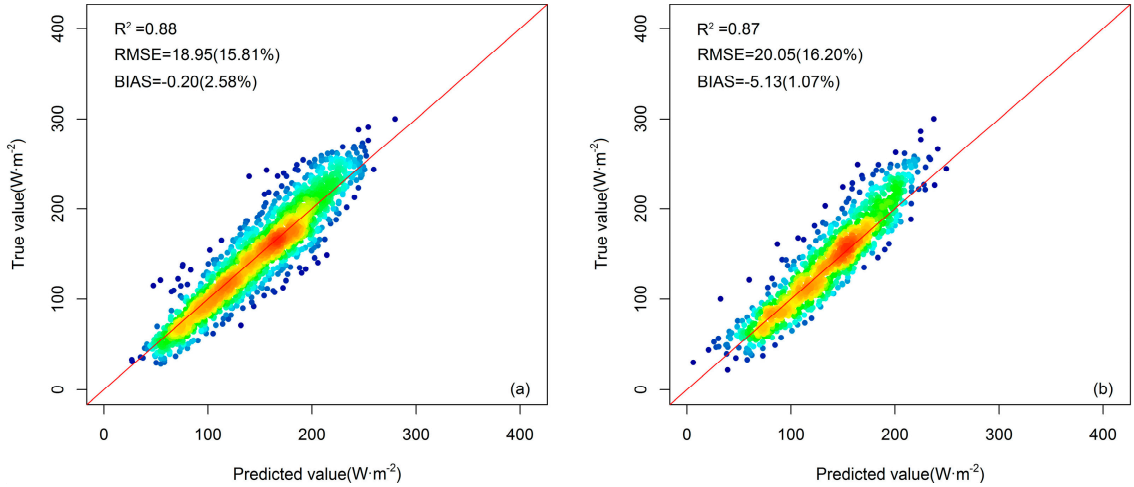

Figure 12. (**a**) Evaluation results of the training set's estimated monthly mean DSR based on the ANN-based DSR model against ground measurements in 2001 and 2002. (**b**) Evaluation results of the validation set's estimated monthly mean DSR based on the ANN-based DSR model against ground measurements in 2003. The number in the parentheses is the percent bias or RMSE value.

Although the DSR estimates based on the GBRT model at both daily and monthly time scales were relatively higher accuracy than those from the ANN-based model, the machine learning methods including GBRT and ANN are sensitive to the choice of parameters. Therefore, the parameters chosen for these two machine learning methods may influence the accuracy of the DSR estimates. In this study, the optimal parameterization scheme was determined by looping in each parameter threshold. Advanced methods for deriving the optimal parameters for both GBRT and ANN should be tested in the future.

3.3. Comparison with Existing DSR Products

3.3.1. Mapping DSR over China

To demonstrate the applicability of the GBRT-based DSR model for regional mapping, the surface monthly mean DSR was estimated based on the GBRT method in the mainland of China in March 2003. Figure 13a shows the estimated results for monthly DSR in March 2003. The GEWEX-SRB and MERRA monthly DSR for the same month are also shown in Figure 13b,c for comparison. According to these three figures, it can be concluded that the spatial distribution of estimated DSR based on the GBRT method is similar to that from the GEWEX-SRB. However, large discrepancies occurred in the comparison with the MERRA. Moreover, the DSR estimates from the GBRT model provide more details compared to the other two existing DSR products.

Figure 13d,e shows the differences between the monthly mean DSR estimates from the GBRT model and those from the GEWEX-SRB and the MERRA, respectively. Before comparison, the DSR

estimates from the GBRT model and the MERRA were projected onto a 1° spatial resolution using bilinear interpolation to match the resolution of the GEWEX-SRB data. As shown in the Figure 13, the differences between the GBRT-based DSR estimates and the GEWEX-SRB DSR product were smaller than that between the GBRT-based DSR estimates and the MERRA DSR product. The maximum differences between the GBRT-based DSR estimates and the GEWEX-SRB DSR product were found in the Tibetan Plateau. The maximum differences between the GBRT-based DSR estimates and the MERRA DSR product were found in southeast China, which were greater than 100 $W \cdot m^{-2}$ at some areas. The large discrepancies in the Tibetan Plateau may be related to the high elevation of the area. Yang et al. [32] pointed out that the discrepancies among the satellite products were always larger in highly variable terrain and smaller for non-variable terrain. The large differences in southeast China were probably due to inappropriate representation of aerosols and clouds, as well as their interactions with the algorithms used for this region [79,80]. In this area, heavy pollution is occurring due to rapid economic development and high population density. However, the DSR comparison of the GBRT model and current existing products were only performed for one month. This may also cause large uncertainties. Therefore, further investigations should be conducted for DSR estimation in the future if long-term DSR estimates are generated based on the GBRT method.

