ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 39, DECEMBER 2022, 1994–2007

• Original Paper •

A Machine Learning-based Cloud Detection Algorithm for the Himawari-8 Spectral Image※

Chao LIU[1,2], Shu YANG[1,2], Di DI[*1,2], Yuanjian YANG[1,2],
Chen ZHOU[3], Xiuqing HU[4], and Byung-Ju SOHN[1,5]

[1]*Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters*, *Nanjing University of Information Science & Technology*, *Nanjing* 210044, *China*

[2]*Key Laboratory for Aerosol-Cloud-Precipitation of China Meteorological Administration*, *School of Atmospheric Physics*, *Nanjing University of Information Science & Technology*, *Nanjing* 210044, *China*

[3]*School of Atmospheric Sciences*, *Nanjing University*, *Nanjing* 210046, *China*

[4]*Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites*, *National Satellite Meteorological Center*, *China Meteorological Administration*, *Beijing* 100081, *China*

[5]*School of Earth and Environmental Sciences*, *Seoul National University*, *Seoul* 151747, *South Korea*

(Received 27 October 2020; revised 17 March 2021; accepted 20 April 2020)

ABSTRACT

Cloud Masking is one of the most essential products for satellite remote sensing and downstream applications. This study develops machine learning-based (ML-based) cloud detection algorithms using spectral observations for the Advanced Himawari Imager (AHI) onboard the Himawari-8 geostationary satellite. Collocated active observations from Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) are used to provide reference labels for model development and validation. We introduce both daytime and nighttime algorithms that differ according to whether solar band observations are included, and the artificial neural network (ANN) and random forest (RF) techniques are adopted for comparison. To eliminate the influences of surface conditions on cloud detection, we introduce three models with different treatments of the surface. Instead of developing independent ML-based algorithms, we add surface variables in a binary way that enhances the ML-based algorithm accuracy by ~5%. Validated against CALIOP observations, we find that our daytime RF-based algorithm outperforms the AHI operational algorithm by improving the accuracy of cloudy pixel detection by ~5%, while at the same time, reducing misjudgment by ~3%. The nighttime model with only infrared observations is also slightly better than the AHI operational product but may tend to overestimate cloudy pixels. Overall, our ML-based algorithms can serve as a reliable method to provide cloud mask results for both daytime and nighttime AHI observations. We furthermore suggest treating the surface with a set of independent variables for future ML-based algorithm development.

Key words: cloud detection, machine learning, surface type, Himawari-8, CALIPSO

Citation: Liu, C., S. Yang, D. Di, Y. J. Yang, C. Zhou, X. Q. Hu, and B.–J. Sohn, 2022: A machine learning-based cloud detection algorithm for the Himawari-8 spectral imager. *Adv. Atmos. Sci.*, **39**(12), 1994–2007, https://doi.org/10.1007/s00376-021-0366-x.

Article Highlights:

- Cloud mask algorithms based on machine learning (ML) techniques are developed for AHI onboard the Himiwari-8 satellite.
- The accuracy for cloud detection can be improved by ~5% by appropriately including surface variables.
- Both our daytime and nighttime RF-based algorithms work slightly better than the current AHI operational product.

1. Introduction

Among various atmospheric components, clouds cover almost two-thirds of the Earth and play an important role in the evolution of weather and climate, especially in determining the global radiative budget (Sakaida et al., 2006; Baker

※ This paper is a contribution to the special issue on Cloud–Aerosol–Radiation–Precipitation Interaction: Progress and Challenges.

* Corresponding author: Di DI
Email: didi@nuist.edu.cn

and Peter, 2008; Geoffroy et al., 2008; Dessler, 2010). Satellite-based instruments are one of the most unique and powerful tools to observe global cloud distributions and variations. Cloud detection (i.e., cloud mask product) is an essential step for downstream satellite applications, e.g., cloud and aerosol property retrievals, data assimilation, and further scientific analyses.

For spaceborne spectral radiometers, cloudy and clear pixels can be distinguished by the different solar reflectance or thermal emissions from cloudy and cloud-free atmospheres, e.g., clouds normally result in lower brightness temperatures (BT) in thermal infrared bands and higher reflectance (R) in solar bands (Rossow and Garder, 1993; Ackerman et al., 1998; Saunders and Kriebel, 1988). Thus, threshold and statistical methods are traditional approaches for cloud detection that use multispectral radiometers. Threshold methods are developed by setting thresholds for R, BTs, and BT differences (BTDs) (Saunders and Kriebel, 1988; Key et al., 1990; Wylie et al., 1994; Ackerman et al., 1998), and are widely used for current radiometer operational algorithms. Operational cloud products from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Ackerman et al., 1998; Platnick et al., 2003; Frey et al., 2008; Baum et al., 2012), the Advanced Very High-Resolution Radiometer (AVHRR) (Stowe et al., 1999; Dybbroe et al., 2005) and others, are all based on threshold methods.

Meanwhile, machine learning (ML) techniques have also been widely considered for cloud detection (Visa et al., 1991; Heidinger et al., 2012; Chen et al., 2018). As a typical binary classification involving multivariate analysis, cloud detection is well suited for ML techniques. ML-based algorithms can learn the hidden relationships of different objects, avoiding artificially defined thresholds or conditions for matching the spectral pattern. Supervised ML methods are mostly applied in cloud detection algorithms based on various satellite measurements, such as the Bayesian algorithm (Heidinger et al., 2012; Karlsson et al., 2015), random forests (RF) (Thampi et al., 2017; Zhang et al., 2019; Wang et al., 2020), support vector machine (Ishida et al., 2018), artificial neural network (ANN) (Hughes and Hayes, 2014; Chen et al., 2018), and others. Those models generally use observed radiative variables and their combinations as the input data for training. Image data have also been considered for cloud detection, which account not only for spectral information, but also cloud textural characteristics (Bai et al., 2016; Le Goff et al., 2017), and have yielded favorable results as well. With the advantages of onboard active instruments, some studies use collocated lidar or radar observations as reference labels to determine pixel cloudiness more accurately, e.g., the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) (Heidinger et al., 2012; Wang et al., 2020) and Cloud Profiling Radar (Gomis-Cebolla et al., 2020). Meanwhile, theoretical results that can provide full atmospheric conditions and corresponding simulated radiances have also been used as efficient training datasets for ML-based model development (Chen et al., 2018). Due to its importance for satellite applications and climate prediction, accurate cloud detection is a worthy area for future research.

Apart from cloud spectral characteristics, the nature of the Earth's surface is another key factor that should be considered in algorithm development because of the significant spatiotemporal variations in surface albedo and emission properties (Platnick et al., 2003). Currently, most ML-based cloud detection models develop independent classifiers for different surface types, e.g., Heidinger et al. (2012) and Wang et al. (2020). Some studies consider the surface by using additional variables as input features (Poulsen et al., 2020), which calls to question how surface features could be most efficiently parameterized. Thus, it is still an open issue as to how to eliminate the influences of surface differences on cloud detection algorithms and ultimately how to fully optimize the cloud mask product.

The Advanced Himawari Imager (AHI) onboard the Himawari-8, a new-generation geostationary satellite launched by the Japan Meteorological Agency, has high temporal and spatial resolutions across 16 spectral bands. New cloud detection algorithms particularly designed for the AHI are needed because there are significant differences between the AHI and other instruments concerning band characteristics. For example, the strong water vapor absorption band centered around 1.38 μm which is useful for thin cirrus detection is not included in AHI. Central wavelengths of some bands sensitive to cloud properties are also modified, e.g., there is an AHI 2.25 μm band that used to be centered around 2.13 μm in previous instruments (Wang et al., 2018). The differences in band spectral response functions also can lead to incomparable measurements from bands with similar central wavelengths.

The Japanese Meteorological Satellite Center team develops a threshold method for the AHI operational cloud mask product (Imai and Yoshida, 2016), and a hit rate of ~0.85 is reported when comparing with MODIS products. As one of the original algorithms for AHI, it has been found that the current AHI cloud mask product may slightly overestimate cloudy pixels compared to MODIS product, i.e. ~25% of MODIS-determined clear pixels may be misclassified as cloudy ones by the AHI product (Lai et al., 2019). However, such evaluations may be biased due to the uncertainties related to the MODIS product, so the performance of the current AHI cloud mask is neither entirely well-known nor perfect enough. Furthermore, the product is only available for daytime cloud detection as solar band observations are used.

Consequently, this study intends to develop ML-based cloud detection algorithms for AHI observations that can provide cloud mask products for both daytime and nighttime observations with a particular focus upon the treatment of the surface. The remainder of the manuscript is organized as follows. Section 2 describes the satellite data and the development of the ML-based methods. The performance of the algorithms is evaluated and discussed in section 3, and section 4 summarizes this study.

2. Data and Methods

The flowchart in Fig. 1 illustrates the general structure of this work and the way pre-processing and prediction steps are carried out. During the pre-processing stage, ML-based models will be trained and tested using different features, ML techniques, and parameters, which will be detailed in this section. Then, an optimal model will be selected and used for predictions. The model evaluations will be presented in section 3.

2.1. *Satellite datasets*

As a member of the Multifunction Transport Satellite series, the geostationary Himawari-8 satellite was launched on 7 October 2014 and is located above 140.7°E for observation of Earth's surface, atmospheric moisture, clouds, and the environment (Bessho et al., 2016; Shang et al., 2017; Wang et al., 2018; Letu et al., 2020). The AHI is an important instrument onboard Himwari-8 and provides images at 16 spectral bands with central wavelengths ranging from 0.46 to 13.3 μm, with a temporal resolution of 10 minutes for full-disk and spatial resolutions ranging from 0.5 to 2.0 km (Min et al., 2017; Wang et al., 2019).

