Net Surface Shortwave Radiation Retrieval Using Random Forest Method With MODIS/AQUA Data

Wangmin Ying, Hua Wu, Zhao-Liang Li

Net Surface Shortwave Radiation Retrieval Using Random Forest Method With MODIS/AQUA Data

Wangmin Ying, Hua Wu and Zhao-Liang Li

***Abstract*—The net surface shortwave radiation (NSSR) at the Earth's surface drives evapotranspiration, photosynthesis and other physical and biological processes. The primary objective of this study is to estimate NSSR in all sky conditions by using narrowband data of the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument onboard the AQUA satellite. The random forest (RF) machine learning method for retrieving NSSR was developed with MODerate resolution atmospheric TRANsmission model (MODTRAN 5) simulated data. The bias, root mean square error (RMSE) and R^2 for the training dataset of the model are 0.04 W m^{-2}, 2.03 W m^{-2} and 1.00, respectively; for testing data, these values are 0.53 W m^{-2}, 5.50 W m^{-2} and 1.00, respectively. Note that the proposed method is better than the traditional method (RMSE 7.29 W m^{-2}) with MODTRAN data, and the sky conditions (clear and cloudy) do not need to be distinguished in the RF method. Seven in situ measurements of the Surface Radiation (SURFRAD) observation network were used to validate the estimated NSSR with MODIS/AQUA data using the proposed RF method, and the bias, RMSE and R^2 of the comparison are -8.4 W m^{-2}, 76.8 W m^{-2} and 0.91, respectively. Approximately 70% of the absolute difference of all the samples are below 50 W m^{-2}. Considering its concise process and relatively improved accuracy, both in regard to model development and validation, it can be concluded that retrieval of NSSR with RF will be an efficient and feasible method in the future.**

***Index Terms*—Net Surface Shortwave Radiation, Random Forest, MODIS/AQUA, MODTRAN, Remote Sensing.**

This work was supported by the National Key R&D Program of China under Grant 2018YFB0504800; partially by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA20030302; and partially by the National Natural Science Foundation of China under Grants 41871267. (Corresponding author: Hua Wu.)

W. Ying is with the State Key Laboratory of Resources and Environment Information System (LREIS), Institute of Geographic Science and Nature Resources Research (IGSNRR), Chinese Academy of Sciences (CAS), Beijing 100101, China, and also with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: qsy@zju.edu.cn).

H. Wu is with the State Key Laboratory of Resources and Environment Information System (LREIS), Institute of Geographic Science and Nature Resources Research (IGSNRR), Chinese Academy of Sciences (CAS), Beijing 100101, China, with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China (e-mail: wuhua@igsnrr.ac.cn).

Z.-L. Li is with the Key Laboratory of Agri-informatics, Ministry of Agriculture/Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China, and also with ICube, UdS, CNRS, 67412 Illkirch, France (e-mail: lizhaoliang@caas.cn).

I. Introduction

NET surface radiation, which controls the energy and water exchange between the land and atmosphere, is one of the most fundamental parameters in various applications [1]. As a main component of net surface radiation, the net surface shortwave radiation (NSSR) is calculated as the difference between surface downward shortwave radiation flux and upward surface shortwave radiation flux in the shortwave spectrum (0.3–5.0 μm); it affects regional and global climate change [2] and is the main driver of surface energy balance and evapotranspiration [3]. Thus, reliable NSSR measurements over large areas at high spatial and temporal resolution is required in many applications.

It is widely recognized that remote sensing technology has become a highly effective and convenient method for retrieving land surface radiation [4-6]. In recent years, numerous methods for accurate estimation of NSSR, including a statistical/empirical method [7] and a physically based method [8-10], have been explored and developed. NSSR can be estimated using the relationship between the upward shortwave flux at the top of atmosphere (TOA) and the shortwave flux absorbed at the surface. This method is applied to various sensors including the Landsat Thematic Mapper (TM) [11] and the Moderate-Resolution Imaging Spectroradiometer (MODIS) onboard Terra satellites [12] as well as the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) [13]. However, these traditional methods usually apply numerous specific formulas, and many coefficients of formulas should be fitted, which are always limited by the initial values [14]. In addition, the formulas used in physically based methods may not be good representations of the real interactions. Consequently, a new concise and accurate method for generating NSSR is needed.

Machine learning algorithms can deal with inherent data variability, providing better recognition of data patterns and making better predictions of independent variables [15]. Machine learning algorithms are excellent at performing adaptive nonlinear fitting, which can be used to approximate any complex functional relationship without prespecifying the type of relationship between dependent and independent variables. Today, machine learning algorithms are widely applied to estimate land surface radiation [16-18]. However, the accuracy of these results is limited, and many machine learning methods are prone to the phenomenon of overfitting. Random forest (RF) is a new machine learning algorithm that evolved from the Bagging algorithm [19], which is good at adaptive nonlinear fitting and is able to avoid model overfitting [20-21]. Zhou et al. suggested that the RF model may be

another feasible way to estimate downward shortwave radiation using satellite observations [22]. However, there are few RF applications for NSSR retrieval.

Retrieving NSSR using traditional methods is a complex process and many parameters should be fitted. Therefore, modeling using the traditional method requires professional skill in data structure design, which also limits the development of its application. For example, Tang's method [12], which is used in NSSR retrieval research [11,13,14,30], has two models: the model used to convert narrowband reflectivity to TOA broadband albedo and the model for retrieving NSSR from TOA broadband albedo using their linear relationship. Furthermore, the sky conditions (clear and cloudy) should also be distinguished in the traditional method, and the complex steps may increase the probability of introducing mistakes. However, since the RF method can learn the correct relationship between dependent and independent variables directly, it is necessary to explore whether estimating NSSR with RF can improve retrieval accuracy in the big-data driven background. The objective of this study is to apply RF, a machine learning method, to the estimation of NSSR from MODIS/AQUA remote sensing data under both clear sky and cloudy conditions and to evaluate the feasibility and accuracy of the proposed method. Moreover, we also compare the proposed RF method with Tang's traditional method for NSSR retrieval.