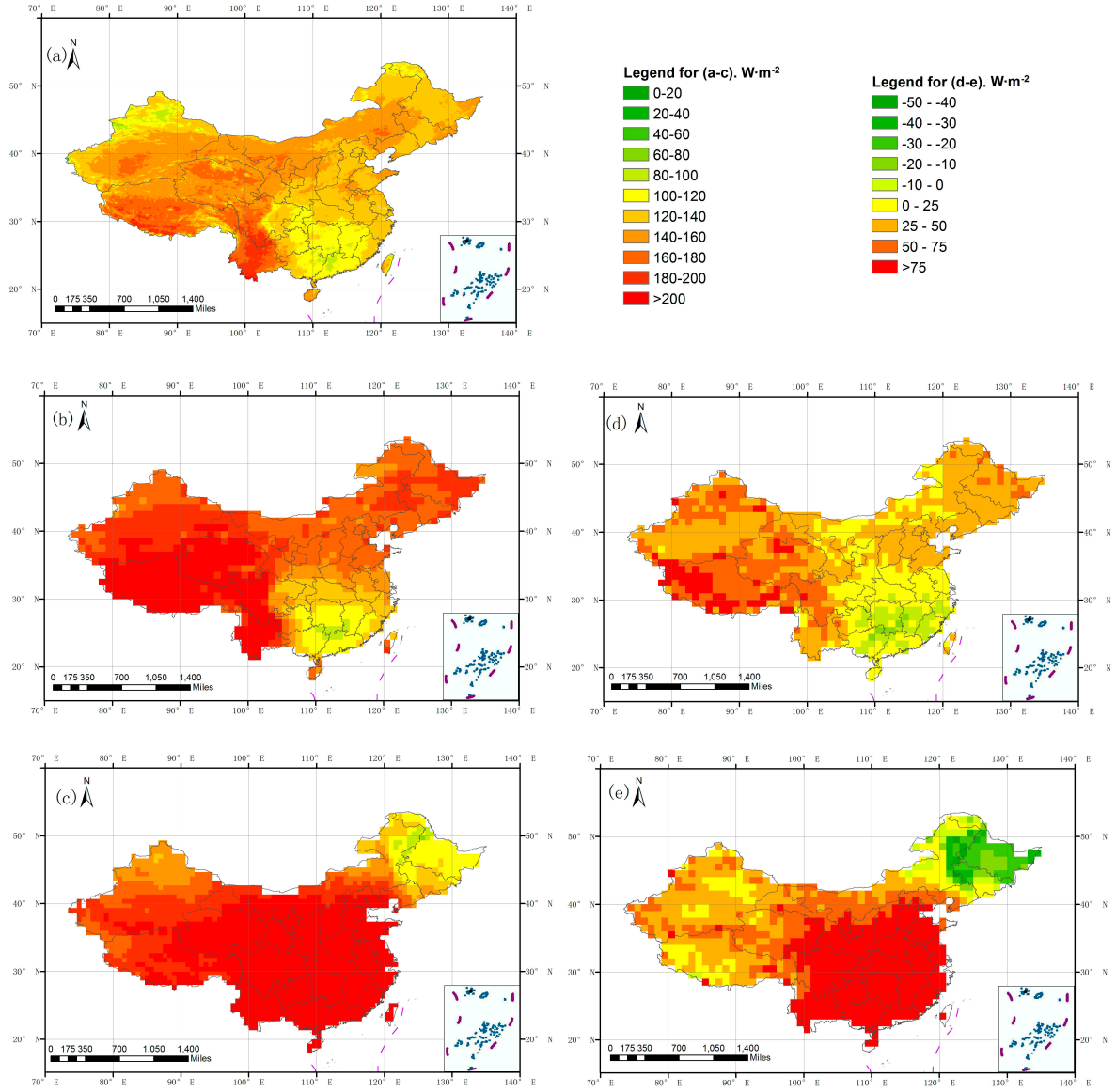

Figure 13. The spatial distribution of the DSR estimates from (**a**) the GBRT model, (**b**) the GEWEX-SRB, and (**c**) the MERRA in March 2003. (**d**) The differences between monthly mean DSR estimates of the GEWEX-SRB and the GBRT model (i.e., the GEWEX-SRB estimates minus the GBRT-based estimates) in March 2003. (**e**) The differences between monthly mean DSR estimates of the MERRA and the GBRT model (i.e., the MERRA estimates minus the GBRT-based estimates) in March 2003.

