Determination of CERES TOA Fluxes Using Machine Learning Algorithms. 
Part I: Classification and Retrieval of CERES Cloudy and Clear Scenes

BIJOY VENGASSERIL THAMPI
Science Systems and Applications, Inc., Hampton, Virginia

TAKMENG WONG, CONSTANTIN LUKASHIN, AND NORMAN G. LOEB
NASA Langley Research Centre, Hampton, Virginia

(Manuscript received 3 October 2016, in final form 10 July 2017)

ABSTRACT

Continuous monitoring of the earth radiation budget (ERB) is critical to the understanding of Earth’s climate and its variability with time. The Clouds and the Earth’s Radiant Energy System (CERES) instrument is able to provide a long record of ERB for such scientific studies. This manuscript, which is the first of a two-part paper, describes the new CERES algorithm for improving the clear/cloudy scene classification without the use of coincident cloud imager data. This new CERES algorithm is based on a subset of the modern artificial intelligence (AI) paradigm called machine learning (ML) algorithms. This paper describes the development and application of the ML algorithm known as random forests (RF), which is used to classify CERES broadband footprint measurements into clear and cloudy scenes. Results from the RF analysis carried using the CERES Single Scanner Footprint (SSF) data for January and July are presented in the manuscript. The daytime RF misclassification rate (MCR) shows relatively large values (>30%) for snow, sea ice, and bright desert surface types, while lower values (<10%) for the forest surface type. MCR values observed for the nighttime data in general show relatively larger values for most of the surface types compared to the daytime MCR values. The modified MCR values show lower values (<4%) for most surface types after thin cloud data are excluded from the analysis. Sensitivity analysis shows that the number of input variables and decision trees used in the RF analysis has a substantial influence on determining the classification error.

1. Introduction

Earth’s radiation budget plays an important role in modulating the global climate. Top of the atmosphere (TOA) observations of incoming solar and outgoing terrestrial fluxes are required to determine Earth’s radiation budget. The Clouds and the Earth’s Radiant Energy System (CERES) is a part of the National Aeronautics and Space Administration (NASA)’s Earth Observing System (EOS) (Terra and Aqua satellites), designed to provide an accurate record of TOA reflected and emitted thermal radiative fluxes (Wielicki et al. 1995). CERES instruments measure broadband radiances in the shortwave (SW, 0.2–5 μm), longwave (LW, 5–200 μm), and window (WN, 8–12 μm) band regions over 20-km footprints at nadir. These CERES-measured directional radiances are converted to radiant fluxes using angular distribution models (ADMs). Compared to previously developed ADMs used for the Earth Radiation Budget Experiment (ERBE) mission (Suttles et al. 1988), the accuracy of the new CERES ADM was greatly improved using coincident scene information (clouds and aerosol) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) measurements (Minnis et al. 2003) and supplementary information from atmospheric reanalysis data products (Wielicki et al. 1996; Loeb et al. 2003). The CERES TOA radiative fluxes, derived from combined CERES and MODIS clouds and aerosol information, form the building block for the higher-level advanced CERES data products.

The CERES mission produces a number of different data products with various levels of complexity, starting with a CERES ERBE-like data product to the most advanced CERES Energy Balanced and Filled (EBAF) data product (Loeb et al. 2009). The CERES ERBE-like data product uses the ERBE algorithm, including ERBE scene identification (ID) (Wielicki and Green

Corresponding author: Bijoy Vengasseril Thampi, bijoy.vengasseril.th@ssaiahq.com

DOI: 10.1175/JTECH-D-16-0183.1
© 2017 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
ERBE ADM, the ERBE time and space averaging method (Brooks et al. 1986), to produce a dataset that is compatible with the historical ERBE mission. The CERES ERBE-like data product does not use any of the MODIS scene information and is based purely on stand-alone CERES broadband data. There are certain advantages and disadvantages to using the CERES ERBE-like data product. The most obvious advantage is that the simplicity and stand-alone nature of this product allow users quick access to the CERES radiance and flux data. The CERES Instrument Working Group (IWG) is currently using the ERBE-like products as part of the calibration–validation effort to determine instrument drift artifacts. Another advantage of the ERBE-like data is that they provide a consistent long-term backup dataset in case of imager instrument premature failure, since the advanced CERES data products can no longer be produced without imager information.

The most noticeable disadvantage of the CERES ERBE-like data is that they are based on a 30-yr-old ERBE algorithm. The ERBE fluxes are known to have larger uncertainty than CERES fluxes due to scene ID and ADM errors. To improve the stand-alone CERES TOA fluxes, these two deficiencies must be corrected. This manuscript describes a new CERES algorithm for improving the clear/cloudy scene classification without using any coincident MODIS data. This new CERES algorithm is based on a subset of the modern artificial intelligence (AI) paradigm called machine learning (ML) algorithms. This paper describes the development and validation of the ML algorithm called random forests (RF) to classify CERES broadband footprint measurements into clear and cloudy scenes. Once the clear-sky scenes are successfully classified using the RF method, the next step involves the conversion of clear-sky CERES TOA radiances to the TOA flux. A follow-on paper will describe the conversion of CERES clear-sky and all-sky directional radiances to clear-sky and all-sky fluxes using a separate ML algorithm called artificial neural network (ANN).