II. DATA

A. *Satellite Data*

The Moderate Resolution Imaging Spectroradiometer (MODIS), one of the sensors in the National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) TERRA satellite, which launched in 1999, and the AQUA satellite, which launched in 2002 [23], provide frequent and comprehensive global imaging in 36 spectral bands with 1-km resolution. MODIS/AQUA passes south to north over the equator in the afternoon and views the entire surface of the Earth every 2 days; these data can improve our understanding of global dynamics and processes occurring on the land and in the lower atmosphere [24].

Various data products are provided by the MODIS instrument onboard the AQUA satellite for many applications in the area of Earth science. Four MODIS products (MYD02, MYD03, MYD05, MYD35) were generated from data collected throughout 2017 and have a spatial resolution of approximately 1 km (https://ladsweb.modaps.eosdis.nasa.gov/). The MYD02 product provides TOA spectral radiance of MODIS/AQUA bands, which would be considered dependent variables in the NSSR retrieval model. Geographic parameters from the MYD03 product, including latitude, longitude, viewing zenith angle (VZA), solar zenith angle (SZA), viewing azimuth angle (VAA), and solar azimuth angle (SAA), were also processed. The MYD05 product contains the atmosphere precipitable water parameter, which is usually considered a dependent variable of the NSSR retrieval model in previous research [11-14]. MYD35 is a cloud mask product, which assigns a clear-sky confidence level (confident clear, probably clear, uncertain clear, cloudy) to each pixel in a remote sensing image. Clear and cloudy samples will be determined by this cloud mask. Confident clear and probably clear pixels were taken as clear, and uncertain clear and cloudy pixels were taken as cloudy in the following analyses.

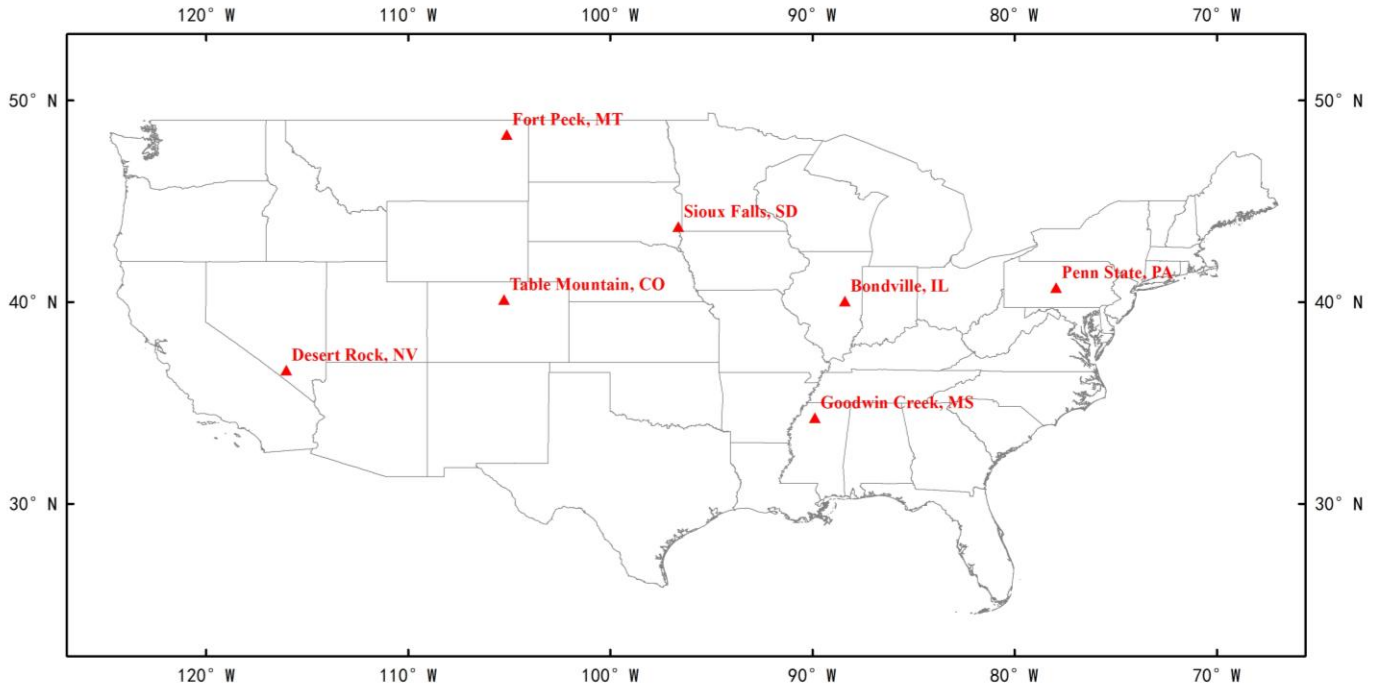

Figure 1. Distribution of seven SURFRAD observation sites.

B. In situ Data

Seven in situ measurements of the Surface Radiation Budget Network (SURFRAD, http://www.srrb.noaa.gov/surfrad) operated by National Oceanic and Atmospheric Administration (NOAA) acquired throughout 2017 were used to evaluate the MODIS/AQUA-derived NSSR. Several variables related to surface shortwave fluxes are measured at SURFRAD stations every minute using pyrheliometers or pyranometers, including net surface shortwave radiation, downwelling global solar radiation, and upwelling solar radiation, together with their quality information [25]. The locations of SURFRAD sites were chosen with the intent of best representing the diverse climates of the United States (Figure 1), and special consideration was given to places where the landform and vegetation are homogeneous over an extended region so that the point measurements would be qualitatively representative of a large area. The seven sites are Bondville, IL; Fort Peck, MT; Goodwin Creek, MS; Table Mountain, CO; Desert Rock, NV; Penn State, PA; and Sioux Falls, SD. These sites represent a variety of surface types, atmospheric conditions and geographic environments [26].