CALIOP onboard CALIPSO is a two-wavelength polarization-sensitive lidar that provides continuous measurements on vertical cloud and aerosol structures (Stephens et al., 2002; Winker et al., 2007). The collocated CALIOP Level 2, 1 km cloud layer product provides a reliable assessment of cloud mask labels for our model training and validation (Heidinger et al., 2012). After collocation, if there are one or more layers of clouds in a CALIOP pixel, it is considered to be a cloudy one. Otherwise, the pixel is defined as a clear one. We find all CALIOP pixels (1 km resolution) within the collocated AHI pixel (spatial resolution of 5 km) with an observational time difference of fewer than 10 minutes, and only homogeneous AHI pixels (i.e., all collocated AHI pixels being cloudy or clear ones) are considered in the training dataset to ensure its quality for training.

A collocated AHI and CALIOP dataset spanning the entirety of 2017 is built, and over 420 000 AHI pixels are labeled. We randomly separate the dataset into a training (70%) and a testing (30%) subset, and the latter is used to tune and to find the optimal ML model parameters. This particular training and testing split is widely used for ML algorithm development to avoid overfitting. CALIOP VFM product and MODIS cloud mask products will also be considered for comparison and evaluation in section 3.

2.2. *Feature selection*

Features, i.e., input parameters for cloud detection, can be chosen by considering the characteristics of different spectral bands with respect to cloudy or clear atmospheres. The analysis of previous cloud mask algorithms (either threshold-based or ML-based models) provides us with the best suggestions on possible feature selection. Table 1 gives examples of features used by recent cloud mask or cloud classification algorithms. Aside from direct radiative variables such as *R*, BTs, and BTDs, auxiliary data such as observational geometries, geolocation information, and surface properties are also used as input datasets. Most of the predictors are chosen from those considered by previous cloud detection algorithms and these predictors have physical support. For example, window band BT (11.2 μm) normally represents cloud top temperature and is one of the most important and widely-used chan-

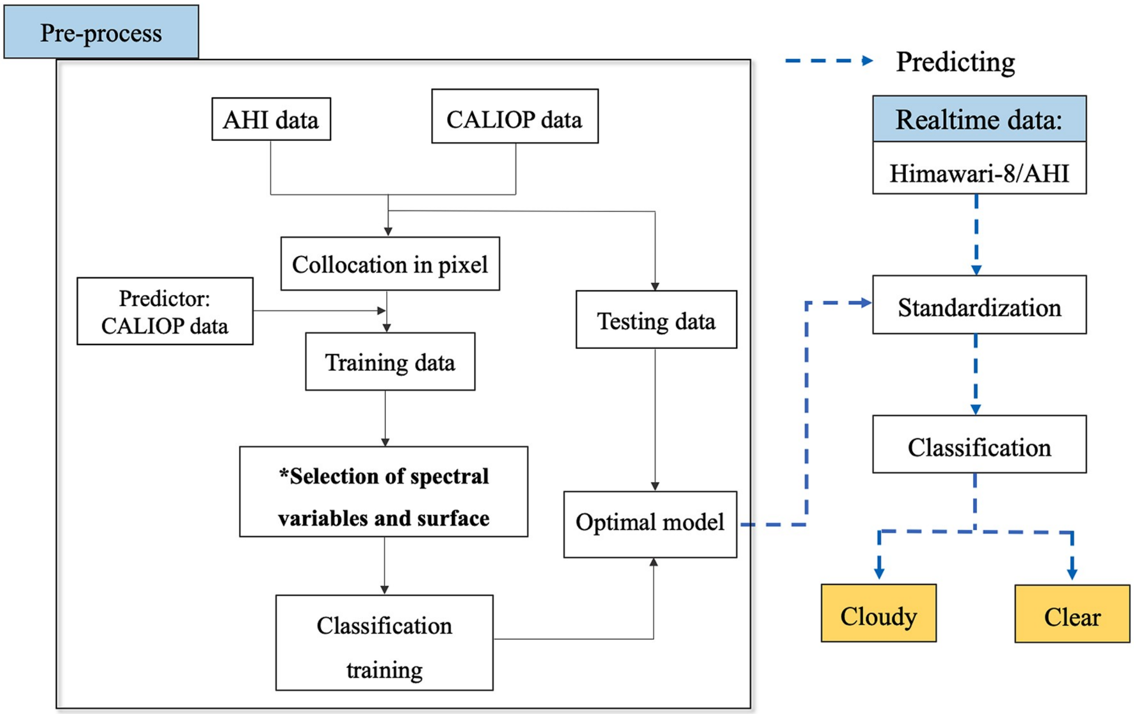

Fig. 1. Flowchart of the ML-based cloud detection algorithm development and prediction.

Table 1. Comparison for some recent ML-based cloud detection and classification algorithms for spectral radiometers.

References	Feature parameters	Auxiliary parameters	Satellite
Lyapustin et al., 2008	*R* (0.64 μm), *R* (0.47 μm), *R* (0.55 μm), *R* (0.86 μm), *R* (1.24 μm), *R* (2.11 μm), BT (11.03 μm)	No	MODIS
Chen et al., 2018	*R* (0.47 μm), *R* (0.55 μm), *R* (0.66 μm), *R* (0.86 μm), *R* (1.24 μm), *R* (2.13 μm)	SZA, VZA, RAZ, Surface elevation	MODIS
Zhang et al., 2019	*R* (0.64 μm), BT (3.85 μm), BT (7.35 μm), BT (8.6 μm), BT (11.2 μm), BT (12.35 μm), BTD (11.2–3.85 μm), BTD (11.2–7.35 μm), BTD (11.2–8.6 μm), BTD (11.2–12.35 μm)	VZA, Ts, Lat, Lon	AHI
Gomis-Cebolla et al., 2020	*R* (0.64 μm), *R* (0.47 μm), *R* (0.45 μm), *R* (0.86 μm), *R* (2.13 μm), *R* (1.38 μm)	No	MODIS
Wang et al., 2020	*R* (0.86 μm), *R* (1.24 μm), *R* (1.38 μm), *R* (1.64 μm), *R* (2.25 μm), BT (8.6 μm),BT (11 μm), BT (12 μm)	VZA, Ts, Lat, Lon	VIIRS

*VZA=View Zenith Angle, SZA= Solar Zenith Angle, RAZ= Viewing zenith angle, *T*s=Surface skin temperature, Lat=Latitude, and Lon=Longitude.

nels to distinguish cloudy and clear pixels (Strabala et al., 1994). Saunders and Kriebel (1988) use BTD (11–12 μm) to detect cirrus clouds because BTDs over clouds are greater than those in the absence of clouds. The low cloud test BTD (3.9–11 μm) is based on the differential absorption of water and ice cloud particles between these wavelengths (Platnick et al., 2003). This study uses only radiative variables (and surface characteristics) as input parameters to avoid possible influences due to collocation, noting that only local afternoon observations are considered because of CALIOP passing time.

Cloud detection algorithms with and without solar band observations are developed as daytime and nighttime versions, respectively. The daytime algorithm refers to the algorithm with solar band reflectance considered, so it is only available during the local daytime. The nighttime algorithms that exclude solar-related parameters can be used for all-time observations. For a fair comparison during algorithm development and validation, all collocated AHI and CALIOP observations are from local daytime. In this way, we used the same dataset for the daytime and nighttime algorithm, and the two algorithms differ on whether the solar band reflectance was used. The following brightness temperatures are considered for both algorithms: BT (3.85 μm), BT (7.35 μm), BT (8.6 μm), BT (11.2 μm), BT (12.35 μm), BTD (3.85–11.2 μm), BTD (11.2–7.35 μm), BTD (8.6–11.2 μm), and BTD (11.2–12.35 μm). The solar band reflectance channels include: *R* (0.64 μm), *R* (0.86 μm), *R* (1.61 μm), and *R* (2.25 μm), and are solely used in the daytime model.

2.3. *Surface treatments*

As mentioned above, the surface is a special but important variable influencing cloud detection. Clear desert pixels might be erroneously detected as cloudy in the daytime due to the higher albedo and emissivity of desert sand (Ackerman et al., 1998), so most algorithms develop independent models for different surface conditions. To better eliminate the negative impact of surface features on cloud detection, three different methods are introduced to treat the surface. Assume that there are *N* surface types (ST), referred to as ST_1, $ST_2 \ldots ST_N$. The first model (Model #1) develops separated ML models for each surface type, and *N* independent models will be achieved, each of which handles only observations from the particular surface type. Model #2 adds an input parameter, i.e., surface type, as a new feature, and each type of surfaces is specified by an integer from 1 to *N*. In this way, only one ML-based model is needed for all observations, but the integers may misrepresent the physical differences among different surfaces. To avoid such misrepresentation by using a single integer, Model #3 is similar to Model #2, but adds *N* binary parameters, i.e., an additional parameter (a binary variable) for each surface type. To be more specific, if the observation is over the *n*th surface, its *n*th surface variable will be defined as one, while all others are zero. The three models differ only on how the surface types are considered, further noting that all radiative features, ML models, and observational datasets are kept the same.

Figure 2 illustrates the structures of the three models. Each quadrangle in the figure represents a ML-based algorithm, and Model #2 and Model #3 also illustrate how surface variables are defined. Considering the coverage of AHI, this study considers four surface types, i.e., ocean, forest, land and desert, and the MODIS Land Cover Climate Modeling Grid Product from MODIS Collection 6 annual surface type product MCD12C1 is considered (Loveland and Belward, 1997; Sulla-Menashe and Friedl, 2018). There are fewer observations over ice or snow surfaces in the covered area, so we have not yet included them in our model.