ML methods focus on the development of algorithms that can be self-trained to grow and adapt when exposed to new data. The benefit of ML algorithms is that they can iteratively learn from a dataset and find hidden insights without being explicitly programmed. ML algorithms have been successfully adapted to remote sensing and rainfall applications (Cutler and Stevens 2006; Gagne et al. 2009; Bunn et al. 2005; Islam et al. 2014; Deloncle et al. 2007; Gislason et al. 2006; Giannakos and Feidas 2013; Rivolta et al. 2006; Grimes et al. 2003; Hsu et al. 1997; Crawford et al. 2003; Ham et al. 2005; Tett et al. 2013; Lawrence et al. 2006; Kühnlein et al. 2014a,b). Loukachine and Loeb (2003, 2004) successfully demonstrated the capability of ANN in estimating CERES anisotropic correction factors from Single Scanner Footprint (SSF) data without using imager data. In recent years, the RF technique has received considerable interest within the scientific community (Breiman 2001). Although RF is one of the most accurate learning algorithms available, its utilization in atmospheric sciences and meteorology (Ham et al. 2005; Gislason et al. 2006; Islam et al. 2014; Cutler et al. 2007; Prasad et al. 2006; Pal 2005; Rodriguez-Galiano et al. 2012; Kühnlein et al. 2014a) is rather limited. RF is an ensemble classification and regression technique based on the assumption that an aggregation of weak classifiers (decision trees) can produce a more accurate prediction than a single classifier (Dietterich 2003). The RF method has a number of useful features like its efficiency on large datasets and its ability to capture nonlinear association patterns between predictors and responses, making it well suited for remote sensing applications.

In this paper, we use the RF-supervised learning method to retrieve the scene information within a CERES Aqua SSF footprint. A global dataset containing TOA radiances, solar and satellite viewing angles, and ancillary data (from CERES Aqua SSF dataset) over the 10 surface type combinations is used in the analysis. A modified open-source Fortran code of RF based on the Breiman and Cutler method is used for classification (Breiman and Cutler 2003). To demonstrate the accuracy of the method, we compare classified CERES footprints (into clear and cloudy classes) with CERES/MODIS cloud cover data. This manuscript is divided into the following sections: Section 2 describes the CERES data used for the study. Section 3 examines the RF methodology used in the scene classification and describes the building of a test and training dataset. Results from the RF analysis are explained in section 4 along with an examination of the RF misclassification rate. The importance of input variables used in the analysis and their sensitivity to the classification accuracy are also discussed in this section. Section 5 delivers the final summary and conclusions. Additional information concerning the decision trees and the RF method used in the study are provided in appendixes A and B, respectively.

2. Data

For this study, we use monthly CERES Aqua Level 2 SSF data (cross-track mode) for the period 2003–14. The SSF product contains TOA broadband radiance measurements in the SW, LW, and WN band regions along with CERES/MODIS-derived cloud and surface information (Wielicki et al. 1996). As a first selection
criterion, only CERES SSF data in the cross-track scanning mode for each month containing coincident MODIS cloud information are selected for the analysis. CERES SSF data, which are also available in the rotating azimuth plane (RAP) scanning mode, are not used in the analysis, since they are available only for very few years compared to the CERES cross-track mode. Another criterion used in the selection of CERES footprints is the percentage of the footprint area not covered by the imager \( A_{\text{unk}} \). To convert radiances to flux, the original CERES ADMs are applied to those satellite footprints that have \( A_{\text{unk}} < 35\% \) (Loeb et al. 2003). In the present analysis, a rather stringent criterion for SSF footprint selection is used \( (A_{\text{unk}} < 10\%) \) to create a more homogenous database for the analysis. The SSF dataset also contains meteorological and surface variables based on the Global Modeling and Assimilation Office (GMAO)’s Goddard Earth Observing System Data Assimilation System, Version 4.0.3 (GEOS-DAS V4.0.3), product. The GMAO data contain parameters such as surface skin temperature, surface wind speed and direction, precipitable water, etc. More details on the parameters available in the CERES Aqua SSF datasets are provided in the CERES Collection Guide (Geier et al. 2003).

3. Random forest (RF) methodology

In this study, we used an ML technique called random forests to classify CERES TOA radiances. RF consists of an ensemble of tree-structured classifiers (Breiman 2001) known as “decision/classification trees” (DTs). Decision trees are very effective statistical methods for classifying complex data structures when no simple relationship (e.g., linear) between input variables and predictors is apparent. A typical decision/classification tree uses successive if–then conditions applied to the input data to ultimately arrive at a unique result. The RF algorithm developed by Breiman and Cutler (2003) constructs an ensemble of decision trees, each built with a “bootstrap” sample (subsample) of the original training data, with randomized splitting of the decision tree at each node. By aggregating results from a large number of decision tree classifiers (so-called bagging), the RF algorithm vastly improves its performance and avoids overfitting (Breiman 1996; Dietterich 1997; Hothorn and Lausen 2003; Svetnik et al. 2003). More details regarding the RF algorithm and decision trees are provided in the appendixes.

a. Training datasets

The dataset used to train an RF classifier has considerable influence in determining the classification accuracy (Pal and Mather 2003; Campbell 1981). It is essential that the training data provide a representative description of each surface type contained within it. Since the CERES Aqua SSF data contain millions of the CERES footprint, it is impossible to use all these data points directly for the RF training. Therefore, compact training sets are created while keeping in mind that a training dataset must represent the complexity and characteristics of the input data. This is achieved by stratifying the data into homogenous groups using selected input variables in such way that the resulting compact training set is able to represent the complexity of the data. As a first step, the SSF data are split into “daytime” and “nighttime” categories and then further into “clear” and “cloudy” using cloud coverage data available in the SSF product. Nighttime data are defined as footprints having a solar zenith angle (SZA) > 90 and shortwave radiances (SWR) equal to 0. The CERES footprint is defined as clear if it has a MODIS cloud area coverage of 0%, while footprints having cloud area coverage > 0% are termed cloudy. A CERES footprint can sometime contain more than one surface (e.g., ocean and land along coastal regions). To maintain the homogeneity of the surface type, only those SSF footprints with surface area coverage > 90% for a single surface type and having sufficient MODIS information \( (A_{\text{unk}} < 10\%) \) are considered in the analysis.