C. Simulated data for modeling

The MODerate resolution atmospheric TRANsmission model (MODTRAN) was developed and continues to be maintained through a longstanding collaboration between Spectral Sciences, Inc. (SSI) and the Air Force Research Laboratory (AFRL). MODTRAN is expert in solving the radiative transfer equation associated with molecular, cloud, aerosol and surface components for emission, scattering, and reflectance [27]. Consequently, MODTRAN is universally used in many sensor data processing systems, particularly for spectral radiance and flux radiation estimations and for removal of atmospheric effects. Here, MODTRAN 5 was used to simulate NSSR at various atmospheric, geometric and surface conditions, which would be used to develop the proposed model.

In our simulations, the spectral range of the MODTRAN model was set to 300 nm-5000 nm and spectral resolution was 2 wavenumbers (cm^{-1}). Nine surface reflectance spectra in the MODTRAN spectral library were employed, including grassland, wetland, sandy loam, broadleaf forest, barren-desert, urban, ocean water, fresh snow, and sea ice. Six atmospheric profiles (tropical, mid-latitude summer, mid-latitude winter, subarctic summer, subarctic winter and US76, with the default atmospheric precipitable water values of 4.11, 2.92, 0.85, 2.08, 0.42, and 1.42 g cm^{-2}, respectively) representing different atmospheres and three types of aerosol models (rural, maritime and troposphere) with default visibility representing different aerosols, were used. Geometric conditions, including six viewing zenith angles (VZA: 0 °, 33.56 °, 44.42 °, 51.32 °56.25 ° and 60 °) and thirty-six solar zenith angles (SZA: 0 °~70 °, interval 2 °), were considered. When VZA was set to these six specific angles, the *1/cos(VZA)* is 1, 1.2, 1.4, 1.6, 1.8 and 2.0, and these values that can help in fitting the coefficients of the following equations. In addition to the clear sky condition, three types of cloud conditions (cumulus, altostratus, and stratus) were also simulated. In total, 139,968 cases were considered in our MODTRAN simulations. After running the simulations, we can get the NSSR from the subtraction between upward flux and downward flux at surface level in the FLUX file. Spectral radiance obtained in the MODTRAN output TP7 file was integrated with MODIS/AQUA narrowband radiances by combining the corresponding spectral response functions.

III. METHODOLOGY

Random forests are an ensemble learning method for regression, which are operated by constructing a multitude of regression trees at training time. Every bootstrap sample is selected from the training set; then, the features used are extracted randomly from all features in a certain proportion while training each mode of the tree [28-29]. Given a training set $X = x_1,..., x_n$ with responses $Y = y_1,..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1,..., B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2. Train a regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{1}$$

RF is also considered a very handy and easy to use algorithm because its default parameters often produce a good predictive results, and its parameters are straightforward and understandable [20-21]. In addition, RF also provides estimates of which participant variables are more important in the regression to quantify the attribution derived from the dependent variables to fit the independent variable.

In the traditional methods, take Tang's method for example, the retrieval of NSSR can be understood from the following formula:

$$NSSR = f(radiance, VZA, SZA, precipitable\ water) \tag{2}$$

The function f in Tang's method consists of two complex models, one of which is a model to convert narrowband reflectivity to TOA broadband albedo. The other model is to retrieve NSSR from TOA broadband albedo using their linear relationship. Consequently, NSSR is estimated using the nonlinear relationship of the dependent variables (band radiance, VZA, SZA, and atmosphere precipitable water). The RF method is essentially used to fit this nonlinear relationship. Like many other machine learning methods (artificial neural network, support vector machine, and so on), the RF method works as a 'black box'. Consequently, it cannot provide the specific form of this nonlinear relationship, but it can learn the correct relationship between dependent variables and NSSR. Unlike other machine learning methods, the RF method is an ensemble learning method, constructed by many regression trees. Every regression tree can train a nonlinear fitting model to estimate NSSR from some of the dependent variables, and the final NSSR is an average of the NSSR values of individual regression trees. The ensemble learning algorithm in the RF

method contributes to the ability to avoid overfitting and to make good predictions, which improves the generation of the model.

Figure 2 shows a flowchart of the RF method proposed in our study. This flowchart shows that the MODTRAN estimated NSSR is regarded as the independent variable in model development. The dependent variables used for the model in our study are MODIS/AQUA six selected radiance bands (the first seven bands except band 6, which is either noisy or nonfunctional shortly after its launch), as well as VZA, SZA and atmospheric precipitable water, which are also the parameters used in the traditional methods in previous research [12]. In the training stage of the RF model, these dependent variables are estimated from MODTRAN, and the parameters in the RF model are set to the default settings. After training, the RF model was applied in retrieving NSSR with MODIS/AQUA satellite data, which would be validated by SURFRAD observing sites.