3.3.2. Validation with Ground Measurements

To further show the accuracy of the DSR estimates based on the GBRT method, we also compared the evaluation results of the GBRT-based daily estimated DSR against ground measurements from CMA in 2003 with those of current existing DSR products from the GEWEX-SRB and the MERRA. As shown in Figure 14, the daily estimated DSR based on the GBRT method correlates very well with the ground measurements, with an RMSE value of 31.65 $W \cdot m^{-2}$ (21.34%) and a bias value of 0.86 $W \cdot m^{-2}$ (1.50%). These values were 40.82 $W \cdot m^{-2}$ (30.93%) and 27.39 $W \cdot m^{-2}$ (17.86%), respectively, for the GEWEX-SRB-based estimates, and 74.2 $W \cdot m^{-2}$ (39.40%) and 57.27 $W \cdot m^{-2}$ (30.06%), respectively, for the MERRA-based estimates. It was obvious that the evaluation results of the GBRT-based DSR model were better than those of the other two products. However, the spatial representativeness of ground measurements is a potential error source for DSR evaluation.

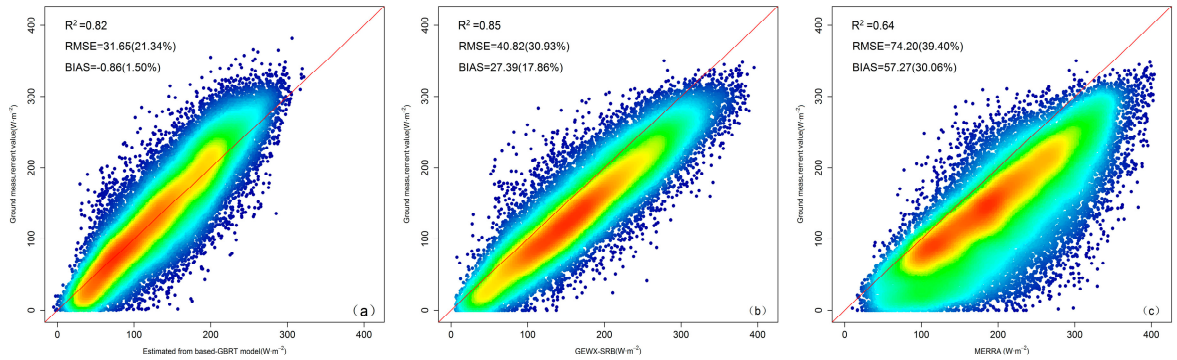

Figure 14. Scatter plots comparing the results from (**a**) the GBRT-based DSR model, as well as the DSR products (**b**) the Global Energy and Water Cycle Experiment-Surface Radiation Budget (GEWEX-SRB) and (**c**) Modern-Era Retrospective analysis for Research and Applications (MERRA) against ground measurements in 2003. The number in the parentheses is the percent bias or RMSE value.

As pointed out by Hakuba et al. [81], the monthly and annual mean representation error at the surface sites with respect to their 1° surroundings are, on average, 3.7% (4 W·m^{-2}) and 2% (3 W·m^{-2}), respectively. The DSR estimates from the GBRT model and current existing radiation products have different spatial resolutions. Therefore, the regional dependence of errors of coarse-resolution satellite products for complex terrain may cause large discrepancies.

4. Conclusions

DSR is an essential parameter in the terrestrial radiation budget and a necessary input for land-surface process models. Although several radiation products using satellite observations have been released, their coarse spatial resolution and low accuracy limit their application. Therefore, high spatio-temporal resolution and high accuracy DSR is still required for many applications. To achieve this goal, a fast, accurate, and robust GBRT method that has the ability to handle different types of input variables and model complex relations was developed to estimate DSR using satellite observations from AVHRR.