Other meteorological and surface type variables are used to further improve the homogeneity of the SSF TOA data for RF training. Altogether, 20 surface types and 10 input variables available in the CERES Aqua monthly SSF data are used in the clear/cloudy scene classification of SSF TOA radiances using the RF method. The input variables used in the analysis are SZA and viewing zenith angles (VZA), relative azimuth angle (RAZ), CERES TOA LW and SW broadband radiances measurements (LWR, SWR, respectively), surface wind speed (WIN), atmospheric precipitable water content (PWC), surface skin temperature (SKT), surface emissivity (EMI), and broadband surface albedo (ALB). The selection of surface types and input variables used in the RF analysis is based on the study of Loukachine and Loeb (2003) using ANN for CERES TOA flux estimation. Out of the 10 input variables, 7 were used in the ANN method of CERES TOA flux estimation, while variables like albedo, emissivity, and wind speed were included because of their positive influence on RF scene classification and are discussed in the sensitivity analysis section. For the nighttime analysis, only eight input variables (CERES TOA SW radiances and surface albedo are not available) are used. CERES window channel (IR) broadband radiances are not used in the present analysis because the next generation of CERES instruments will not include a
window channel. To make a compact training dataset, surface types having similar surface characteristics (out of the 20 surface types available in the CERES SSF database) are combined to reduce the total number of surface types to 10. For example, surface types like evergreen needleleaf forest and evergreen broadleaf forest are combined to form the surface type “evergreen forest” (EF), while deciduous needleleaf forest and deciduous broadleaf forest are combined to form “deciduous forest” (DF). Similarly, surface types like “open shrublands” and “tundra” are combined to form dark deserts, while the surface type “bare soil and rocks” is categorized as a bright desert. The 10 modified surface types used in the analysis are as follows: EF, DF, woody savannas and shrublands (WS), dark desert (DD), bright desert (BD), water bodies (WB), grasslands (GR), croplands and cities (CC), permanent and fresh snow (SN), and sea ice (SI).

Details regarding these surface types, which are based on the International Geosphere–Biosphere Programme (IGBP) surface map, can be found in the CERES SSF Collection Guide (Geier et al. 2003) and are explained in detail by Loukachine and Loeb (2004). A training set is then constructed for each surface type.

To create the training datasets, 10 yr of monthly CERES Aqua SSF cross-track instantaneous footprint data for the years 2003–12 are used. Daytime training datasets for the 10 surface types are created by independently stratifying the clear and cloudy SSF data using four input variables for the intervals shown in Table 1. Because of the large range of variation in TOA SW radiances \([-\text{0-350} \text{ W m}^{-2} \text{ sr}^{-1}\)] compared to LW radiances \([-\text{10-150} \text{ W m}^{-2} \text{ sr}^{-1}\]), different bin intervals are chosen for SWR \((10-30 \text{ W m}^{-2} \text{ sr}^{-1})\) and LWR \(10 \text{ W m}^{-2} \text{ sr}^{-1}\) radiances (Table 1). The data are first split into different bins having bin width \(\text{SZA} = \text{VZA} = \text{RAZ} = 1^\circ\). For each surface type, the training dataset can contain 90 SZA bins, 67 VZA bins, and 180 RAZ bins (with a total of 1 085 400 angular bin combinations possible). For each angular bin combo, radiances values are estimated first by separating them into clear and cloudy, and subsequently averaging them for radiance bin intervals shown in Table 1. Hence, each radiance bin contains an average of all the footprint radiances belonging to that interval along with corresponding mean values of variables like emissivity, albedo, etc. For each angular bin interval, these radiances bins are then serially numbered 1, 2, 3, etc., with class 1 allocated to the mean clear-sky radiance residing in the lowest radiance bin (average of all radiances in that bin interval) and the highest class number allocated to the mean cloudy-sky radiance residing in the highest radiance bin for each surface type. Since the number of clear-sky radiance bins is typically small compared to the number of cloudy-sky radiances (the range of clear-sky radiances is smaller than the range for cloudy-sky radiances) for each angular bin, the total number of clear-sky classes will always be fewer than cloudy-sky classes.

For each angular bin interval, the mean values of all the input variables are estimated and a scene/class number is assigned to each bin. The scene/class value represents whether the data in each bin are clear or cloudy. A varying sampling threshold is used in the present analysis for averaging the data, since the number of data points available for RF training and testing varies with surface type. A sampling threshold of three data points per bin is chosen for surface types with a lower number of data points (like forest, savannas, etc.), while a higher value of 10 is used for surface types like water and snow. A nighttime training set is built in a similar fashion using only eight input variables, since surface albedo and SW radiances values are not available during nighttime. It should be noted that the building of a nighttime training set involves using SZA and RAZ values. Using SZA and RAZ values in the nighttime analysis, the latitude/regional variabilities in the data (from the CERES instrument on board the Aqua satellite, which follows a sun-synchronous orbit) can be better captured. This analysis is repeated for all 12 calendar months from January to December so that a total of 120 training sets each for daytime and nighttime analysis are available. Once the training sets are constructed, RF training is carried out using a “decision tree forest” made of a large number of decision trees \(N_{\text{tree}}\) for each surface type. Details on the RF process and decision tree construction are provided in the appendixes. In the present RF analysis, the RF algorithm builds a forest of \(N_{\text{tree}} = 1000\) using five input variables \(m_{\text{tr}}\) at random for a decision tree node split.