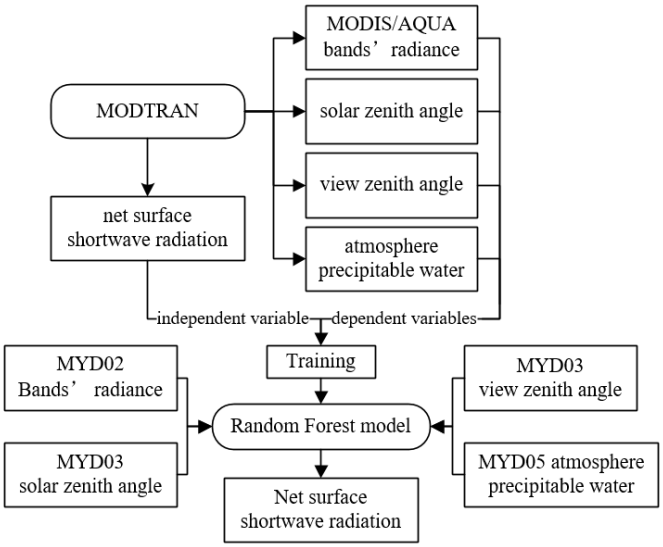

Figure 2. Flowchart of the proposed random forest model.

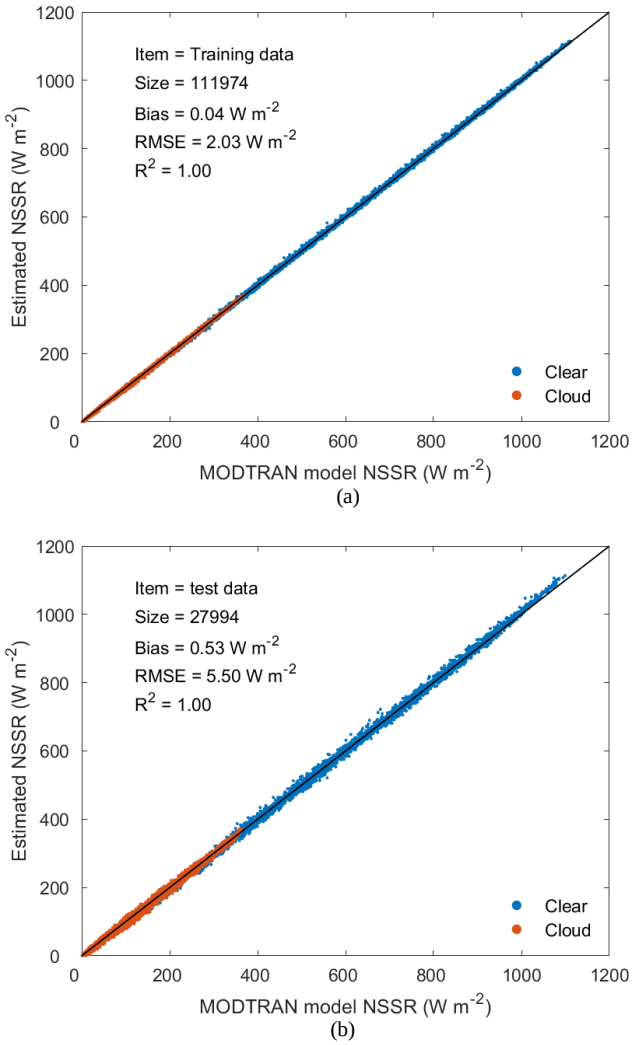

Figure 3. Comparison of NSSR values estimated using the random forest method with MODTRAN-modeled NSSRs. (a) training dataset, (b) test dataset.

IV. Results

A. Algorithm Accuracy with Simulated Data

Figure 3 shows a comparison of the estimated NSSR using the random forest method with the MODTRAN modeled NSSR. Note that the figure includes all of the simulated conditions of the different geographic and atmospheric environments. The dataset was randomly stratified into two groups, with 80% made part of the training dataset (Figure 3a, size: 111,974 cases) and 20% made part of the test dataset (Figure 3b, size: 27.994 cases). The NSSR estimated for the training dataset has an overall bias value of 0.04 W m^{-2}, an RMSE value of 2.03 W m^{-2} and an R^2 value of 1.00. These values are 0.53 W m^{-2}, 5.50 W m^{-2} and 1.00, respectively, for the test dataset. It is found that most scatter points are near the 1:1 line and the error statistics are reasonable. Compared to previous research [30], the RMSE is lower, which illustrates the accuracy and robustness of our RF method in model development.

B. Ground Station Validation with In situ Data

The model produced using the RF machine learning method was developed with MODTRAN simulated data, the efficiency and universality of which should be validated by applying it to MODIS/AQUA data and ground-based measurements. Though the quality of SURFRAD station data is controlled by the data provider, abnormal measurements were excluded before validation. Note that the criterion for the outlier is that the standard deviation of the data is greater than the threshold within five minutes. Clouds, wind and other complex parameters can influence data measurements, leading to inaccurate fluctuating data for the corresponding period [31].

Figure 4 shows a comparison of the estimated NSSR obtained using the RF method with measurements from the seven stations. The bias, RMSE and R^2 of all the samples are -8.4 W m^{-2}, 76.8 W m^{-2} and 0.91, respectively. Approximately 70% of the absolute difference of all the samples are below 50 W m^{-2}. Note that the proposed RF method can estimate NSSR

in all sky conditions, i.e., the scatters from clear sky and cloudy conditions were obtained from the same model. To further avoid the misclassification of cloud and clear in cloud mask product of MYD35, samples with extreme low or high NSSR values were checked visually again, which can help guarantee the reality of sky condition. The scatters acquired under clear skies are distributed evenly around the 1:1 line, with a bias, RMSE and R^2 of -26.2 W m^{-2}, 65.5 W m^{-2} and 0.85 (not shown in the figure), respectively. The bias, RMSE and R^2 of samples collected under cloudy skies are 16.3 W m^{-2}, 90.2 W m^{-2} and 0.34, respectively, and distribution of these scatters is chaotic. Consequently, it can be concluded that the estimation from clear skies is better than estimation from cloudy skies, especially considering that the NSSR value in clear sky conditions is universally larger than cloud condition.