2.4. *Machine learning technologies*

Two popular supervised ML methods are considered, ANN and RF (Swami and Jain, 2013), because their performances have been well justified (Chen et al., 2018; Gomis-Cebolla et al., 2020; Wang et al., 2020). We pay more attention to the construction of the algorithms, e.g., preparation of the training dataset, feature selection, and surface treatment, as opposed to the particular ML techniques, because the latter is responsible for fewer differences in the results. Thus, we only consider ANN and RF in this algorithm, and

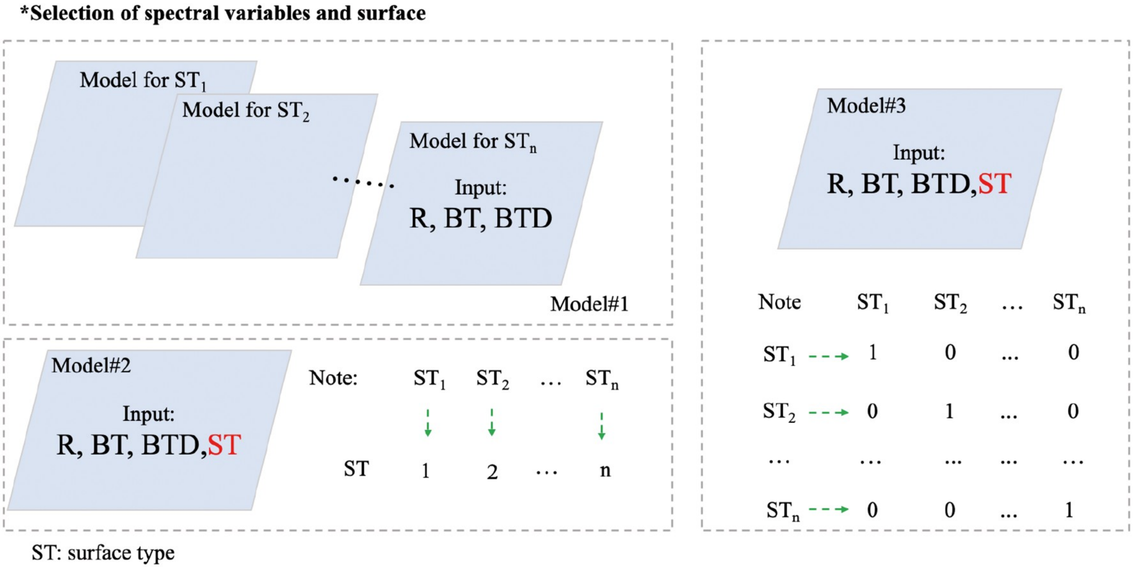

Fig. 2. Schematic diagram of the three models for the treatments for observations over different surfaces.

others may be tested similarly in future studies.

ANN is a multilayer perceptron model consisting of an input layer, a hidden layer, and an output layer by seizing the non-linear relationship between input and output variables. As a simple binary classification problem, we consider only one hidden layer, and the sigmoid function is used as the activation function. The neuron number is chosen between 5 and 100 in steps of five to find the optimal value. RF considers an ensemble of decision trees and uses bagging to train the model (Breiman, 2001). Two important parameters in the RF model are the number of trees and the maximum depth of the tree. We test the number of trees varying from 100 to 500 in steps of 100, and the maximum depth from 10 to 50. Table 2 lists the tuning parameters for the two ML models.

The optimal parameters of the ML algorithms are obtained by grid searches. Here, we define “accuracy” as the ratio of the number of pixels (samples), which are correctly detected by our algorithm (according to CALIOP results), to that of the total pixels. Figure 3 gives the accuracy values of the ANN and RF algorithms with different parameters. The accuracies for the best ANN daytime and nighttime model are 0.88 and 0.80, respectively. The accuracy of the best daytime model for RF is as large as 0.94, and that for the nighttime model is 0.87. Evidently, the algorithm performance is not significantly sensitive to the model parameters with the accuracy generally varying around approximately 0.03. For the daytime algorithm, the best neuron node parameter for ANN is 11, and the best ntrees and mdepth parameters for RF are 200 and 20, respectively. Larger mdepth may lead to overfitting, so two relatively small parameters are used to guarantee the robustness of the models (Scornet, 2018). For the nighttime model, the best neuron node parameter for ANN is eight, and the best ntrees and mdepth parameters for RF are 100 and 10.

Through feature selection, the contribution of each feature to the algorithm is calculated. For ANN, a “f_classif score” is obtained based on the analysis of variance. The higher the score of an interest field, the more the feature contributes to the cloud detection. For the RF algorithm, the importance of a feature can be illustrated by the “mean decrease gini”, and larger ”mean decrease gini” values correspond to features that are more “useful” for the detection.

Table 2. Contingency matrix of evaluation of cloud detection results by comparing with CALIOP results.

Scenario	AHI cloudy	AHI clear
CALIOP cloudy	TP	FN
CALIOP clear	FP	TN

Figure 4 shows the f_classif score and mean decrease gini for the two models to demonstrate the feature importance. The tests are performed using the entire training dataset including observations from all surfaces. The six most influential parameters in the daytime ANN algorithm are BTD (11.2–7.35 μm), BT (12.25 μm), BT (11.2 μm), BT (8.6 μm), BTD (3.85–11.2 μm), and *R* (0.64 μm). For the RF-based algorithm, BTD (11.2–7.35 μm), BT (12.25 μm), BTD (3.85–11.2 μm), BT (11.2 μm), *R* (0.64 μm), and BTD (11.2–8.6 μm) are the six more important inputs. Clearly, the physically important bands and combinations all rank relatively high here. It should also be noticed that the two water vapor bands, i.e., BT (7.35 μm) and BT (3.85 μm), contribute less to the two ML-based algorithms. Meanwhile, we have also considered geolocation and solar-viewing geometries for tests, and their contributions are relatively limited. Thus, we retain only radiative information in the model.

Figure 5 gives the feature contribution under three surface models based on the RF algorithm. Figures 5a–5d are feature contributions of four specific surface types based on Model #1, and Fig. 5e and Fig. 5f are for Model #2 and

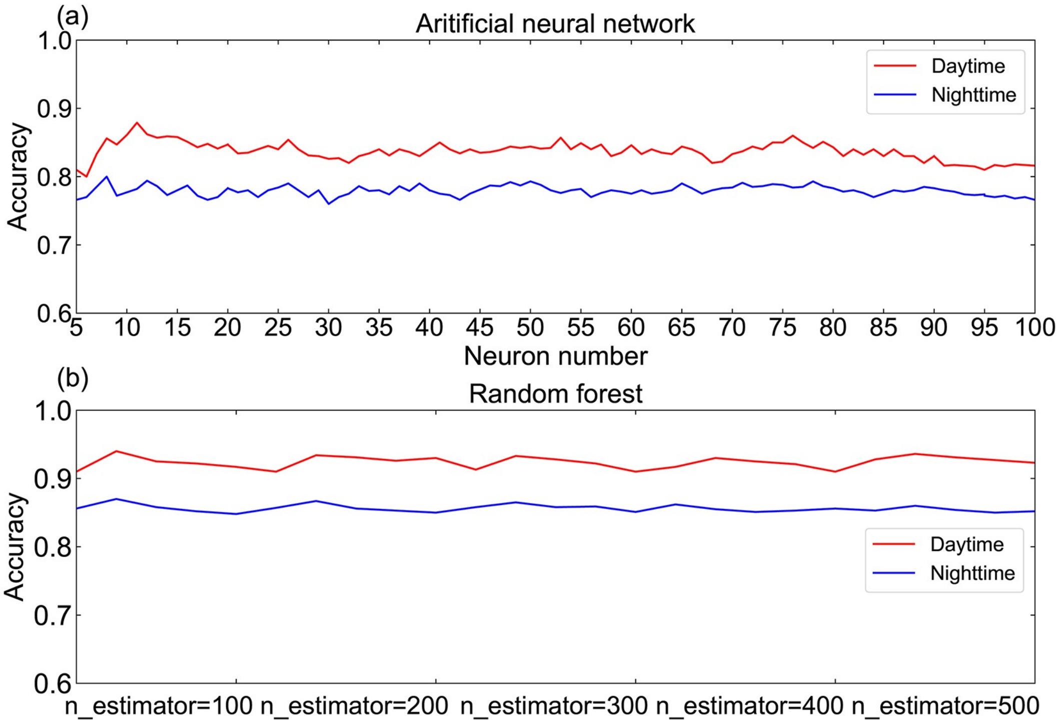

Fig. 3. Accuracy scores for the ANN and RF algorithms with different parameters.

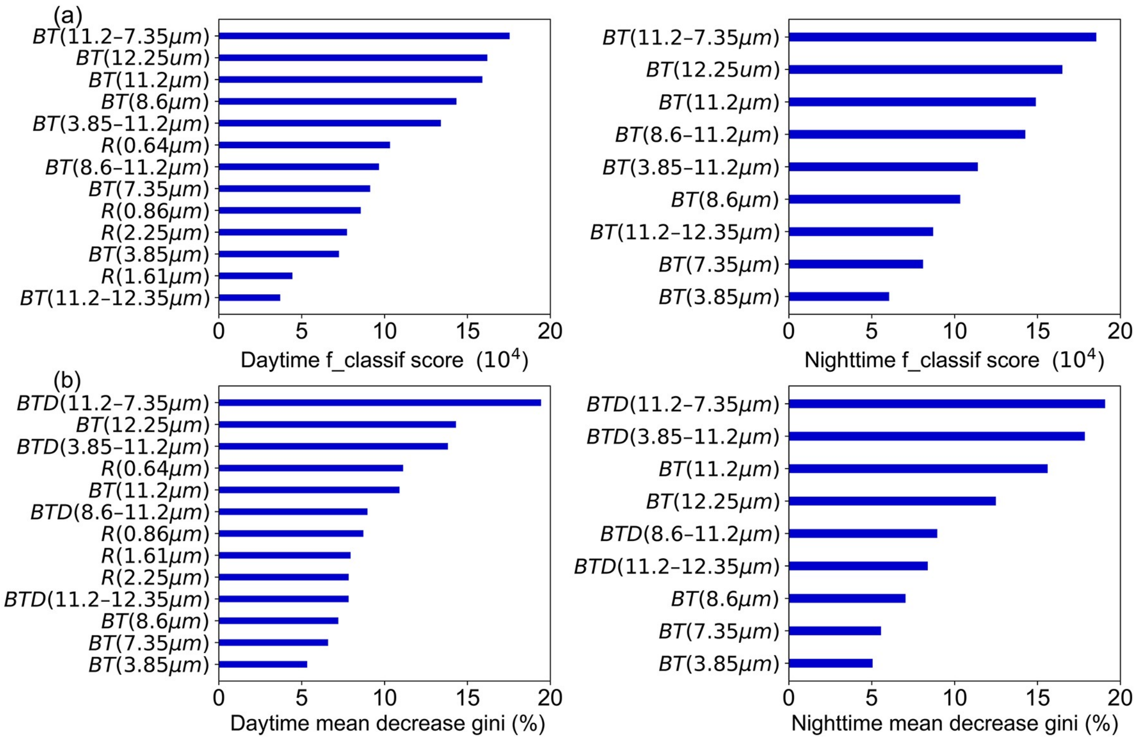

Fig. 4. F_classif scores (top panels) in the ANN algorithm and mean decrease gini (bottom panels) in the RF algorithm for both daytime (left) and nighttime (right) algorithms.