### Test dataset

Using the test dataset, the performance of the trained RF can be studied. The test data are obtained from the
monthly CERES Aqua SSF cross-track data for the period 2013–14. The test dataset contains the same input variables as the training set with regular values associated with the SSF footprint being used in each available bin, instead of using mean values (as in the training dataset). In the selection of test data, similar criteria used in the selection of input data for training were followed. The appropriate scene/class number (clear or cloudy) is provided to all the test data points depending on which classification of cloud coverage (clear or cloudy) and radiance bin interval they belong. The classification of the test dataset into clear and cloudy classes is then carried out using the RF algorithm. The RF output containing the classification of test data into clear/cloudy classes is then compared with its original class number to test the classification accuracy. The results from the RF analysis using test data for January and July are explained in the next section.

4. Results

a. Confusion matrix

In supervised machine learning, a confusion matrix (error matrix) is a specific table layout of classification that allows for visualization of the performance of an algorithm on a set of test data for which the true values are usually known (Gislason et al. 2006; Pal 2005). Each row of the confusion matrix represents instances in a predicted class, while each column represents instances in an actual class (or vice versa). Since a confusion matrix contains information about actual and predicted classes by the classifier (random forest), it can be used to evaluate the performance of such systems. Tables 2a and 2b show the confusion matrix for the RF classification using the training and test datasets (July), respectively, for the “water” surface type. The classification of 10 individual classes [clear sky (1–4) and cloudy sky (5–10)] using the RF method is given in Tables 2a and 2b.

In Tables 2a and 2b, the diagonal axis (in bold) represents the cases that are correctly classified by the RF (“true cases”), while off-diagonal values are those that are misclassified (“false cases”). In Table 2, class groups (1, 5) (2, 6) (3, 7) (4, 8) represent the clear-sky and cloudy-sky class data belonging to the same SW radiance bin; that is, class 1 (clear sky) and class 5 (cloudy sky) data belong to the same SW radiance bin interval of 0–10 W m⁻² sr⁻¹, class 2 (clear sky) and class 6 (cloudy sky) data belong to the same SW radiance bin interval of 0–10 W m⁻² sr⁻¹, class 3 (clear sky) and class 7 (cloudy sky) data belong to the same SW radiance bin interval of 0–10 W m⁻² sr⁻¹, and class 4 (clear sky) and class 8 (cloudy sky) data belong to the same SW radiance bin interval of 0–10 W m⁻² sr⁻¹.
(with a bin interval of 10–20 W m\(^{-2}\) sr\(^{-1}\)), etc. From Table 2a, it can be seen that 5030 data points belonging to class 1 (clear) are correctly classified, while 376 data points are incorrectly classified (cloudy) as class 5. Similarly, 5504 data points belonging to class 5 are correctly classified (cloudy), while 406 data points are incorrectly classified as class 1 (clear). This shows that data belonging to class 1 are misclassified by RF usually into class 5, while those from class 5 are misclassified into class 1. A similar pattern can be seen for other clear-sky–cloudy-sky class groups. This happens because some clear- and cloudy-sky class groups occupy similar radiance bins, thereby creating the possibility of misclassification between these classes. The output from RF analysis shows that (Tables 2a and 2b) misclassification mainly occurs between classes belonging to the same radiance bins. However, classes 9 and 10 (cloudy) in Table 2 belong to radiance bins having larger radiance values compared to clear-sky classes and are usually classified correctly. Classes in the training and test datasets (for all surface types) are constructed in such a way that the largest two class numbers represent thick cloudy points, while the rest of the classes usually represent the thin cloudy–clear points. An analysis of the confusion matrix for other surface types also shows a similar pattern for the training and test datasets.

### b. Classification of CERES TOA radiances

To validate the results, we examine the accuracy of the RF method in classifying the test data into their corresponding class (either clear or cloudy). The misclassification rate (MCR) for each surface type is defined as the percentage of total data points that are incorrectly classified. This is different from precision, which is the MCR for individual classes within the dataset for each surface type, while the misclassification rate deals with a broad class group of clear and cloudy. We used CERES Aqua monthly SSF footprint data from January and July of 2013 and 2014 as our test datasets. It should be mentioned that the number of SSF footprints available in the training and test datasets varies depending on the surface type and month. Tables 3a and 3b show the total number of clear-sky and cloudy-sky test data points (daytime and nighttime) used in the RF

### Table 3a. RF classification (daytime) of test data points (clear sky and cloud sky) for different surface types and corresponding misclassification rate (MCR) for the calendar months of January and July (2013–2014).

<table>
<thead>
<tr>
<th>Surface type</th>
<th>January</th>
<th>July</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clear</td>
<td>MCR(%)</td>
</tr>
<tr>
<td>bdesert</td>
<td>397247</td>
<td>25.0</td>
</tr>
<tr>
<td>crops</td>
<td>67304</td>
<td>24.2</td>
</tr>
<tr>
<td>ddesert</td>
<td>122511</td>
<td>13.9</td>
</tr>
<tr>
<td>dforest</td>
<td>4329</td>
<td>19.2</td>
</tr>
<tr>
<td>eforest</td>
<td>5947</td>
<td>35.9</td>
</tr>
<tr>
<td>grass</td>
<td>107040</td>
<td>18.8</td>
</tr>
<tr>
<td>savannas</td>
<td>44874</td>
<td>27.0</td>
</tr>
<tr>
<td>seaice</td>
<td>12259</td>
<td>6.5</td>
</tr>
<tr>
<td>snow</td>
<td>306723</td>
<td>2.9</td>
</tr>
<tr>
<td>water</td>
<td>381583</td>
<td>19.3</td>
</tr>
</tbody>
</table>