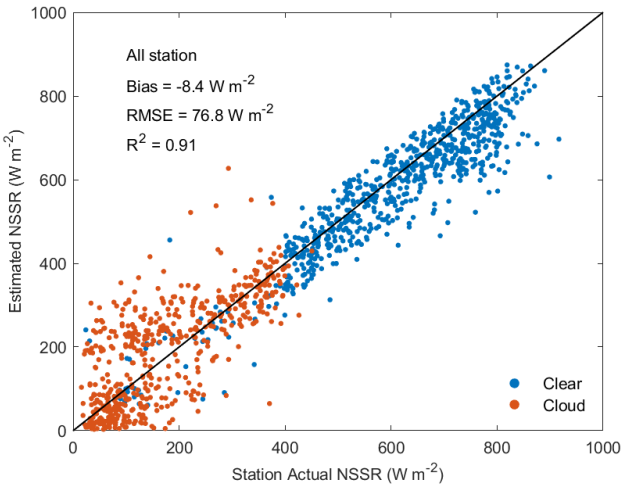

Figure 4. Comparison of RF-estimated NSSR with seven SURFRAD in situ measurements over the year of 2017.

The error statistics of the NSSR retrieval using the proposed method at the seven SURFRAD sites were also evaluated (Table 1). The surface types of the seven sites are obtained from Geostationary Operational Environmental Satellite (GOES) footprints [25]. We found that the grassland stations (Bondville, IL; Fort Peck, MT; Goodwin Creek, MS; and Sioux Falls, SD) had a better accuracy. However, the error of comparison in the Desert Rock, NV site is not very good, probably because the surface type (Arid shrubland) of this site is relatively complex. Relatively more cloud conditions contribute to the relatively large RMSE of station Table Mountain CO.

Table 1. Error statistics of NSSR retrieval using the RF method at seven SURFRAD sites.

Site Name	BIAS (W m^{-2})	RMSE (W m^{-2})	R^2	Surface Type
Bondville, IL	9.1	67.6	0.93	Grassland
Fort Peck, MT	15.8	78.7	0.90	Grassland
Goodwin Creek, MS	-1.4	64.3	0.93	Grassland
Table Mountain, CO	-20.2	85.7	0.88	Sparse grassland
Desert Rock, NV	-56.6	78.3	0.80	Arid shrubland
Penn State, PA	23.4	78.4	0.90	Cropland
Sioux Falls, SD	-17.0	79.5	0.90	Grassland

Compared to previous research at the same stations [26], the accuracy of the proposed method is better. The error of comparison can be explained by the uncertainty of satellite radiance, atmospheric precipitable water, station measurement noise, and so on. It should also be pointed out that there is a spatial scale difference between satellite and ground-based data in our study, which also contributes to the error of the comparison. In other words, overall good accuracy implies that the proposed method is feasible for the estimation of NSSR.

C. Comparison with the Traditional Method

For comparison, we also built the model using Tang's method according to previous research [12, 30] using the following formulas

$$NSSR = \alpha' \cdot \frac{E_0 \mu_s}{D^2} - \beta' \cdot r \cdot \frac{E_0 \mu_s}{D^2} \tag{3}$$

in which *NSSR* represents net surface shortwave radiation, E_0 is the solar irradiance at the TOA, μ_s represents the cosine of the solar zenith angle and D is the Earth-Sun distance in astronomical units. The intercept α' and slope β' are variables that depend on atmospheric precipitable water, SZA and various sky conditions. The r in Eq. (3) represents the TOA shortwave broadband albedo, and an algorithm converting the narrowband reflectivity to broadband albedo was developed:

$$r = b_0 + \sum_{i=1}^{6} b_i \rho_i \tag{4}$$

$$b_i = c_{1i} + c_{2i} / (1 + \exp((1/\mu_v - c_{3i}) / c_{4i})) \tag{5}$$

$$\rho_i = \frac{\pi \cdot L_i \cdot D^2}{\mu_s \cdot \bar{E}_i} \tag{6}$$

where μ_v is the cosine of the viewing zenith angle and c_i is a constant that is mainly dependent on SZA. ρ represents the TOA narrowband reflectivity of the sensor, which can be estimated from TOA band radiances (L_i) according to the assumption that the land surface is Lambertian. $\bar{E}_i$ in Eq. (6) is the mean TOA solar irradiance at band i, which can be estimated by combining spectral irradiance and the corresponding spectral response. Table 2 presents the $\bar{E}_i$ and detailed information of the six selected MODIS/AQUA radiance bands.

Table 2. Spectral range, center spectra and the mean TOA solar irradiance of the six selected MODIS/AQUA bands.

Band	Spectral range (nm)	Center spectra (nm)	E_i (W m^{-2} μm)
1	620-670	645.8	1595.0
2	841-876	856.9	973.9
3	459-479	466.1	2063.5
4	545-565	553.9	1872.6
5	1230-1250	1241.5	235.6
7	2105-2155	2114.0	95.3

As seen from the above, retrieval of NSSR using Tang's method contains two models, one of which is a model to convert narrowband reflectivity to TOA broadband albedo (Eq. (4) - Eq. (6)). The other model is carried out to retrieve NSSR from TOA broadband albedo using their linear relationship (Eq.

(3)). The bias, RMSE and R^2 of the first model are 0, 0.011 and 1.00, respectively, which can demonstrate the robustness of the conversion algorithm. As shown in Figure 5, the bias, RMSE and R^2 for comparison of estimated NSSR from TOA broadband albedo using Eq. (3) with the MODTRAN model are 0 W m^{-2}, 7.29 W m^{-2} and 1.00, respectively. These results contain both clear day and cloudy (cumulus, altostratus and stratus) sky conditions. Note that the relationship between broadband albedo and NSSR is dependent on sky condition, independent of surface types. The results also show that most scatter points, especially clear sky points, are near the 1:1 line and that values of cloud points are often below 400 W m^{-2}. The complex interaction of the cloud situation contributes to uncertainty of simulated radiance and flux, leading to larger deviations for cloud. Note that the figures include all the simulated conditions of different geographic and atmospheric environments.