Model #3, respectively. For ocean surfaces, both solar and infrared bands, e.g., BTD (3.85–11.2 μm), *R* (0.64 μm), and BTD (11.2–7.35 μm), correspond to higher feature importance values, and, for other surfaces, longwave infrared BT differences such as BTD (11.2–7.35 μm), BTD (11.2–12.35 μm), and BTD (11.2–8.6 μm) are more important. For Model #2 and Model #3, the surface variables don't rank high (not in the top six), but clearly show different impacts on the models by modifying the orders of the radiative parameters. The relative performances of the three surface models

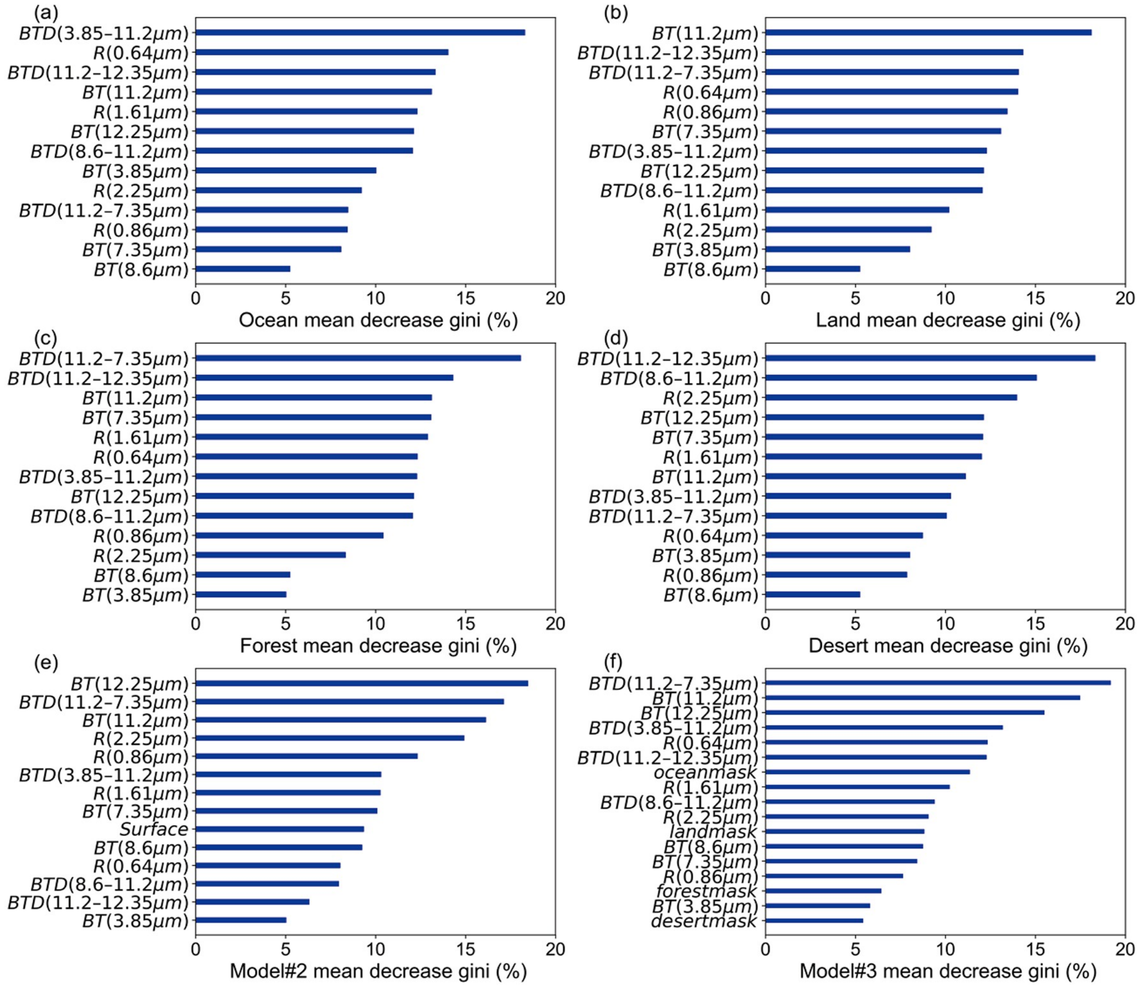

Fig. 5. Mean decrease gini (bottom panels) in the RF algorithm for different surface models: (a–d) for Model #1 with four surface types, (e) for Model #2, and (f) for Model #3.

will be evaluated in the following section.

It is worth mentioning that CALIPSO is in the afternoon orbit, so the collocated AHI and CALIOP observations include only local afternoon data, which could bring uncertainties for the all-time cloud detection algorithm. To avoid such influences, our algorithms consider only direct radiance observations, auxiliary information such as pixel position and viewing geometries (especially solar zenith) are not included. In this way, the algorithms would be less dependent on the time of observation, viewing geometries, or the spatiotemporal distributions of the clouds.

3. Results and discussion

With the optimal algorithms for both ANN and RF determined, we quantitatively evaluate them as well as the AHI operational product by comparing their results with active CALIOP observations and MODIS results. To keep the dataset independent, we collocate AHI and CALIOP observations from the first five days of each month in 2018. CALIOP determined cloudy pixels are defined as "positive" events, and clear ones are marked as "negative" events. Again, we refer to the CALIOP results as the truth. Then, "true positive (TP)" means that CALIOP and our ML-based algorithms (or the AHI operational product) consistently detect a pixel as a cloudy one, and "true negative (TN)" refers to pixels that are detected as clear ones by both instruments. "False positive (FP)" corresponds to pixels that are detected as clear by CALIOP but as cloudy by the AHI algorithms, and "false negative (FN)" is defined similarly but for pixels recognized as cloudy and clear by CALIOP and AHI respectively. These definitions of TP, TN, FP, and FN can be better understood by Table 2. Then, the performance of algorithms can be generally represented with two indices, the true positive rate (TPR) and false positive rate (FPR), and they are given by:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{2}$$

TPR is also called sensitivity and shows the fraction of cloudy pixels correctly identified by our algorithms or operational product. FPR is known as specificity, which shows the fraction of clear pixels that are not identified. A better algorithm can be defined by that with a TPR closer to 1, and a FPR closer to 0. It is noticed that for similar variables such as the relative operating characteristic, the Pierce skill score has also been widely used for model evaluations, which results in similar conclusions, so we will just keep the problem simple by using TPR and FRP.

Figure 6 shows the TPR and FPR values of the AHI operational product and our daytime ML-based algorithms for all surface results [panel (a)] as well as for the four specific surface types [panels (b)–(e)]. Again, markers closer to the upper left corner represent better algorithm performance. The TPR values are mostly over 0.9, and the FPR values are under 0.2, indicating a slight overestimation of cloudy pixels. As can be seen from Fig. 6a, the ANN-based results are better than those of the operational results while they are not as good as the RF-based ones. The RF Model #3 results are the best among all algorithms with a TPR of 0.97 (all surface types). For most surface types, our ML-based algorithms, especially the RF-based ones, are better than the AHI operational product. For results over oceans, the AHI operational product and our algorithm achieve similar TPR and FPR values with differences less than 0.02. This shows that the surface parameters screening into the model in the form of independent binary variables is helpful to improve the ML algorithm performance.

Figure 7 is similar to Fig. 6 but for the nighttime algorithms. As expected, the nighttime algorithms do not perform as well as the daytime ones but only by a slight margin, i.e., TPRs are approximately 0.03 smaller. Surprisingly, the nighttime RF-based algorithms also result in a larger TPR and a smaller FPR compared to the AHI product, noting once again, that the AHI is based on a daytime threshold method with solar band tests. Overall, both our daytime and nighttime algorithms improve the accuracy for detection and reduce the false identification rate.

To understand the ML-based algorithm performance in detail, Figure 8 illustrates an example of cloud mask results from the AHI and MODIS operational products as well as our daytime RF-based algorithms. The observation is taken at 0510 UTC 4 June 2018. The top panels show the AHI RGB image, the AHI operational cloud mask (Bessho et al., 2016), and the MODIS MYD35 cloud mask (Ackerman et al., 1998). Each pixel in the AHI and MODIS operational products is classified into one of the four categories, i.e., confident clear, probably clear, probably cloudy, and cloudy. In our investigation, the probably cloudy and probably clear pixels are recognized as cloudy and clear ones respectively. The blue and gray colors represent clear and cloudy pixels, respectively. If the MODIS cloud mask is understood as the reference, the AHI product tends to underestimate clear sky pixels, and this agrees with its large FPR values in Fig. 6. Because the RF-based algorithms work better than the ANN-based ones, only RF-based results are presented in the bottom panels. Generally, the results from the three RF-based algorithms agree well with each other and are consistent with the MODIS cloud mask (the top center panel).