### Table 3b. As in (a), but for nighttime.

<table>
<thead>
<tr>
<th>Surface type</th>
<th>January</th>
<th>July</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clear</td>
<td>MCR(%)</td>
</tr>
<tr>
<td>bdesert</td>
<td>216082</td>
<td>12.2</td>
</tr>
<tr>
<td>crops</td>
<td>92083</td>
<td>27.6</td>
</tr>
<tr>
<td>ddesert</td>
<td>147753</td>
<td>31.0</td>
</tr>
<tr>
<td>dforest</td>
<td>13961</td>
<td>26.9</td>
</tr>
<tr>
<td>eforest</td>
<td>40054</td>
<td>41.8</td>
</tr>
<tr>
<td>grass</td>
<td>116956</td>
<td>27.6</td>
</tr>
<tr>
<td>savannas</td>
<td>17566</td>
<td>28.3</td>
</tr>
<tr>
<td>seaice</td>
<td>14190</td>
<td>7.1</td>
</tr>
<tr>
<td>snow</td>
<td>301740</td>
<td>17.4</td>
</tr>
<tr>
<td>water</td>
<td>358756</td>
<td>15.6</td>
</tr>
</tbody>
</table>

Unauthenticated | Downloaded 06/06/22 05:38 AM UTC
analysis for January and July (2013–14), and the corresponding misclassification rate (%) for each surface type. From Table 3a it can be seen that the MCR values for both clear-sky and cloudy-sky daytime data (for January and July) are relatively lower compared to the corresponding nighttime dataset (Table 3b). For the daytime dataset, the misclassification rate shows a large difference between the clear and cloudy data. The clear-sky MCR values for January show small values for snow (~3%) and sea ice (6.5%), while relatively large values (>25%) are shown for savannas, forest, and bright desert. In July, smaller MCR values (<7%) are observed for snow and evergreen forest, and larger values (>20%) are observed for dark forest, bright desert, and crops. An analysis of cloudy-sky MCR values for both months shows lower values for water, crops, and forests (both bright and dark) surface data, and larger values for snow, sea ice, and desert surface types.

For the nighttime dataset, the lowest clear-sky MCR values are observed for snow, sea ice, and water surfaces for January and July, while the largest values are observed for desert and forest surfaces. For cloudy-sky data, the smallest (<15%) MCR values for both January and July are observed for grass, savannas, and evergreen forests, while the largest MCR values (>49%) are observed for sea ice and snow. From Table 3b, it can be seen that the nighttime MCR values are relatively higher for most of the surface types compared that in Table 3a. From Tables 3a and 3b, it can be seen that the lowest clear-sky MCR values are observed for snow and sea ice surface types, while the lowest cloudy-sky MCR values are observed for forest surface types. It should also be noted that the data volume available for training and testing is relatively low for forest and land surface types compared to water and snow surfaces. There is also a difference between clear-sky and cloudy-sky data volume, since clear-sky SSF footprints are fewer in number compared to cloudy footprints for most of the surface types (except for bright desert). The observed increase in MCR for the nighttime data may be due to the absence of CERES SW TOA radiance and albedo values in the input dataset. The influence of classification accuracy on the number of input variables used in the RF analysis is explained in the sensitivity analysis section. An intercomparison of the clear/cloudy scene classification by the CERES ERBE-like method and the RF method for coincident CERES footprints belonging to January and July (2013) is carried out. The analysis shows that a total misclassification observed for all available data (which include all 10 surface types) is on the order of ~20.1% and ~10.6% for the RF method and ~23.2% and ~15.8% for the ERBE-like method for January and July 2013, respectively. This study shows that CERES scene classification using the RF method provides better results (for most of the surface types) than using the ERBE-like approach.

The misclassification rate for cloudy-sky data is further analyzed using cloud optical depth and cloud fraction observed in the CERES footprint. Figure 1a shows the cumulative percentage plot between misclassified cloudy-sky (as clear) test data points (%) and the corresponding effective cloud optical depth (COD) values along the y axis. Here, the cloudy-sky data from CERES SSF can have an optical thickness as low as ~0.05 and can be as high as ~128.
Also, the COD mentioned here is the effective cloud optical depth as a result of either a single layer or multiple layers of clouds. More than 50% of the misclassified data points in Fig. 1a have COD < 1 with 90% of the data points lying below COD < 5. It should also be noted that sometimes CERES footprints can be covered by broken cloud fields, which usually cover only a fraction of the footprint area. The magnitude of the TOA radiances in a CERES SSF cloudy-sky footprint is sensitive to cloud properties (cloud fraction and cloud optical depth) in that bin. Studies have shown that for overcast conditions, the albedo is approximately linear in ln(COD) when the variability in the cloud optical depth is small (Cahalan et al. 1994; Loeb et al. 2005). For an SSF footprint having either overcast or broken cloud fields, the CERES SW TOA radiance is dependent on both COD and cloud fraction \( F \). Therefore, a combination of cloud fraction and COD in a single parameter provides a better understanding of the "cloud strength" in a particular CERES footprint. Plots in Fig. 1 compare misclassified data points of the "cloud strength" in a particular CERES footprint. Higher values (negative) of both parameters indicate the strength of a cloud field in a CERES footprint. Lower values (negative) of both parameters indicate lower cloud strength (thin clouds), while positive values (COD > 1) typically indicate a strong cloud presence. From Fig. 1b it can be seen that the majority of misclassified points (>80%) have lower cloud strength. This indicates that the majority of cloudy-sky footprints misclassified as clear by the RF algorithm contain either optically thin clouds or clouds with very small coverage. Since optically thin clouds have a small radiometric influence on TOA radiances, it is difficult to distinguish whether such footprints are clear or cloudy using the RF method. It should also be mentioned that the radiative impact of such optically thin clouds is minimal compared to optically thicker clouds.