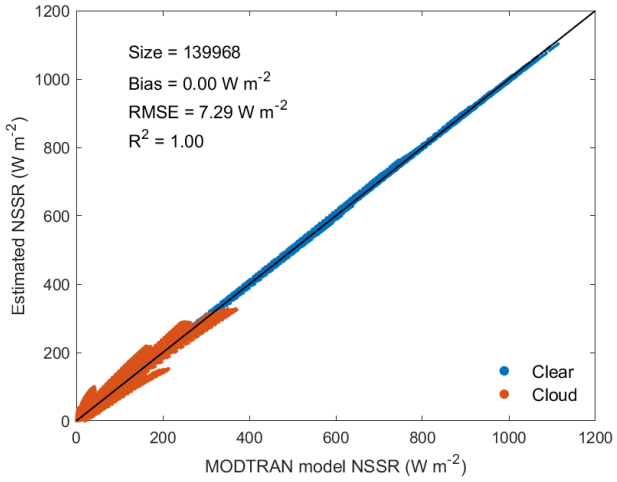

Figure 5. Comparison of the estimated NSSR using the traditional method with a MODTRAN-modeled NSSR value.

The number of our simulations is much larger than in previous articles, which means that our experiment is more universal and can represent complex realities better. The RF method can greatly improve the accuracy of MODTRAN simulated data (Figure 3) compared to Tang's method. It must be noted that many parameters should be fitted and that sky conditions (clear and cloudy) should be distinguished in the traditional method. While the random forest machine learning method can simplify the procedure greatly, this is also a main advantage of the RF model.

An experimental comparison between NSSR estimated by Tang's method and ground-based data from 2017 was also carried out (Figure 6), which had an overall bias value of -5.2 W m^{-2}, an RMSE value of 71.8 W m^{-2} and an R^2 value of 0.92. The bias, RMSE and R^2 for the clear sky samples are 11.3 W m^{-2}, 61.6 W m^{-2} and 0.87 and for the cloudy sky samples are -28.0 W m^{-2}, 84.0 W m^{-2} and 0.42 (not shown in the figure). The accuracy of the clear sky samples is comparable to that of the proposed RF method, but the bias of the cloudy samples is much larger than that of the RF method. Figure 5 shows that the model of the cloudy samples obtained by the traditional method does not perform well, which contributes to the poorer accuracy during validation of the station cloudy sky samples. Table 3 demonstrates the error statistics of NSSR retrieval using the traditional method in seven SURFRAD sites. Unlike the RF method, the accuracy of the grassland stations is not very good, but other surface type stations performed better. The relatively higher number of cloudy samples and the worse cloud model in traditional method contribute to the worse accuracy of the grassland stations. In addition, the underfitting phenomenon of other surface types in the RF method may be overcome if MODTRAN can better represent these surface types.

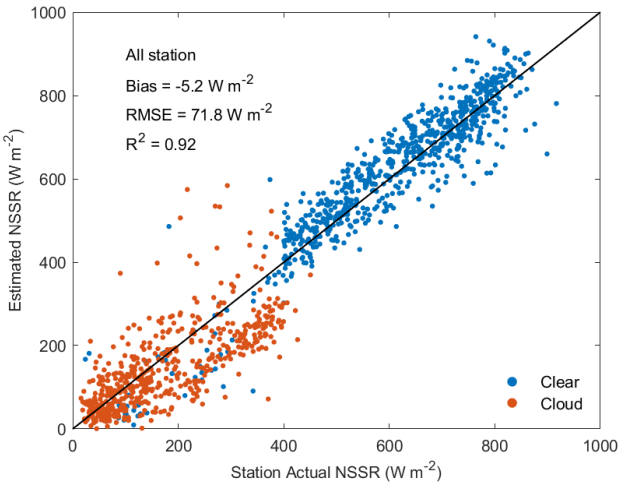

Figure 6. Comparison of NSSR estimated by the traditional method with seven SURFRAD in situ measurements throughout the year of 2017.

Table 3. Error statistics of NSSR retrieval using the traditional method at seven SURFRAD sites.

Site Name	BIAS (W m^{-2})	RMSE (W m^{-2})	R^2	Surface Type
Bondville, IL	-3.0	62.2	0.94	Grassland
Fort Peck, MT	-8.4	69.1	0.92	Grassland
Goodwin Creek, MS	21.3	74.3	0.91	Grassland
Table Mountain, CO	-2.8	77.6	0.90	Sparse grassland
Desert Rock, NV	-12.2	65.4	0.86	Arid shrubland
Penn State, PA	9.7	69.0	0.92	Cropland
Sioux Falls, SD	-28.9	87.4	0.88	Grassland

In a word, the error statistics of the RF machine learning method is slightly better than Tang's method, especially in cloudy conditions. If more MODTRAN simulations representing various real conditions are carried out, the accuracy of the RF method would be improved, considering its powerful nonlinear data fitting ability. Considering its simple process and relatively good accuracy, it can be concluded that retrieval of NSSR with the RF machine method will be an effective technique in the future.

V. Conclusion

Net surface shortwave radiation is one of the most fundamental parameters in various applications and is the main driver of surface energy balance and evapotranspiration. In this study, we developed a concise and accurate method for retrieving NSSR using an RF machine learning method with MODIS/AQUA satellite data under both clear sky and cloudy conditions. NSSR estimated using the RF method was

evaluated using SURFRAD in situ measurements. Comparative results illustrate the accuracy of the proposed method.