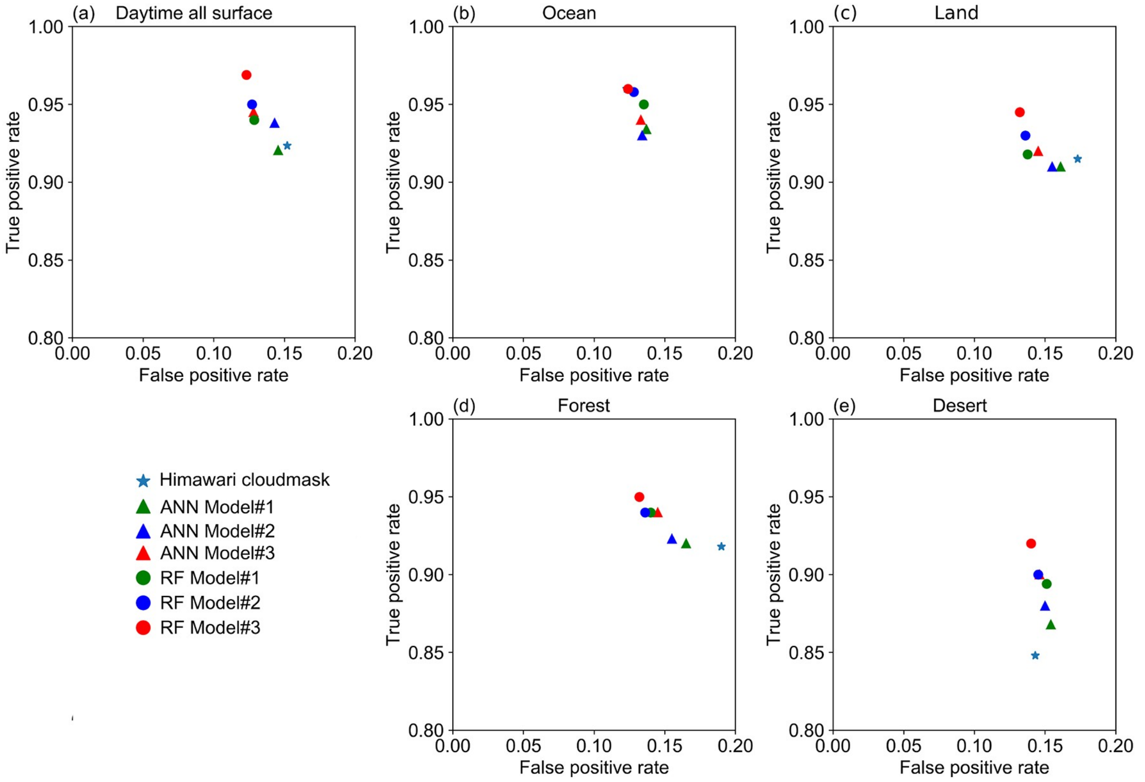

Fig. 6. Comparison of the true positive rate (TPR) and false positive rate (FPR) of the two ML-based daytime algorithms and the AHI operational product, (a) for all surface results and (b–e) for the four different surface types.

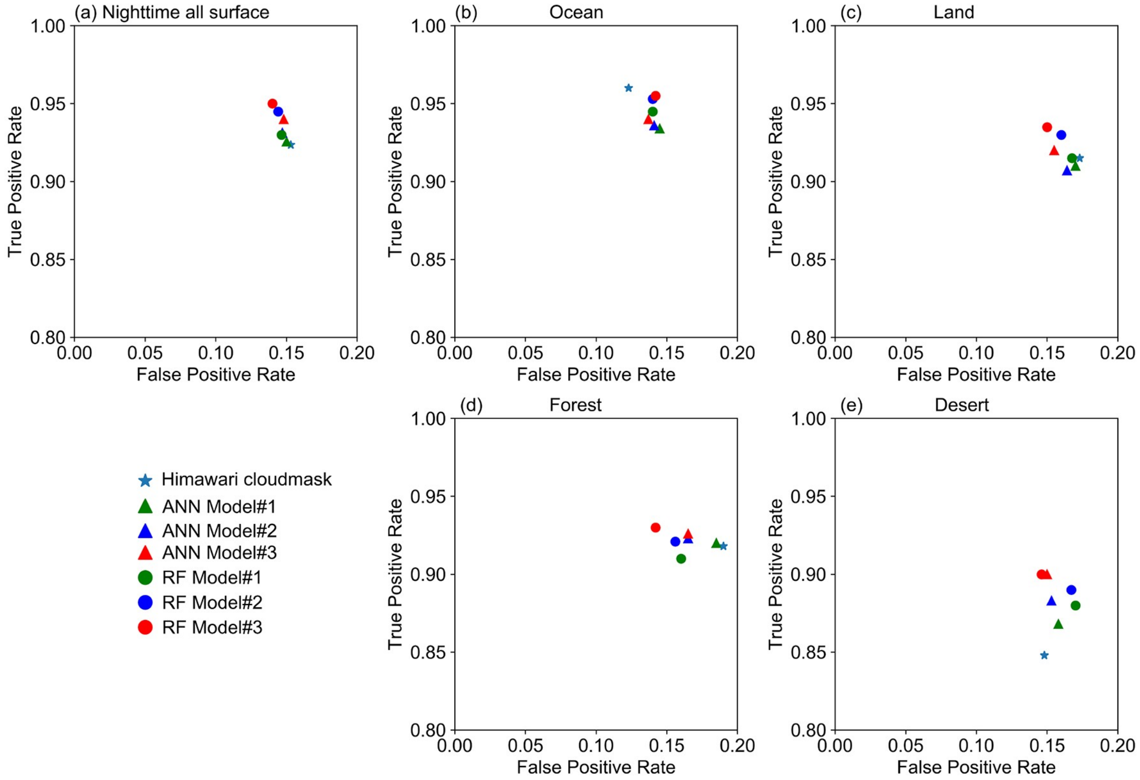

Fig. 7. Same as Fig. 6 but for nighttime algorithms.

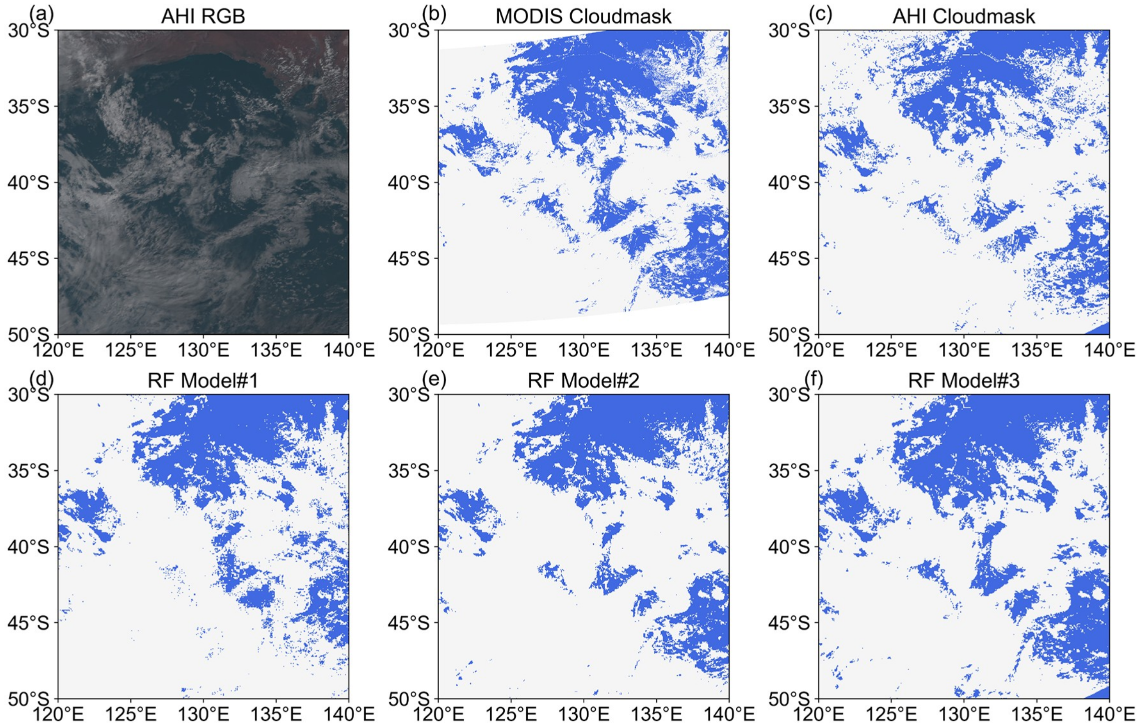

Fig. 8. An example for the MODIS and AHI operational cloud product and our results for the scene at 0510 UTC 4 June 2018, (a) RGB image, (b) MODIS cloud mask, (c) AHI cloud mask, and (d–f) RF-based results. Cloudy pixels are marked by gray, and clear ones are marked by blue.