To better understand the accuracy of the RF method, optically thin cloudy-sky data points (having low cloud strength) misclassified as clear and clear-sky data points misclassified as cloudy (optically thin clouds) are to be removed from the misclassified dataset. For cloudy-sky footprints misclassified as clear, this can be achieved by removing all of the points residing above a certain threshold value of cloud strength from the analysis. However, this method can be used only for cloudy-sky data; it cannot be used for clear-sky data. Another approach we developed involves removing those misclassified data points whose TOA SW (for daytime) and TOA LW (for nighttime) mean radiance values exceed the corresponding training set mean value by a threshold value \( \Delta r \), where \( \Delta r \) is defined as the absolute difference in TOA radiance between the misclassified test data point and the corresponding training data point belonging to the same angular (SZA, VZA, RAZ) and radiance bins. In this analysis using daytime data, we considered only those misclassified points for which \( \Delta r > 10 \text{ W m}^{-2} \text{ sr}^{-1} \). The logic behind the assumption is that if a test data point is indeed misclassified as clear (or vice versa) by the RF algorithm, then the SW (LW for nighttime) radiance value for such a data point should not be closer to the corresponding mean radiance value in the training dataset (belonging to the same angular and radiance bins). In the present RF analysis, clear-sky data points misclassified as cloudy and cloudy-sky data points misclassified as clear mostly occur when the clear-sky and cloudy-sky radiances (clear and cloudy class radiances, respectively) are closely aligned, which is expected for CERES footprints that are either clear or contain optically thin clouds. Using the above-mentioned criteria (\( \Delta r > 10 \text{ W m}^{-2} \text{ sr}^{-1} \)), a modified...
misclassification rate is calculated by removing these data points and is shown in Table 4. For the nighttime dataset, a similar exclusion criterion involving a TOA LW radiance value ($\Delta r > 5 \text{ W m}^{-2} \text{ sr}^{-1}$) is used in calculating the modified misclassification rate.

Using this methodology, the modified misclassification rate (for clear-sky and cloudy-sky data) is estimated for January and July and is shown in Tables 4a and 4b. From Table 4a it can be seen that the modified MCR values (daytime) for most surface types are relatively low ($\lesssim3\%$) for clear-sky data and cloudy-sky data. However, a modified MCR shows values $>4\%$ for bright desert (for both months) and snow in January. Compared to daytime data, modified MCR values for most surface types show low values ($<2\%$) for both clear-sky and cloudy-sky data during nighttime.

**FIG. 2.** Bar diagram of the VI score estimated for the 10 input variables ($x$ axes) and (top left to bottom right) the 10 different surface types used in the daytime RF classification for July. The $y$ axis shows the VI score.
(Table 4b). Relatively large MCR values (>3%) are observed for surface types like snow and sea ice (for cloudy-sky data) during nighttime. In general, the cloudy-sky data show relatively large MCR values compared to the clear-sky data. Large values of modified MCR (cloudy-sky data) observed for the snow and bright desert during daytime could be due to the contribution of increased surface reflectance to the TOA.
SW radiance compared to other surface types. As an exercise, an attempt was made to understand how the modified MCR changes the CERES ERBE classification results and how they compare to the RF modified MCR results. An intercomparison of the modified MCR between the CERES ERBE-like method and the RF method using the daytime dataset shows lower modified MCR values for the RF method compared to the ERBE-like method for most surface types, except for snow and sea ice. The mean modified MCR values (for all available data, which include 10 surface types) are on the order of ~1.16% and ~0.29% for the RF method and ~0.89% and ~0.47% for the ERBE-like method for January and July 2013, respectively. A slightly larger value of the modified MCR for the RF method compared to the ERBE-like method for January is mainly due to the contribution from snow and sea ice surface types. It should also be mentioned here that such thin cloud screening is not possible using historical ERBE data because of the lack of imager information.

c. Variable importance

One of the main disadvantages associated with the machine learning method is the difficulty associated with interpreting the relationship between the input predictor variables and output classification. However, a by-product of the RF analysis known as the variable importance (VI) score provides quantifiable information about the importance of input variables and provides insight into the relation between input variables and output classification (Breiman and Cutler 2003; Liaw and Wiener 2002). The RF algorithm estimates the VI score for each input variable, where a higher VI score for a particular input variable means greater influence on the classification. However, the VI score discussed here was not used in the selection of input variables but used to analyze the impact of input variables to the RF classification in terms of a quantifiable value. Using the RF algorithm, the VI score for all the input variables used in the analysis for daytime and nighttime is estimated and plotted in Figs. 2 and 3, respectively. More details on the VI score and its calculation are provided in appendix B.

Figure 2 is a bar diagram of the VI score estimated for the 10 surface types using daytime SSF data for July. The x axis shows the 10 input variables used in the RF classification (daytime) and the y axis shows the VI scores (multiplied by 100 for better representation). For all surface types, the VI score is <10 for most input variables except for SWR and LWR. The VI score of SWR exceeds 65 for all 10 surface types, indicating that SWR is the most important variable and has significant influence in determining the RF classification. Compared to SWR, LWR has a VI score of 20–50 for most surface types, indicating that LWR is the second-most important variable. From Fig. 2, it can be seen that the variables SWR and LWR contribute the most to the RF classification (daytime). All other input variables have very low VI score (<10), indicating they have a lower contribution to the classification compared to SWR and LWR. However, for some surface types (like sea ice, snow and water), variables like SZA, PWC, and SKT have relatively high VI scores. This indicates that depending on the surface type, input variables wield varying influence on the RF classification. VI scores for the eight input variables used in the nighttime analysis are shown in Fig. 3. Figure 3 shows that for nighttime RF analysis, LWR has the highest VI score (>80) over 10 surface types while other input variables have a VI score < 20. It is interesting to note that some input variables (like skin temperature, precipitable water, etc.) have higher VI scores (>10) during nighttime analysis compared to daytime analysis (mostly below 10). Compared to daytime analysis, higher VI scores were observed for variables SKT (~21) and PWC (~32) over snow. This shows that while daytime RF classification is largely influenced by the TOA radiances, nighttime RF classification is influenced by both TOA radiance and geophysical variables like precipitable water vapor and surface properties.