The MODTRAN 5 model is used to simulate MODIS/AQUA band radiation, TOA broadband albedo and NSSR in different atmospheric, geometric and surface conditions. These simulated data can help build the new method. The random forest method can greatly simplify the procedure and has the ability to perform adaptive, nonlinear data fitting. The bias, RMSE and R^2 for the training dataset of the model are 0.04 W m^{-2}, 2.03 W m^{-2} and 1.00, respectively, and for the testing data are 0.53 W m^{-2}, 5.50 W m^{-2} and 1.00. The proposed method was also compared to the Tang method. Tang's method contains two models, one of which is a model to convert narrowband reflectivity to TOA broadband albedo; the bias, RMSE and R^2 for all samples in this model are 0, 0.011 and 1.00, respectively. Another model is carried out to retrieve NSSR from TOA broadband albedo using their linear relationship, and the bias, RMSE and R^2 for comparison of estimated NSSR with the MODTRAN-modeled NSSR are 0 W m^{-2}, 7.290 W m^{-2} and 1.00, respectively. Note that sky conditions (clear and cloudy) should be distinguished in the development of Tang's method but not in the RF method. Consequently, the RF method has a stronger and more effective ability to estimate NSSR due to its greater accuracy and concise model development. Although 139,968 cases were simulated, more complex situations (including bidirectional reflectance phenomenon) are required for better, more realistic representations, which may reduce errors and improve the universality of the built model.

The proposed RF method of NSSR was also validated by applying it to MODIS/AQUA data and ground-based measurements. The bias, RMSE and R^2 for the seven SURFRAD stations are -8.4 W m^{-2}, 76.8 W m^{-2} and 0.91, respectively. The error of the models could be explained by satellite channel radiance noise, station measurement noise, uncertainty of cloud mask, atmospheric precipitable water, and so on. Though the overall accuracy of the RF method is comparative with Tang's method, the accuracy of NSSR in cloudy conditions using the RF method is much better than that of the traditional method, which suggests that our applied method is universal and accurate. It can also be concluded that retrieval of NSSR with the RF machine method will be an effective technique in the future due to the simple process and relatively good accuracy.

Future studies will focus on estimating NSSR and other radiation fluxes by training hundreds of station measurements and auxiliary data (including satellite data, meteorological data, reanalyzed data, and so on) with different machine learning methods instead of MODTRAN simulated data. In addition, the spatial and temporal features of global NSSR distribution will be analyzed in future research.

REFERENCES

[1] G. L. Stephens, J. L. M. Wild, C. A. Clayson, N.G. Loeb, S. Kato, T. L'Ecuyer, P. W. Stackhouse, M. Lebsock and T. Andrews, "An update on earth's energy balance in light of the latest global observations," *Nat. Geosci*., vol. 23, pp. 691-696, Oct. 2012.

[2] M. Verma, J. B. Fisher, K. Mallick, Y. Ryu, H. Kobayashi, A. Guillaume, G. Moore, L. Ramakrishnan, V. Hendrix, S. Wolf, M. Sikka, G. Kiely, G. Wohlfahrt, B. Gielen, O. Roupsard, P. Toscano, A. Arain and A Cescatti, "Global surface net-radiation at 5 km from MODIS Terra," *Remote Sens*., vol. 8, no. 9, pp. 739-759, Sep. 2016.

[3] M. Mira, A. Olioso, B. Gallego-Elvira, D. Courault, S. Garrigues, O. Marloie, O. Hagolle, P. Guillevic and G. Bouletf, "Uncertainty assessment of surface net radiation derived from Landsat images," *Remote Sens. Enviro*., vol. 175, pp. 251-270, Mar. 2016.

[4] S. L. Liang, K. C. Wang and M. Wild, "Review on estimation of land surface Radiation and energy budgets from ground measurement, remote sensing and model simulations," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*., vol. 3, no. 3, pp. 225-240, May. 2010.

[5] K. Mallick, A. Jarvis, G. Wohlfahrt, G. Kiely, T. Hirano, and A. Miyata, S. Yamamoto, and L. Hoffmann, "Components of near-surface energy balance derived from satellite soundings – Part 1: Noontime net available energy," *Biogeosciences*., vol. 12, pp. 433-451, Jan. 2015.

[6] R. T. Pinker, R. Frouin and Z. Li, "A review of satellite methods to derive surface shortwave irradiance," *Remote Sens. Enviro.*, vol. 51, no. 1, pp. 108-124, Jan. 1995.

[7] J. D. Tarpley, "Estimating incident solar radiation at the surface from geostationary satellite data," *J. Appl. Meteor.*, vol. 18, pp. 1172-1181, May. 1979.

[8] K. Masuda, H. G. Leigton and Z. Q. Li, "A new parameterization for the determination of solar flux absorbed at the surface from satellite measurements," *J. Climate*., vol. 8, pp. 1615−1629, Jun. 1995.

[9] A. Berk, L. S. Bernstein, G. P. Anderson, P. K. Acharya, D. C. Robertson, J. H. Chetwynd and S. M. Adler-Golden, "MODTRAN cloud and multiple scattering upgrades with application to AVIRIS," *Remote Sens. Enviro*., vol. 65, no. 3, pp. 367-375, Sep. 1998.

[10] Z. Q. Li, H. G. Leighton, K. Masuda and T. Takashima, "Estimation of shortwave flux absorbed at the surface from TOA reflected flux," *J. Climate*., vol. 6, pp. 317-330, Feb. 1993.

[11] D. D. Wang, S. L. Liang, T. He, "Mapping high-resolution surface shortwave net radiation from Landsat data," *IEEE Geosci. Remote Sens. Lett.*, vol, 11, pp. 459–463, Nov. 2013.