Figure 9 directly illustrates the differences between the MODIS/AHI operational results and those from our algorithms, and the upper and lower panels use MODIS and AHI results as references, respectively. In Fig. 9, the green pixels correspond to those identified as cloudy ones by the MODIS/AHI operational product but as clear ones in our

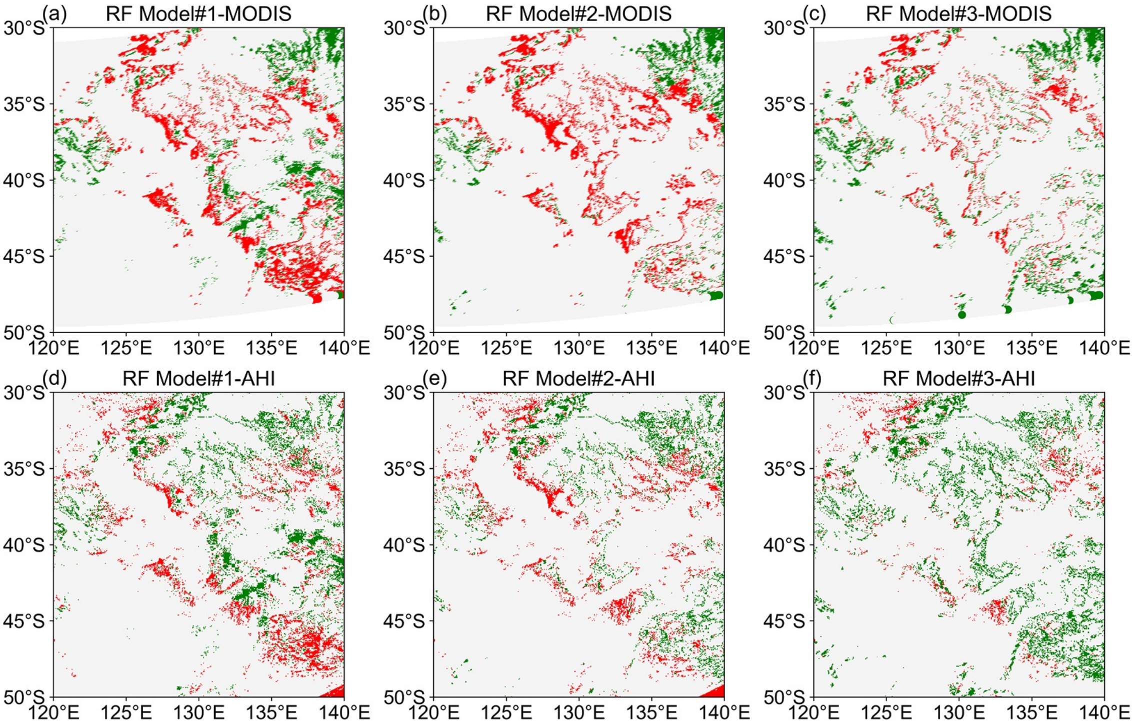

Fig. 9. Differences between our cloud detection results and MODIS/AHI operational results. The upper panels use the MODIS results as references, and the lower panels use AHI results as references. Green represents pixels that were detected as clear ones by our method but cloudy ones by MODIS or AHI, and red represents pixels that were detected as cloudy ones by our method but clear ones by MODIS or AHI.

ML-based algorithms, and the red represents pixels that are identified as clear ones by the MODIS/AHI but cloudy ones by our algorithms. In short, the gray regions indicate pixels correctly identified by our ML-based algorithms, while blue or red ones are misclassified ones. Compared with the MODIS product, the RF Model #1 and Model #2 slightly overestimate the cloudy pixels, whereas our Model #3 shows much fewer misclassified pixels. Compared with the AHI operational product, the RF-based Model #1 overestimates cloudy pixels more than the other two methods. Model #2 and Model #3 overestimate clear pixels, especially in the region around ~ 45°S and 135°–140°E. The results suggest that the current AHI operational algorithm erroneously detects more cloudy pixels, and our ML-based algorithms, especially Model #3, overcome this overestimation. For Model #3, disagreements are mostly noticed around cloud edges, and this may be caused by collocation errors and cloud movement or development. Thus, the agreement between our RF-based algorithm and MODIS results should be even better than presented.

An example is given in Fig. 10 for the nighttime algorithms, and the results are based on observations at 2000 UTC 3 November 2018. Since there is no AHI operational cloud mask for nighttime observations, only the MODIS product is compared. The infrared cloud image in the AHI 11.2 μm band is given in Fig. 10a. Overall, our ML-based results agree well with the MODIS results, and Model #1 seems slightly better for this case. The red boxes in the figure illustrate regions, over which our results disagree more significantly with the MODIS results. Figure 11 is similar to Fig. 9 for the relative differences. Massive red pixels that appear in Fig. 11 indicate that our nighttime algorithms tend to overestimate cloudy pixels as well. However, with possible misclassifications from MODIS itself, those differences cannot entirely be attributed to the problem of our algorithms.

To better evaluate the results with the "truth" determined from active lidar observations, Fig. 12 compares our results with the collocated CALIOP product. The CALIOP VFM product shows the vertical profiles of the atmosphere, and various types of cloudy conditions are depicted. We present two examples from 2018. Again, our results are consistent with the CALIOP and MODIS products. Specifically, in the low latitudes (13°–17°N), some pixels are misidentified as cloudy ones in the AHI operational product, but most of them are correctly detected by our algorithms.

Last, Fig. 13 shows a full disk comparison to show the consistency of our algorithms for large-scale cloud coverage, along with some details. Our algorithm performs reasonably well in detecting large-scale cloud coverage. Due to the absence of solar bands at local night (black regions in Figs. 13b and 13c), the AHI product and our daytime results miss a small region on the disk, whereas our nighttime algorithm can still provide reasonable cloud masks.

In summary, when compared to both active CALIOP and classic MODIS results, our RF-based algorithms, with the surface treated in the form of independent binary variables, provide the most reliable cloud mask results for the AHI observations. We further note that both the daytime

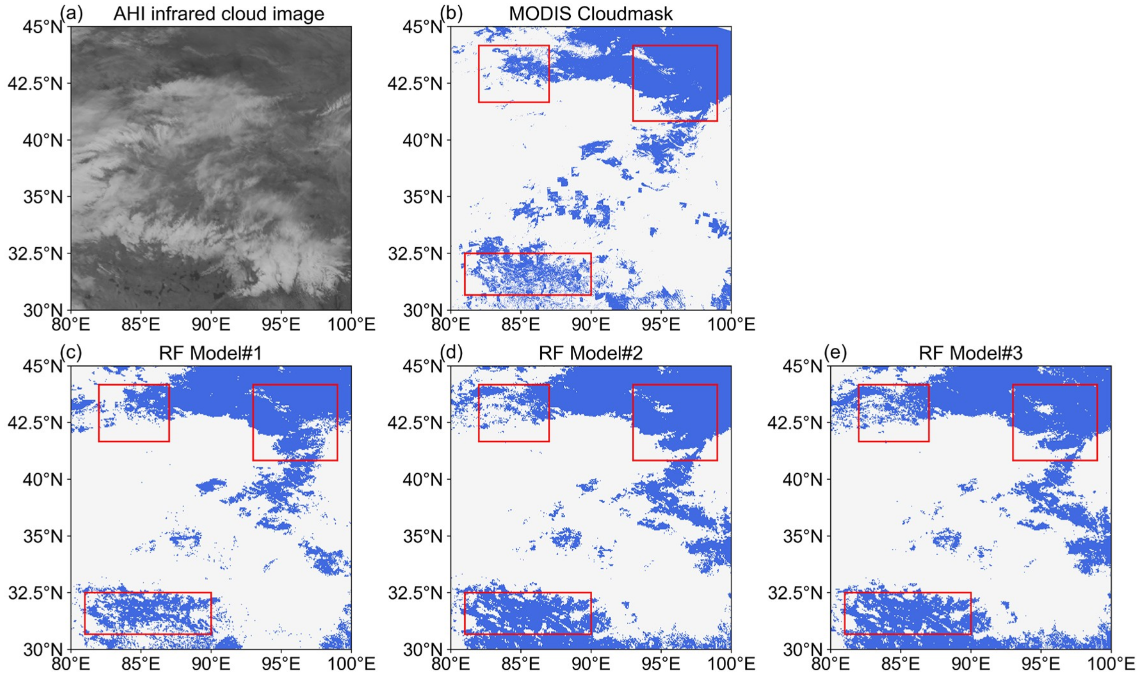

Fig. 10. Same as Fig. 8 but for ML-based nighttime algorithm results and a sense at 2000 UTC 3 November 2018.

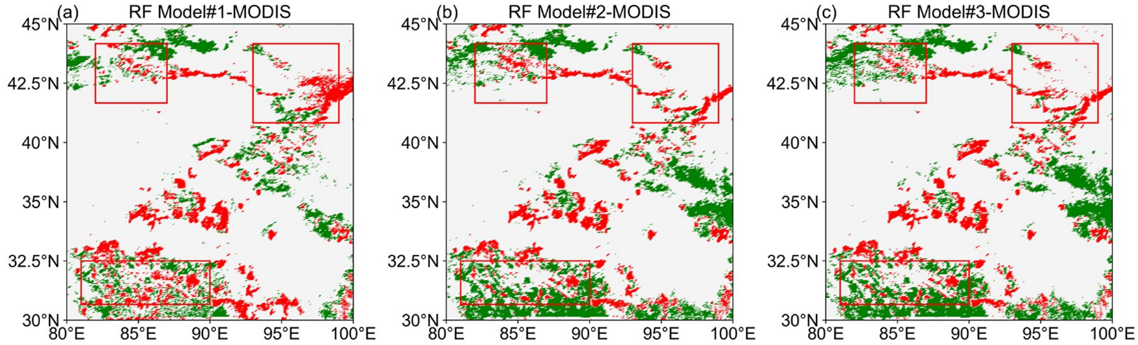

Fig. 11. Same as Fig. 9 but for differences of Fig. 10 results.

and nighttime algorithms outperform the current AHI threshold method-based results.