d. Sensitivity analysis

The successful classification using the RF method depends essentially on the number of internal parameters and input variables used in the classification. In this section, a sensitivity analysis is carried out to understand the influence of various internal factors that control the RF classification. One of the important user-defined parameters that may influence the RF classification is the number of decision trees built and the number of input variables used to build the decision trees. In the first analysis, the sensitivity of the RF algorithm to the number of decision trees built is studied while keeping the number of input variables used to build the decision trees constant. RF analysis using the training dataset (July for a water surface) is then carried out by varying the number of decision trees (Ntree = 1–1000) while keeping the value of mtry constant. This analysis is repeated by varying the mtry values (from 1 to 9) and classification error is estimated for each instance. Results from the analysis are shown in Fig. 4 with the classification error (%) plotted along the y axis for a number of decision trees and for different values of mtry. Figure 4 shows that the error values decrease as Ntree and mtry increase. The error decreases (from ~70%) with increasing Ntree values to reach a lower value.
an average of 3% around \( N_{\text{tree}} = 100 \) and then it enters a plateau region. The error values are relatively high when \( m_{\text{try}} = 1 \) and \( N_{\text{tree}} < 10 \). For \( m_{\text{try}} > 4 \), the error values converge and show little variation with an increase in \( N_{\text{tree}} \) values. The error values show a considerable decrease between \( N_{\text{tree}} = 1 \) and 150, after which they show very little variation to an increase in \( N_{\text{tree}} \) values.

In the next sensitivity analysis, the sensitivity of the RF algorithm to the total number of input variables in the dataset is analyzed. This analysis is carried out by running the RF code with the number of input variables increasing from 1 to 10 in steps and estimating the classification error each time. This analysis use a fixed value of \( N_{\text{tree}} = 1000 \) and is carried out for 10 surface types using the daytime data. Figure 5 shows the variation in classification error (y axis) against the number of input variables (x axis) for the 10 surface types used in the daytime RF classification. From the figure, it can be seen that the classification error decreases as the total number of input variables (\( M \)) increases from 1 to 10. When \( M = 1 \), the error values are the highest (ranging from \( \sim 10\% \) to 45%) for all surface types. For lower values of \( M (<5) \), the error is the highest for surface types like sea ice, snow, and bright desert, while the error is the lowest for savannas and evergreen forest. The error gradually decreases as the number of input variables in the analysis is increased, and it reaches a lower value of \( \sim (1\%–5\%) (M = 10) \) for all surface types. A sensitivity analysis using nighttime data also shows a similar pattern.

5. Conclusions

This manuscript describes a new CERES methodology for improving the stand-alone CERES TOA fluxes without the use of coincident cloud imager data. This paper describes the development and validation of the machine learning (ML) algorithm called “random forests” to classify CERES broadband footprint measurements into “clear” and “cloudy” scenes. Using the RF method, we developed a methodology for the scene classification (clear or cloudy) of the CERES SSF footprint using TOA radiances and available ancillary data. A monthly training dataset for both daytime and nighttime analysis is built for 10 surface types using 10 input variables and 10 years (2003–12) of CERES Aqua SSF data. Similarly, a dataset to test the efficiency of the RF classification is also built using 2 years (2013–14) of CERES Aqua SSF data. Using a modified Breiman and Cutler open-source RF code, a supervised classification analysis is carried out over CERES footprints with known cloud cover information. RF analysis shows large MCR values for surface types like snow, sea ice, and
bright desert while lower MCR values are observed for forest, grass, and savannas. Compared to daytime data, MCR values are relatively larger for nighttime data. The modified MCR values estimated for January and July (for both clear-sky and cloudy-sky datasets) are less than 5% for most surface types (except for bright desert and snow). The importance of various input variables used in the RF analysis was computed (VI score) for all 10 surface types. The VI score was found to be higher (>80) for TOA broadband radiances (SW and LW) for daytime and (LW only) nighttime data, indicating a relatively higher influence of these variables on the RF classification. Sensitivity analyses performed using the RF technique show that the number of input variables used in the decision tree construction and the increasing number of decision trees used significantly reduces the error associated with the classification.

Acknowledgments. CERES SSF data were obtained from the NASA Langley Research Center EOSDIS Distributed Active Archive Center. The authors thank NASA LaRC and Science Systems and Applications, Inc. (SSAI), for providing the necessary support during this work (performed as a part of a NASA CERES project). The authors are thankful to Dr. Leo Breiman and Dr. Adele Cutler for providing the RF FORTRAN code used in the analysis, and to David Rutan (SSAI) for proofreading of the manuscript.

APPENDIX A

Decision Trees

A DT is a type of supervised learning methodology that can be used as a tool in classification and regression. The objective of a DT is to create a model that predicts the output variable by learning simple decision rules inferred from the input data. A typical decision tree can be represented in the form of a flowchart-like tree structure in which each internal node represents a test on an attribute, with each tree branch representing an outcome of the test and each leaf node representing a class label or output. Decision trees are usually built from the top down (from a root node) that involves partitioning the data into subsets and paths from root to leaf node representing classification rules. At each node, the splitting rule is determined to maximize the “purity” or “homogeneity” of the resulting subset nodes in such a way that resulting nodes contain instances with similar or homogenous values. Figure A1 depicts the typical structure of a binary decision tree.