The estimated DSR was evaluated using the ground measurements from CMA and was compared with one remote sensing DSR product (the GEWEX-SRB) and one reanalysis DSR product (the MERRA). The daily estimated DSR had an overall R^2 value of 0.82, an RMSE of 27.71 W·m^{-2} (38.38%), and a bias of −2.53 W·m^{-2} (1.37%) under clear sky conditions, and an R^2 of 0.64, an RMSE of 42.97 W·m^{-2} (34.57%), and a bias of −2.83 W·m^{-2} (1.45%) under cloudy sky conditions. Comparison of the DSR estimates with the reanalyzed and the retrieved DSR values from satellite observation showed that the estimated DSR values are reasonably accurate but with higher spatial resolution. However, the DSR comparison of the GBRT model and current existing products was only performed for one month, which may cause large uncertainties. Beside this, measurement errors (e.g., instrument sensitivity, drift, and urbanization effects) and spatial representativeness of surface measurements are potential sources of error in DSR estimation [81]. Therefore, further investigations should be conducted for DSR estimation in the future if long-term DSR estimates are generated based on the GBRT method.

The strengths of GBRT are accuracy, speed, and robustness [51]. To show the advantages of GBRT, an ANN model was built. The results were compared between the GBRT-based DSR model and the ANN-based DSR model under clear and cloudy sky conditions, as shown in Section 3.2. The daily validation analysis showed that the maximum RMSE for GBRT-based and ANN-based clear sky model was less than 28 W·m^{-2}, but the bias of the GBRT-based clear sky model (−2.53 W·m^{-2}) was less than that of the ANN-based clear sky model (−3.67 W·m^{-2}). Similar results were also found for the cloudy sky model. The ANN has two known disadvantages: it needs a relatively long processing time to train a model with many input variables, and it behaves unpredictably when overestimation

occurs during the training stage [82]. In contrast, the GBRT was evaluated as a promising machine learning approach in terms of processing speed and accuracy. All experiments were conducted on a Windows 7 Intel(R) Core(TM) i7-6700 CPU, 3.4 GHz, 20.00 GB RAM processor. The means for the elapsed time of completion of the GBRT clear sky model and the GBRT cloudy sky model were within 10 seconds. Therefore, we conclude that the GBRT method performs better than the ANN method for DSR estimation in this study. As it is well known, the mechanisms of machine learning methods are often considered to be black boxes, and the training procedure is sensitive to the choice of parameters. These limitations may influence the accuracy of the DSR estimates.

The contributions of this study demonstrate that the GBRT is efficient and practical for estimating DSR using remote sensing and ground observation data. Simultaneously, this method has a very good development procedure for defining training data and generating parameters. The method also has more extensive applicability than other current methods. The proposed GBRT-based method can also be used for the retrieval of other land surface variables.

Acknowledgments: This work was supported in part by the National Key Research and Development Program of China (No. 2016YFA0600102 and No. 2017YFA0603002) and in part by the National Natural Science Foundation of China under grant 41571340. The surface observation data of surface incident solar radiation was downloaded from CMA (http://cdc.nmic.cn/home.do). The AVHRR data was downloaded from NOAA National Climatic Data Center (NCDC) (website: https://www.ncdc.noaa.gov/cdr).