4. Conclusions

This study develops ML-based algorithms to distinguish between cloudy and clear pixels for AHI observations. The training, testing, and validating datasets for our algorithm development are from the collocated CALIOP product, and only radiometer-based radiative and surface variables are used as input for the ML prediction. Two ML-based algorithms are separately developed, referred to as the daytime and nighttime models, which differ on whether the solar band observations are included. Meanwhile, this study pays special attention to eliminating influences of different surface types on cloud detection and finds that optimizing the elimination of surface influence is an important consideration and an efficient strategy to further improve the model. Overall, the RF-based algorithms outperform the current AHI operational product by improving the TPR by ~5% and reducing the FPR by ~3%. Such advantages are achieved through the application of the ML model, including careful preparations of the training dataset, special treatments for surfaces, and more careful considerations of radiative observations and their combinations.

Future research will focus on further optimization for our algorithm. We would consider methods that use additional auxiliary information, such as latitude, longitude, and surface albedo which may be worked upon with a much larger training dataset. The ML classification algorithms, such as deep learning algorithms, may also be considered in the future. Meanwhile, our method may be easily applied to cognate instruments, such as the Advanced Geosynchronous Radiation Imager (AGRI) onboard China's geostationary satellite Fengyun-4A. In conclusion, as cloud detection is one of the most fundamental products for satellite applications, our mod-

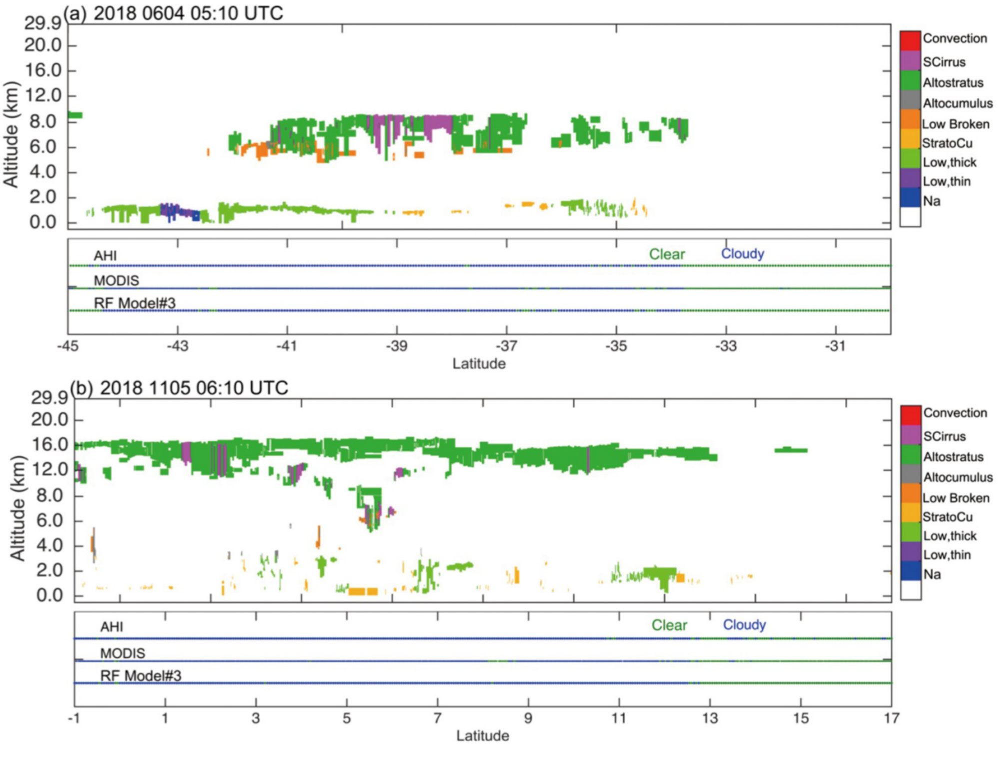

Fig. 12. Two examples of AHI, MODIS, and our results compared with the CALIPSO level 2 VFM product.

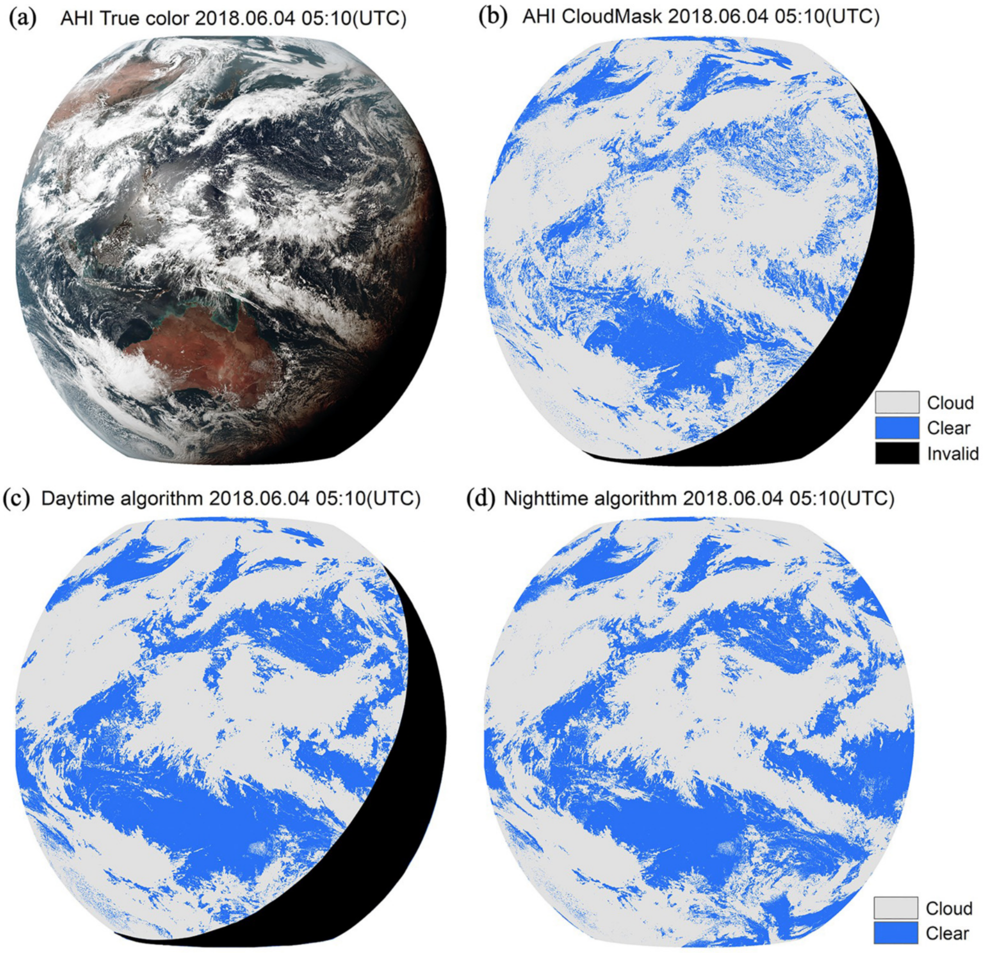
(a) AHI True color 2018.06.04 05:10(UTC)
(b) AHI CloudMask 2018.06.04 05:10(UTC)
Cloud
Clear
Invalid
(c) Daytime algorithm 2018.06.04 05:10(UTC)
(d) Nighttime algorithm 2018.06.04 05:10(UTC)
Cloud
Clear

Fig. 13. A full disk comparison among AHI operational cloud mask (upper right) and our ML-based results.

els will definitely benefit downstream applications such as remote sensing, data assimilation, and climate studies.

Acknowledgements. We thank Drs. Letu HUSI and Huazhe SHANG for valuable discussions and help on accessing the AHI cloud products. This study is financially supported by the National Key Research and Development Program of China (Grant No. 2018YFC1506502), National Natural Science Foundation of China (Grant No. 41975025), Natural Science Foundation of Jiangsu Province (Grant No. BK20190093), Defense Industrial Technoloy Development Program, and "Qinlan" program of Jiangsu Province.

REFERENCES

Ackerman, S. A., K. I. Strabala, W. P. Menzel, R. A. Frey, C. C. Moeller, and L. E. Gumley, 1998: Discriminating clear sky from clouds with MODIS. *J. Geophys. Res.*, **103**, 32141−32157, https://doi.org/10.1029/1998JD200032.

Bai, T., D. R. Li, K. M. Sun, Y. P. Chen, and W. Z. Li, 2016: Cloud detection for high-resolution satellite imagery using machine learning and multi-feature fusion. *Remote Sens.*, **8**, 715, https://doi.org/10.3390/RS8090715.

Baker, M. B., and T. Peter, 2008: Small-scale cloud processes and climate. *Nature*, **451**, 299−300, https://doi.org/10.1038/nature06594.

Baum, B. A., W. P. Menzel, R. A. Frey, D. C. Tobin, R. E. Holz, S. A. Ackerman, A. K. Heidinger, and P. Yang, 2012: MODIS cloud-top property refinements for Collection 6. *J. Appl. Meteorol. Climatol.*, **51**, 1145−1163, https://doi.org/10.1175/JAMC-D-11-0203.1.

Bessho, K., and Coauthors, 2016: An introduction to Himawari-8/9-Japan's new-generation geostationary meteorological satellites. *J. Meteor. Soc. Japan*, **94**, 151−183, https://doi.org/10.2151/jmsj.2016-009.

Breiman, L., 2001: Random forests. *Machine Learning*, **45**, 5−32, https://doi.org/10.1023/A:1010933404324.

Chen, N., W. Li, C. Gatebe, T. Tanikawa, M. Hori, R. Shimada, T. Aoki, and K. Stamnes, 2018: New neural network cloud mask algorithm based on radiative transfer simulations. *Remote Sens. Environ.*, **219**, 62−71, https://doi.org/10.1016/j.rse.2018.09.029.

Dessler, A. E., 2010: A determination of the cloud feedback from climate variations over the past decade. *Science*, **330**, 1523−1527, https://doi.org/10.1126/science.1192546.

Dybbroe, A., K.-G. Karlsson, and A. Thoss, 2005: NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part I: Algorithm description. *J. Appl. Meteorol.*, **44**, 39−54, https://doi.org/10.1175/JAM-2188.1.