![Figure A1](image_url)

**Fig. 5.** Variation in error (y axis) against the total number of input variables (x axis) for the 10 surface types used in the RF classification.
Fig. A1. (top) Schematic of a DT with feature values ($x$), threshold values ($a$–$e$) and class labels ($A$–$C$). (bottom) General architecture of the RF algorithm.
In Fig. A1, decision tree nodes (oval shape) represent tests performed by the tree, while the terminal or leaf nodes (square shape) represent the outcome. A DT splits a dataset into smaller and smaller subsets while growing an associated decision tree in incremental steps. The final result is a completed tree with decision nodes and leaf nodes as shown Fig. A1. Decision trees are used mainly for classification problems because of the prediction error and affects the classification accuracy. Trees sometimes produce high variance, which increases become unstable due to variations in the data. Decision trees are combined to form a strong ensemble classifier. The main disadvantage of decision trees is that they can create complex models that do not generalize the data well (overfitting) and can sometimes become unstable due to variations in the data. Decision trees sometimes produce high variance, which increases the prediction error and affects the classification accuracy. This problem can be mitigated to an extent by using decision trees within an ensemble. A common approach is to produce several different decision trees from a single training dataset and to use some aggregation technique to combine the predictions of all these trees.

APPENDIX B

Random Forests

Random forest is an ensemble learning method for classification and regression using decision tree predictors. In ensemble learning method, instead of using a single classifier, a large number of individual classifiers are combined to form a strong ensemble classifier. Breiman (1996) showed that prediction/classification error can be reduced by aggregating the results over a large number of unstable, weak classifier like decision trees. Since its introduction, the RF has been extremely successful and widely used as a general-purpose classification and regression method (Gagne et al. 2009; Williams 2014; Kusiak and Verma 2011; Deloncle et al. 2007; Gislason et al. 2006). Development of RF was influenced by the work of many on the fields of decision trees, the random subspace method, and random split selection (Amit and Geman 1997; Dietterich 2000; Ho 1998; Bernard et al. 2012). For a more comprehensive review, readers can peruse published materials (Breiman 2001; Dietterich 2003; Criminisi et al. 2011; Svetnik et al. 2003; Prasad et al. 2006; Cutler et al. 2007; Criminisi and Shotton 2013) as well as the freely available online resources (https://www.stat.berkeley.edu/~breiman(RandomForests/cc_home.htm).

Figure A1 (bottom panel) shows a flow diagram depicting the typical steps involved in the RF classification. There are many ways in which the multiple decision trees can be constructed. Each decision tree is grown as follows:

- Let $N$ be the total number of cases in the training dataset. Then create a random sample of cases with $M$ input variables such that ~66% of cases are used and ~33% of cases are left out (out of bag).
- A subset of input variables $m_{try} \leq M$ are selected at random from all the predictor variables. Out of $m_{try}$ variables, one variable that provides the best binary split on a decision tree node is chosen. At the next node, choose another from the $M$ input variables at random and repeat the process until a leaf node is reached. The value of $m_{try}$ is held constant during the forest growing.
- Each decision tree is fully grown and not pruned during the construction.
- In a decision tree, an input is entered at the top and as it traverses down the tree, the data are bucketed into smaller and smaller sets and assigned the class label of the terminal node it ends up in. This procedure is iterated over all the decision trees in the forest, and the average vote of all the decision trees is reported as the RF prediction.

Bagging improves the accuracy of the prediction when random feature selection is used (Breiman 2001). Classification error caused by RF depends on the correlation between any two decision trees and the strength of an individual decision tree in the forest: increasing the correlation increases the classification error, while increasing the strength of the individual trees decreases it. Reducing $m_{try}$ decreased both the correlation and strength of DTs, while increasing $m_{try}$ increases both. This is the only adjustable parameter to which the classification accuracy of the RF is sensitive. The optimal value of $m_{try}$ can be expressed mathematically as $m_{try} = \log_2(M_{tot} + 1)$ or $m_{try} = (M_{tot})^{1/2}$, where $M_{tot}$ is the total number of input variables. The value of $m_{try}$ is then increased or decreased until a minimum prediction error is reached and this particular value of $m_{try}$ can be used to carry out future classifications. RF do not require a separate cross validation or a test to get an unbiased estimate of the classification error (Breiman 2001). When building a decision tree, a bootstrap sample is built for each decision tree and about one-third of the cases are left out of the training set. These left-out data are called out-of-bag (OOB) data and are used to get an unbiased estimate of the classification error (OOB error). Studies show that RF out-of-bag error estimates are generally very close to actual prediction error estimated using new data (Breiman 1996).

The importance of variables in a classification can be studied using the input feature selection available in the
RF code (Breiman 2001). The importance of a particular input variable is estimated by analyzing how much the prediction error increases when a particular variable in the OOB data is permuted randomly while all other variables are left unchanged. To calculate the importance of the $k$th input feature, the value of the $k$th feature is permuted among the OOB data and the difference in the OOB error before and after the permutation over all the decision trees in that forest is estimated. The variable importance measured for the $k$th variable can be expressed mathematically as (Verikas et al. 2011)

$$VI_k = \frac{1}{N} \sum_{n=1}^{N} (R_{n}^{\text{oob}} - R_{n,k}^{\text{oob}}),$$

where $n = 1, 2, \ldots, N$ are the bootstrap samples used and $R_{n}^{\text{oob}}$ and $R_{n,k}^{\text{oob}}$ are the OOB error before and after permuting the $k$th variable, respectively. Generally, input variables that produce large VI values are more important to the classification process than those that generate small VI values.

REFERENCES