Author Contributions: Xiaotong Zhang and Shunlin Liang designed the experiment. Xiaotong Zhang and AoLin Jia collected the required data. YunJun Yao and Kun Jia preprocessed the data. Lu Yang and Xiaotong Zhang performed the experiment. Lu Yang, Xiaotong Zhang, and Shunlin Liang conducted the analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lu, N.; Liu, R.; Liu, J.; Liang, S. An algorithm for estimating downward shortwave radiation from GMS 5 visible imagery and its evaluation over China. *J. Geophys. Res. Atmos.* **2010**, *115*. [CrossRef]
2. Gupta, S.K.; Ritchey, N.A.; Wilber, A.C.; Whitlock, C.H.; Gibson, G.G.; Stackhouse, P.W., Jr. A Climatology of Surface Radiation Budget Derived from Satellite Data. *J. Clim.* **1999**, *12*, 2691–2710. [CrossRef]
3. Gautam, R.; Hsu, N.C.; Lau, K.M.; Kafatos, M. Aerosol and rainfall variability over the Indian monsoon region: Distributions, trends and coupling. *Ann. Geophys.* **2009**, *27*, 3691–3703. [CrossRef]
4. Wild, M.; Ohmura, A.; Gilgen, H.; Roeckner, E. Validation of General Circulation Model Radiative Fluxes Using Surface Observations. *J. Clim.* **1995**, *8*, 1309–1324. [CrossRef]
5. Running, S.W.; Thornton, P.E.; Nemani, R.; Glassy, J.M. *Global Terrestrial Gross and Net Primary Productivity from the Earth Observing System*; Springer: New York, NY, USA, 2000; pp. 44–57.
6. Running, S. W.; Nemani, R.; Glassy, J.M.; Thornton, P.E. *MODIS Daily Photosynthesis (PSN) and Annual Net Primary Production (NPP) Product (MOD17). Algorithm Theoretical Basis Document*; NASA Goddard Space Flight Center: Greenbelt, MD, USA, 1999.
7. Mondol, J.D.; Yohanis, Y.G.; Norton, B. Solar radiation modelling for the simulation of photovoltaic systems. *Renew. Energy* **2008**, *33*, 1109–1120. [CrossRef]
8. Perez, R.; Seals, R.; Zelenka, A. Comparing satellite remote sensing and ground network measurements for the production of site/time specific irradiance data. *Sol. Energy* **1997**, *60*, 89–96. [CrossRef]
9. Blanc, P.; Gschwind, B.T.; Lefèvre, M.; Wald, L. The HelioClim Project: Surface Solar Irradiance Data for Climate Applications. *Remote Sens.* **2011**, *3*, 343–361. [CrossRef]
10. Kaplanis, S.; Kaplani, E. A model to predict expected mean and stochastic hourly global solar radiation *I*(*h*;*nj*) values. *Renew. Energy* **2007**, *32*, 1414–1425. [CrossRef]
11. Wong, L.T.; Chow, W.K. Solar radiation model. *Appl. Energy* **2001**, *69*, 191–224. [CrossRef]
12. Salcedo-Sanz, S.; Casanova-Mateo, C.; Pastor-Sánchez, A.; Sánchez-Girón, M. Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization—Extreme Learning Machine approach. *Sol. Energy* **2014**, *105*, 91–98. [CrossRef]
13. Mellit, A.; Benghanem, M.; Arab, A.H.; Guessoum, A. A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach. *Sol. Energy* **2005**, *79*, 469–482. [CrossRef]