Frey, R. A., S. A. Ackerman, Y. H. Liu, K. I. Strabala, H. Zhang, J. R. Key, and X. Wang, 2008: Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for Collection 5. *J. Atmos. Oceanic Technol.*, **25**, 1057−1072, https://doi.org/10.1175/2008JTECHA1052.1.

Geoffroy, O., J.-L. Brenguier, and I. Sandu, 2008: Relationship between drizzle rate, liquid water path and droplet concentration at the scale of a stratocumulus cloud system. *Atmospheric Chemistry and Physics*, **8**, 4641−4654, https://doi.org/10.5194/acp-8-4641-2008.

Gomis-Cebolla, J., J. C. Jimenez, and J. A. Sobrino, 2020: MODIS probabilistic cloud masking over the Amazonian evergreen tropical forests: A comparison of machine learning-based methods. *Int. J. Remote Sens.*, **41**, 185−210, https://doi.org/10.1080/01431161.2019.1637963.

Heidinger, A. K., A. T. Evan, M. J. Foster, and A. Walther, 2012: A naive bayesian cloud-detection scheme derived from *CALIPSO* and applied within PATMOS-X. *J. Appl. Meteorol. Climatol.*, **51**, 1129−1144, https://doi.org/10.1175/JAMC-D-11-02.1.

Hughes, M. J., and D. J. Hayes, 2014: Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sensing*, **6**, 4907−4926, https://doi.org/10.3390/rs6064907.

Imai, T., and R. Yoshida, 2016: Algorithm theoretical basis for Himawari-8 cloud mask product. Meteorological Satellite Center Tech. Note, 61, 17 pp.

Ishida, H., Y. Oishi, K. Morita, K. Moriwaki, and T. Y. Nakajima, 2018: Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions. *Remote Sens. Environ.*, **205**, 390−407, https://doi.org/10.1016/j.rse.2017.11.003.

Karlsson, K.-G., E. Johansson, and A. Devasthale, 2015: Advancing the uncertainty characterisation of cloud masking in passive satellite imagery: Probabilistic formulations for NOAA AVHRR data. *Remote Sens. Environ.*, **158**, 126−139, https://doi.org/10.1016/J.RSE.2014.10.028.

Key, J., 1990: Cloud cover analysis with Arctic Advanced Very High Resolution Radiometer data: 2. classification with spectral and textural measures. *J. Geophys. Res.*, **95**, 7661−7675, https://doi.org/10.1029/JD095iD06p07661.

Lai, R. Z., S. W. Teng, B. Q. Yi, H. Letu, S. H. Tang, and C. Liu, 2019: Comparison of cloud properties from Himawari-8 and FengYun-4A geostationary satellite radiometers with MODIS cloud retrievals. *Remote Sensing*, **11**, 1703, https://doi.org/10.3390/rs11141703.

Le Goff, M., J.-Y. Tourneret, H. Wendt, M. Ortner, and M. Spigai, 2017: Deep learning for cloud detection. *Proc. 8th International Conf. of Pattern Recognition Systems* (*ICPRS 2017*), Madrid, IET, 1−6, https://doi.org/10.1049/cp.2017.0139.

Letu, H., and Coauthors, 2020: High-resolution retrieval of cloud microphysical properties and surface solar radiation using Himawari-8/AHI next-generation geostationary satellite. *Remote Sens. Environ.*, **239**, 111583, https://doi.org/10.1016/J.RSE.2019.111583.

Loveland, T. R., and A. S. Belward, 1997: The international geosphere biosphere programme data and information system Global Land Cover Data Set (DIScover). *Acta Astronautica*, **41**, 681−689, https://doi.org/10.1016/S0094-5765(98)00050-2.

Lyapustin, A., Y. Wang, and R. Frey, 2008: An automatic cloud mask algorithm based on time series of MODIS measurements. *J. Geophys. Res.*, **113**, D16207, https://doi.org/10.1029/2007JD009641.

Min, M., and Coauthors, 2017: Developing the science product algorithm testbed for Chinese next-generation geostationary meteorological satellites: Fengyun-4 series. *Journal of Meteorological Research*, **31**, 708−719, https://doi.org/10.1007/S13351-017-6161-Z.

Platnick, S., M. D. King, S. A. Ackerman, W. P. Menzel, B. A. Baum, J. C. Riedi, and R. A. Frey, 2003: The MODIS cloud products: Algorithms and examples from Terra. *IEEE Trans.*

Geosci. Remote Sens., **41**, 459–473, https://doi.org/10.1109/TGRS.2002.808301.

Poulsen, C., U. Egede, D. Robbins, B. Sandeford, K. Tazi, and T. Zhu, 2020: Evaluation and comparison of a machine learning cloud identification algorithm for the SLSTR in polar regions. *Remote Sens. Environ.*, **248**, 111999, https://doi.org/10.1016/j.rse.2020.111999.

Rossow, W. B., and L. C. Garder, 1993: Cloud detection using satellite measurements of infrared and visible radiances for ISCCP. *J. Climate*, **12**, 2341–2369, https://doi.org/10.1175/1520-0442(1993)006<2341:CDUSMO>2.0.CO;2.

Sakaida, F., K. Hosoda, M. Moriyama, H. Murakami, A. Mukaida, and H. Kawamura, 2006: Sea surface temperature observation by Global Imager (GLI)/ADEOS-II: Algorithm and accuracy of the product. *Journal of Oceanography*, **62**, 311–319, https://doi.org/10.1007/S10872-006-0056-4.

Saunders, R. W., and K. T. Kriebel, 1988: An improved method for detecting clear sky and cloudy radiances from AVHRR data. *Int. J. Remote Sens.*, **9**, 123–150, https://doi.org/10.1080/01431168808954841.

Scornet, E., 2018: Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, **60**, 144–162, https://doi.org/10.1051/proc/201760144.

Shang, H. Z., L. F. Chen, H. Letu, M. Zhao, S. S. Li, and S. H. Bao, 2017: Development of a daytime cloud and haze detection algorithm for Himawari ‐ 8 satellite measurements over central and eastern China. *J. Geophys. Res.*, **122**, 3528–3543, https://doi.org/10.1002/2016JD025659.

Stephens, G. L., and Coauthors, 2002: The CloudSat Mission and the A-Train: A new dimension of space-based observations of clouds and precipitation. *Bull. Amer. Meteor. Soc.*, **83**, 1771–1790, https://doi.org/10.1175/BAMS-83-12-1771.

Stowe, L. L., P. A. Davis, and E. P. McClain, 1999: Scientific basis and initial evaluation of the CLAVR-1 global clear/cloud classification algorithm for the Advanced Very High Resolution Radiometer. *J. Atmos. Oceanic Technol.*, **16**, 656–681, https://doi.org/10.1175/1520-0426(1999)016<0656:SBAIEO>2.0.CO;2.

Strabala, K. I., S. A. Ackerman, and W. P. Menzel, 1994: Cloud properties inferred from 8–12 μm data. *J. Appl. Meteorol. Climatol.*, **33**, 212–229, https://doi.org/10.1175/1520-0450(1994)033<0212:CPIFD>2.0.CO;2.

Sulla-Menashe, D., and M. A. Friedl, 2018: *User Guide to Collection 6 MODIS Land Cover* (*MCD12Q1 and MCD12C1*) *Product*. USGS.

Swami, A., and R. Jain, 2013: Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, **12**, 2825–2830.

Thampi, B. V., T. Wong, C. Lukashin, and N. G. Loeb, 2017: Determination of CERES TOA fluxes using machine learning algorithms. Part I: Classification and retrieval of CERES cloudy and clear scenes. *J. Atmos. Oceanic Technol.*, **34**, 2329–2345, https://doi.org/10.1175/JTECH-D-16-0183.1.

Visa, A., K. Valkealahti, and O. Simula, 1991: Cloud detection based on texture segmentation by neural network methods. *Proc. IEEE International Joint Conference on Neural Networks*. Singapore, IEEE, 1001–1006, https://doi.org/10.1109/IJCNN.1991.170529.

Wang, C. X., S. Platnick, S. K. Meyer, Z. B. Zhang, and Y. P. Zhou, 2020: A machine-learning-based cloud detection and thermodynamic-phase classification algorithm using passive spectral observations. *Atmospheric Measurement Techniques*, **13**, 2257–2277, https://doi.org/10.5194/amt-2019-409.

Wang, J. J., C. Liu, M. Min, X. Q. Hu, Q. F. Lu, and H. Letu, 2018: Effects and applications of satellite radiometer 2.25-μm channel on cloud property retrievals. *IEEE Trans. Geosci. Remote Sens.*, **56**, 5207–5216, https://doi.org/10.1109/TGRS.2018.2812082.

Wang, X., M. Min, F. Wang, J. P. Guo, B. Li, and S. H. Tang, 2019: Intercomparisons of cloud mask products among Fengyun-4A, Himawari-8, and MODIS. *IEEE Trans. Geosci. Remote Sens.*, **57**, 8827–8839, https://doi.org/10.1109/TGRS.2019.2923247.

Winker, D. M., W. H. Hunt, and M. J. McGill, 2007: Initial performance assessment of CALIOP. *Geophys. Res. Lett.*, **34**, https://doi.org/10.1029/2007GL030135.

Wylie, D. P., W. P. Menzel, H. M. Woolf, and K. T. Strabala, 1994: Four years of global cirrus cloud statistics using HIRS. *J. Climate*, **7**, 1972–1986, https://doi.org/10.1175/1520-0442(1994)007%3C1972:FYOGCC%3E2.0.CO;2.

Zhang, C. W., X. Y. Zhuge, and F. Yu, 2019: Development of a high spatiotemporal resolution cloud-type classification approach using Himawari-8 and CloudSat. *Int. J. Remote Sens.*, **40**, 6464–6481, https://doi.org/10.1080/01431161.2019.1594438.