[12] B. H. Tang, Z. L. Li and R. H. Zhang, "A direct method for estimating net surface shortwave radiation from MODIS data," *Remote Sens. Enviro*., vol. 103, no. 1, pp. 115-126, Jul. 2006.

[13]. T. He, S. L. Liang, D. D. Wang, Q. Q. Shi, M. L. Goulden, "Estimation of high-resolution land surface net shortwave radiation from AVIRIS data: Algorithm development and preliminary results," *Remote Sens. Enviro.*, vol. 167, pp. 20-30, Apr. 2015.

[14] W. M. Ying, H. Wu and Z. L. Li, "Net surface shortwave radiation retrieval using VIIRS Data," IGARSS 2018., pp. 2623-2626, Jul. 2018

[15] D. V. Mahalakshmi, A. Paul, D. Dutta, M. M. Ali, R. S. Reddy, C. Jha, J. R. Sharma and V. K. Dadhwal, "Estimation of net surface radiation from eddy flux tower measurements using artificial neural network for cloudy skies," *Sustainable Environment Research*, vol. 26, pp. 44-50, Apr. 2016.

[16] D. V. Mahalakshmi, A. Paul, D. Dutta, M. M. Ali, C. S. Jha and V. K. Dadhwal, "Net surface radiation retrieval using earth observation satellite data and machine learning algorithm," *ISPRS J. Photogramm. Remote Sens*, vol. 2, no. 8, pp. 9-12, Sep. 2014.

[17] B. Jiang, Y. Zhang, S. L. Liang, X. T. Zhang and Z. Q. Xiao, "Surface daytime net radiation estimation using artificial neural networks," *Remote Sens.*, vol. 6, no. 11, pp. 11031-11050, Nov. 2014.

[18] L. Yang, X. T. Zhang, S. L. Liang, Y. J. Yao, K. Jia and A. Jia, "Estimating surface downward shortwave radiation over China based on the gradient boosting decision tree method," *Remote Sens*., vol. 10, pp. 185-207, Jan. 2018.

[19] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp.123-140, 1996.

[20] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization." *Mach. Learn.*, vol. 40, pp.139-157, 2000.

[21] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Comput.*, vol. 9, no. 7, pp. 1545-1588, 1997.

[22] Q. T . Zhou, A. Flores, N. F. Glenn, R. Walters and B. S. Han, "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*., vol. 12, no. 8, Aug. 2017.

[23] R. Shrivastava1, I. S. Iyer1, M. N. Hegde and R. B. Oza, "Application of remotely sensed data in the estimation of net radiation at the earth's surface in clear sky conditions," *Amer. J. Remote Sens.*, vol. 6, no. 1, pp. 23-28, Mar. 2018.

[24] Z. Q. Chen, C. M. Hu and M. K. Frank, "Monitoring turbidity in Tampa Bay using MODIS/Aqua 250-m imagery," *Remote Sens. Enviro.*, vol. 109, pp. 207-220, 2007.

[25] D. D. Wang, S. L. Liang, T. He and Q. Q. Shi, "Estimation of daily surface shortwave net radiation from the combined MODIS data," *IEEE Trans. Geosci. Remote Sensing.*, vol. 56, no. 10, pp. 5519-5529, Oct. 2015.
[26] K. I. Anand and C. G. Pierre, "Net surface shortwave radiation from GOES imagery—product evaluation using ground-based measurements from SURFRAD," *Remote Sens.*, vol. 7, pp. 10788-10814, Aug. 2015.
[27] G. P. Anderson, A. Berk, J. H. Chetwynd, J. Harder, and E. P. Shettle, "Using the MODTRAN5 radiative transfer algorithm with NASA satellite data: AIRS and SORCE," Proc. of SPIE., vol. 6565, pp. 65651O-65651O-11, 2007.
[28] P. O. Gislason, J. A. Benediktsson and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, pp. 294-300, 2006.
[29] R. I. Louisr, M. P. Anantham, N. M. Stephen and P. Matthew, "Estimating potential habitat for 134 eastern US tree species under six climate scenarios," *For. Ecol. Manage.*, vol. 254, pp. 390-406, 2008.
[30] X. Y. Zhang and L. L. Li, "Estimating net surface shortwave radiation from Chinese geostationary meteorological satellite FengYun-2D (FY-2D) data under clear sky," *Opt. Express.*, vol. 24, no. 6, pp. A476, Feb. 2016.
[31] D. D. Wang, S. L. Liang, T. He, Y. F. Cao and B. Jiang, "Surface shortwave net radiation estimation from FengYun-3 MERSI data," *Remote Sens.*, vol. 7, pp. 6224-6239, May. 2015.

Wangmin Ying received a B.S. degree in marine science from Zhejiang University, Zhejiang, China, in 2013. He is currently pursuing the M.S. degree in cartography and geographical information systems from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research mainly includes the retrieval and validation of net surface shortwave radiation and machine learning application to remote sensing.

Hua Wu received a B.S. degree in photogrammetric engineering and remote sensing from Wuhan University, Wuhan, China, in 2003, a M.S. degree in cartography and geographical information systems from Beijing Normal University, Beijing, China, in 2006, and a Ph.D. degree in cartography and geographical information systems from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, in 2010.

He is currently an Associate Research Fellow with the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research mainly includes the retrieval, validation, and scaling of remotely sensed products.

Zhao-liang Li received a Ph.D. degree in 1990. Since 1992, he has been a research scientist at CNRS, Illkirch, France. He joined the Institute of Agricultural Resources and Regional Planning in 2013. He has participated in many national and international projects such as the NASA-funded MODIS, the EC-funded EAGLE program, and ESA funded programs such as SPECTRA. His main fields of expertise are thermal infrared radiometry, large-scale parameterization of land surface processes, and the assimilation of satellite data to land surface models. He has published more than 100 papers in international refereed journals.