14. Houghton, J.T. Climate change 2001: The scientific basis. *Neth. J. Geosci.* **2001**, *87*, 197–199.
15. Zhang, X.; Liang, S.; Song, Z.; Niu, H.; Wang, G.; Tang, W.; Chen, Z.; Jiang, B. Local Adaptive Calibration of the Satellite-Derived Surface Incident Shortwave Radiation Product Using Smoothing Spline. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1156–1169. [CrossRef]
16. Zhang, X.; Liang, S.; Wang, G.; Yao, Y.; Jiang, B.; Cheng, J. Evaluation of the Reanalysis Surface Incident Shortwave Radiation Products from NCEP, ECMWF, GSFC, and JMA Using Satellite and Surface Observations. *Remote Sens.* **2016**, *8*, 225. [CrossRef]
17. 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. [CrossRef]
18. Lu, X. Estimation of the Instantaneous Downward Surface Shortwave Radiation Using MODIS Data in Lhasa for All-Sky Conditions. Master's Thesis, Clark University, Worcester, MA, USA, 2016.
19. Barzin, R.; Shirvani, A.; Lotfi, H. Estimation of daily average downward shortwave radiation from MODIS data using principal components regression method: Fars province case study. *Int. Agrophys.* **2017**, *31*, 23–34. [CrossRef]
20. Liu, H.; Pinker, R.T. Radiative fluxes from satellites: Focus on aerosols. *J. Geophys. Res.* **2008**, *113*. [CrossRef]
21. Stone, P.H.; Risbey, J.S. On the limitations of general circulation climate models. *Geophys. Res. Lett.* **2013**, *17*, 2173–2176. [CrossRef]
22. Betts, A.K.; Zhao, M.; Dirmeyer, P.A.; Beljaars, A.C.M. Comparison of ERA40 and NCEP/DOE near-surface data sets with other ISLSCP-II data sets. *J. Geophys. Res. Atmos.* **2006**, *111*. [CrossRef]
23. Rossow, W.B.; Zhang, Y.C. Calculation of surface and top of atmosphere radiative fluxes from physical quantities based on ISCCP data sets: 2. Validation and first results. *J. Geophys. Res. Atmos.* **1995**, *100*, 1167–1197. [CrossRef]
24. Gui, S.; Liang, S.; Wang, K.; Li, L.; Zhang, X. Assessment of Three Satellite-Estimated Land Surface Downwelling Shortwave Irradiance Data Sets. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 776–780. [CrossRef]
25. Jia, B.; Xie, Z.; Dai, A.D.; Shi, C.; Chen, F. Evaluation of satellite and reanalysis products of downward surface solar radiation over East Asia: Spatial and seasonal variations. *J. Geophys. Res. Atmos.* **2013**, *118*, 3431–3446. [CrossRef]
26. Zhang, X.; Liang, S.; Wild, M.; Jiang, B. Analysis of surface incident shortwave radiation from four satellite products. *Remote Sens. Environ.* **2015**, *165*, 186–202. [CrossRef]
27. Kim, H.-Y.; Liang, S. Development of a hybrid method for estimating land surface shortwave net radiation from MODIS data. *Remote Sens. Environ.* **2010**, *114*, 2393–2402. [CrossRef]
28. Yang, K.; Koike, T.; Ye, B. Improving estimation of hourly, daily, and monthly solar radiation by importing global data sets. *Agric. For. Meteorol.* **2006**, *137*, 43–55. [CrossRef]
29. Cano, D.; Monget, J.M.; Albuisson, M.; Guillard, H.; Regas, N.; Wald, L. A method for the determination of the global solar radiation from meteorological satellite data. *Sol. Energy* **2010**, *37*, 31–39. [CrossRef]
30. Tang, W.J.; Yang, K.; Qin, J.; Min, M. Development of a 50-year daily surface solar radiation dataset over China. *Sci. China Earth Sci.* **2013**, *56*, 1555–1565. [CrossRef]
31. Perez, R.; Moore, K.; Wilcox, S.; Renné, D.; Zelenka, A. Forecasting solar radiation—Preliminary evaluation of an approach based upon the national forecast database. *Sol. Energy* **2007**, *81*, 809–812. [CrossRef]
32. Yang, K.; He, J.; Tang, W.; Qin, J.; Cheng, C.C.K. On downward shortwave and longwave radiations over high altitude regions: Observation and modeling in the Tibetan Plateau. *Agric. For. Meteorol.* **2010**, *150*, 38–46. [CrossRef]
33. Dongdong, W.; Shunlin, L.; Tao, H.; Qinqing, S. Estimation of Daily Surface Shortwave Net Radiation from the Combined MODIS Data. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 5519–5529. [CrossRef]
34. Rigollier, C.; Lefèvre, M.; Wald, L. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. *Sol. Energy* **2004**, *77*, 159–169. [CrossRef]
35. Muneer, T.; Younes, S.; Munawwar, S. Discourses on solar radiation modeling. *Renew. Sustain. Energy Rev.* **2007**, *11*, 551–602. [CrossRef]
36. Mueller, R.W.; Matsoukas, C.; Gratzki, A.; Behr, H.D.; Hollmann, R. The CM-SAF operational scheme for the satellite based retrieval of solar surface irradiance—A LUT based eigenvector hybrid approach. *Remote Sens. Environ.* **2009**, *113*, 1012–1024. [CrossRef]

37. Li, Z.; Leighton, H.; Cess, R.D. Surface net solar radiation estimated from satellite measurements: Comparisons with tower observations. *J. Clim.* **1993**, *6*, 1764–1772. [CrossRef]
38. 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*, 4975–4988. [CrossRef]
39. López, G.; Batlles, F.J. Estimating Solar Radiation from MODIS Data. *Energy Procedia* **2014**, *49*, 2362–2369. [CrossRef]
40. Mellit, A.; Eleuch, H.; Benghanem, M.; Elaoun, C.; Pavan, A.M. An adaptive model for predicting of global, direct and diffuse hourly solar irradiance. *Energy Convers. Manag.* **2010**, *51*, 771–782. [CrossRef]
41. Jiang, Y. Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. *Energy Policy* **2008**, *36*, 3833–3837. [CrossRef]
42. Voyant, C.; Muselli, M.; Paoli, C.; Nivet, M.L. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. *Energy* **2011**, *36*, 348–359. [CrossRef]
43. Wang, T.; Yan, G.; Chen, L. Consistent retrieval methods to estimate land surface shortwave and longwave radiative flux components under clear-sky conditions. *Remote Sens. Environ.* **2012**, *124*, 61–71. [CrossRef]
44. Qin, J.; Chen, Z.; Yang, K.; Liang, S.; Tang, W. Estimation of monthly-mean daily global solar radiation based on MODIS and TRMM products. *Appl. Energy* **2011**, *88*, 2480–2489. [CrossRef]
45. Zhou, Q.; Flores, A.; Glenn, N.F.; Walters, R.; Han, B. A machine learning approach to estimation of downward solar radiation from satellite-derived data products: An application over a semi-arid ecosystem in the U.S. *PLoS ONE* **2017**, *12*. [CrossRef] [PubMed]
46. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial Neural Networks: A Tutorial. *Computer* **1996**, *29*, 31–44. [CrossRef]
47. McInerney, D.O.; Nieuwenhuis, M. A comparative analysis of kNN and decision tree methods for the Irish National Forest Inventory. *Int. J. Remote Sens.* **2009**, *30*, 4937–4955. [CrossRef]
48. Mubiru, J.; Banda, E.J.K.B. Estimation of monthly average daily global solar irradiation using artificial neural networks. *Sol. Energy* **2008**, *82*, 181–187. [CrossRef]
49. Lam, J.C.; Wan, K.K.W.; Yang, L. Solar radiation modelling using ANNs for different climates in China. *Energy Convers. Manag.* **2008**, *49*, 1080–1090. [CrossRef]
50. Kanamitsu, M.; Ebisuzaki, W.; Woollen, J.; Yang, S.K.; Hnilo, J.J.; Fiorino, M.; Potter, G.L. NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **2002**, *83*, 1631–1643. [CrossRef]
51. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
52. Ma, Y.Z.; Liu, X.N.; Xu, S. The description of Chinese radiation data and their quality control procedures. *Meteorol. Sci.* **1998**, *2*, 53–56. (In Chinese)
53. Moradi, I. Quality control of global solar radiation using sunshine duration hours. *Energy* **2009**, *34*, 1–6. [CrossRef]
54. Tang, W.; Yang, K.; He, J.; Qin, J. Quality control and estimation of global solar radiation in China. *Sol. Energy* **2010**, *84*, 466–475. [CrossRef]
55. Minnis, P.; Bedka, K.; Yost, C.R.; Bedka, S.T.; Scarino, B.A.; Khlopenkov, K.; Khaiyer, M.M. *A Consistent Long-Term Cloud and Clear-Sky Radiation Property Dataset from the Advanced Very High Resolution Radiometer (AVHRR), Climate Algorithm Theoretical Basis Document*; NOAA Climate Data Record Program CDRP-ATBD-0826 Rev.1; CDR Program Library: South Carolina, SC, USA, 2016; Available online: http://www.ncdc.noaa.gov/cdr/operationalcdrs.html (accessed on 20 August 2017).
56. Bhatt, R.; Doelling, D.R.; Scarino, B.R.; Gopalan, A.; Haney, C.O.; Minnis, P.; Bedka, K.M. A Consistent AVHRR Visible Calibration Record Based on Multiple Methods Applicable for the NOAA Degrading Orbits. Part I: Methodology. *J. Atmos. Ocean. Technol.* **2016**, *33*, 2499–2515. [CrossRef]
57. Doelling, D.R.; Bhatt, R.; Scarino, B.R.; Gopalan, A.; Haney, C.O.; Minnis, P.; Bedka, K.M. A Consistent AVHRR Visible Calibration Record Based on Multiple Methods Applicable for the NOAA Degrading Orbits. Part II: Validation. *J. Atmos. Ocean. Technol.* **2016**, *33*, 2517–2534. [CrossRef]
58. Minnis, P.B.; Kristopher; The NOAA CDR Program. *NOAA Climate Data Record (CDR) of Cloud and Clear-Sky Radiation Properties*; Version 1.0; NOAA National Centers for Environmental Information, 2015. https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00876 (accessed on 20 August 2017).

59. Doelling, D.M.; Patrick; The NOAA CDR Program. *NOAA Climate Data Record (CDR) of Visible and Near Infrared Reflectance from GOES and AVHRR*; Version 1.0[C00860]; NOAA National Centers for Environmental Information., 2015. Available online: https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00860 (accessed on 20 August 2017).
60. Rienecker, M.M.; Suarez, M.J.; Gelaro, R.; Todling, R.; Bacmeister, J.; Liu, E.; Bosilovich, M.G.; Schubert, S.D.; Takacs, L.; Kim, G.K. MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Clim.* **2011**, *24*, 3624–3648. [CrossRef]
61. Zhang, T.; Stackhouse, P.W.; Gupta, S.K.; Cox, S.J.; Mikovitz, J.C.; Srb, N.G. *The Effect of Cloud Fraction on the Radiative Energy Budget: The Satellite-Based GEWEX-SRB Data vs. the Ground-Based BSRN Measurements*; American Geophysical Union: Washington, DC, USA, 2011.
62. Wang, Y.; Feng, D.; Li, D.; Chen, X.; Zhao, Y.; Niu, X. A mobile recommendation system based on logistic regression and Gradient Boosting Decision Trees. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1896–1902.
63. Dror, G.; Koren, Y.; Maarek, Y.; Szpektor, I. I Want to answer; who has a question?: Yahoo! answers recommender system. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1109–1117.
64. Son, J.; Jung, I.; Park, K.; Han, B. Tracking-by-Segmentation with Online Gradient Boosting Decision Tree. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3056–3064.
65. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 308–324. [CrossRef]
66. Chung, Y.S. Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accid. Anal. Prev.* **2013**, *61*, 107–118. [CrossRef] [PubMed]
67. Xia, Y.; Jungangb, C.H.E.N. Traffic Flow Forecasting Method Based on Gradient Boosting Decision Tree. In Proceedings of the 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology, Taiyuan, China, 24–25 June 2017.
68. Ding, C.; Wang, D.; Ma, X.; Li, H. Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. *Sustainability* **2016**, *8*, 1100. [CrossRef]
69. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. 2001. *Technometrics* **2001**, *45*, 267–268.
70. Ridgeway, G. Generalized Boosted Models: A Guide to the GBM Package. *Update* **2005**, *1*, 1–12.
71. Suzuki, K. *Artificial Neural Networks—Methodological Advances and Biomedical Applications*; Intech: Rijeka, Croatia, 2011; Available online: https://www.intechopen.com/books/citations/artificial-neural-networks-methodological-advances-and-biomedical-applications (accessed on 10 September 2017).
72. Yadav, A.K.; Chandel, S.S. Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model. *Renew. Energy* **2015**, *75*, 675–693. [CrossRef]
73. Şahin, M. Comparison of modelling ANN and ELM to estimate solar radiation over Turkey using NOAA satellite data. *Int. J. Remote Sens.* **2013**, *34*, 7508–7533. [CrossRef]
74. Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, USA, 28 March–1 April 1993; Volume 581, pp. 586–591.
75. Qian, Y.; Jia, Z.; Jiong, Y.U.; Yang, F. Application of BP-ANN to classification of hyperspectral grassland in desert. *Comput. Eng. Appl.* **2011**, *47*, 225–228.
76. Hatzianastassiou, N.; Matsoukas, C.; Fotiadi, A.; Pavlakis, K.G. Global distribution of Earth's surface shortwave radiation budget. *Atmos. Chem. Phys. Discuss.* **2005**, *5*, 2847–2867. [CrossRef]
77. Sobrino, J.A.; Raissouni, N.; Simarro, J.; Nerry, F. Atmospheric water vapor content over land surfaces derived from the AVHRR data: Application to the Iberian Peninsula. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1425–1434. [CrossRef]
78. Sobrino, J.A.; Jimenez, J.C.; Raissouni, N.; Soria, G. A simplified method for estimating the total water vapor content over sea surfaces using NOAA-AVHRR channels 4 and 5. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 357–361. [CrossRef]

79. Xia, X.A.; Wang, P.C.; Chen, H.B.; Liang, F. Analysis of downwelling surface solar radiation in China from National Centers for Environmental Prediction reanalysis, satellite estimates, and surface observations. *J. Geophys. Res. Atmos.* **2006**, *111*. [CrossRef]
80. Wu, F.; Fu, C. Assessment of GEWEX/SRB version 3.0 monthly global radiation dataset over China. *Meteorol. Atmos. Phys.* **2011**, *112*, 155. [CrossRef]
81. Hakuba, M.Z.; Folini, D.; Sanchez-Lorenzo, A.; Wild, M. Spatial representativeness of ground-based solar radiation measurements. *J. Geophys. Res. Atmos.* **2013**, *118*, 8585–8597. [CrossRef]
82. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139. [CrossRef]