Low Cloud Detection in Multilayer Scenes Using Satellite Imagery with Machine Learning Methods

JOHN M. HAYNES,[a] YOO-JEONG NOH,[a] STEVEN D. MILLER,[a] KATHERINE D. HAYNES,[a] IMME EBERT-UPHOFF,[a] AND ANDREW HEIDINGER[b]

[a] *Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado*
[b] *NOAA/NESDIS, Madison, Wisconsin*

(Manuscript received 22 June 2021, in final form 17 October 2021)

ABSTRACT: The detection of multilayer clouds in the atmosphere can be particularly challenging from passive visible and infrared imaging radiometers since cloud boundary information is limited primarily to the topmost cloud layer. Yet detection of low clouds in the atmosphere is important for a number of applications, including aviation nowcasting and general weather forecasting. In this work, we develop pixel-based machine learning–based methods of detecting low clouds, with a focus on improving detection in multilayer cloud situations and specific attention given to improving the Cloud Cover Layers (CCL) product, which assigns cloudiness in a scene into vertical bins. The random forest (RF) and neural network (NN) implementations use inputs from a variety of sources, including GOES Advanced Baseline Imager (ABI) visible radiances, infrared brightness temperatures, auxiliary information about the underlying surface, and relative humidity (which holds some utility as a cloud proxy). Training and independent validation enlists near-global, actively sensed cloud boundaries from the radar and lidar systems on board the *CloudSat* and *CALIPSO* satellites. We find that the RF and NN models have similar performances. The probability of detection (PoD) of low cloud increases from 0.685 to 0.815 when using the RF technique instead of the CCL methodology, while the false alarm ratio decreases. The improved PoD of low cloud is particularly notable for scenes that appear to be cirrus from an ABI perspective, increasing from 0.183 to 0.686. Various extensions of the model are discussed, including a nighttime-only algorithm and expansion to other satellite sensors.

SIGNIFICANCE STATEMENT: Using satellites to detect the heights of clouds in the atmosphere is important for a variety of weather applications, including aviation weather forecasting. However, detecting low clouds can be challenging if there are other clouds above them. To address this, we have developed machine learning–based models that can be used with passive satellite instruments. These models use satellite observations at visible and infrared wavelengths, an estimate of relative humidity in the atmosphere, and geographic and surface-type information to predict whether low clouds are present. Our results show that these models have significant skill at predicting low clouds, even in the presence of higher cloud layers.

KEYWORDS: Clouds; Cloud retrieval; Lidars/Lidar observations; Radars/Radar observations; Remote sensing; Satellite observations; Nowcasting; Machine learning

1. Introduction

Passive remote sensing instruments, like the Advanced Baseline Imager (ABI; Schmit et al. 2017) on the Geostationary Operational Environmental Satellite (GOES), are particularly effective at revealing attributes of the topmost layer of clouds, such as cloud-top height, particle size, and water phase. However, these instruments lack the ability to reliably observe below the top cloud layer in multilayer scenes, causing information on lower layers to be limited or nonexistent. Accurately identifying low cloud presence has important aviation and meteorological implications. Low clouds can be hazardous for aviation, as they reduce visibility and can contain supercooled water that freezes on aircraft control surfaces. Additionally, accurate low cloud identification would provide more accurate sky conditions to improve weather forecasts, to aid boundary layer cloud studies, and to more accurately assess the role of these clouds in Earth's radiation budget.

While there is a rich body of literature on cloud detection using satellite data, a focus on low cloud detection from satellites is significantly more limited, much of which is focused on polar-orbiting satellites (e.g., Platnick et al. 2017; Wang et al. 2016; Wind et al. 2010). Most recent work on cloud masking has focused on identifying cloud types at any vertical level (e.g., Qin et al. 2019; Shang et al. 2018; Stöckli et al. 2019); however, there are two notable exceptions. First, Leinonen et al. (2019) used a conditional generative adversarial network (CGAN) to assess cloud vertical structure from Moderate Resolution Imaging Spectroradiometer (MODIS), producing a wide range of realistic-looking vertical cloud structures, including low clouds. Second, Andersen and Cermak (2018) used geostationary satellite data to predict fog and low cloud in the Namib. Using infrared channels only from the Spinning Enhanced Visible and Infrared Imager on board the *Meteosat-11* satellite, they classified scenes using a decision tree algorithm with sequential applications of spectral thresholds.

Corresponding author: John Haynes, john.haynes@colostate.edu

DOI: 10.1175/JTECH-D-21-0084.1

Evaluating this algorithm against net radiation measurements from a station network revealed an overall correctness of classification of 97% for the region.

The continuous high-resolution spatial images from the ABI, updated as frequently as every 30 s over the mesoscale sectors and every 10–15 min for the full disk, opens doors to developing new methods of detecting low clouds. The Cloud Cover Layers (CCL) product provides a vertical depiction of the presence of cloud layers and fractional cloud cover for each layer. CCL is an official cloud product for the GOES-R ABI[1] and Visible Infrared Imaging Radiometer Suite (VIIRS;[2] Hillger et al. 2013) on board the Joint Polar Satellite System (JPSS) satellites. The intent of CCL is to classify clouds in a given scene by their vertical extent, using a variety of prescribed thresholds based on pressure levels, or at flight levels for aviation community users. It is the only official ABI/VIIRS product that focuses on quantifying the vertical distribution of cloud cover.

As originally formulated, CCL was primarily a function of the retrieved cloud-top height (CTH), and included limited information on cloud vertical extent. Adding statistically based retrievals of cloud geometric thickness (CGT) has since allowed clouds to span multiple vertical levels between CTH and an inferred base, improving the representation of lower cloud layers (Noh et al. 2017). Leveraging CGT has allowed CCL to include three additional layer designations: L + M, H + M, and H + M + L, where L = low cloud, M = middle cloud, and H = high cloud. *All references to CCL that follow refer to this updated CCL algorithm that accounts for CGT.* This CCL algorithm is applicable to both ABI and VIIRS, and has been implemented in the Clouds from AVHRR Extended (CLAVR-x) system, which forms NOAA's operational cloud algorithm processing framework (Heidinger and Straka 2013). In this paper, we use a recent version of CCL that assigns pressure-based layer cutoffs of 631 and 350 hPa to differentiate between the L/M and M/H layers.

Figure 1 shows the occurrence of clouds in these various CCL categories from the combined perspective of two spaceborne active sensors, the *CloudSat* radar and the *CALIPSO* lidar (details of the specific datasets used and how they are combined to derive cloud boundaries are provided in section 2). The combined radar and lidar designations of cloud height are available in 240 m bins, suffering far less ambiguity than ABI in resolving multiple cloud layers. These cloud profiles are compared against *GOES-16* ABI-based CCL designations of cloud height that have been matched in space and time (also per section 2). It is clear from Fig. 1 that CCL is biased against low clouds when high clouds are present, and it is noteworthy that the CCL algorithm cannot produce the high-over-low (H+L) designation, since it assumes a single, vertically contiguous cloud layer. As a result, many scenes that are H+L or H+M+L are instead placed into the coalesced H+M category, which CCL overpopulates relative to the radar/lidar observations. While geostationary and polar-orbiting satellites with passive

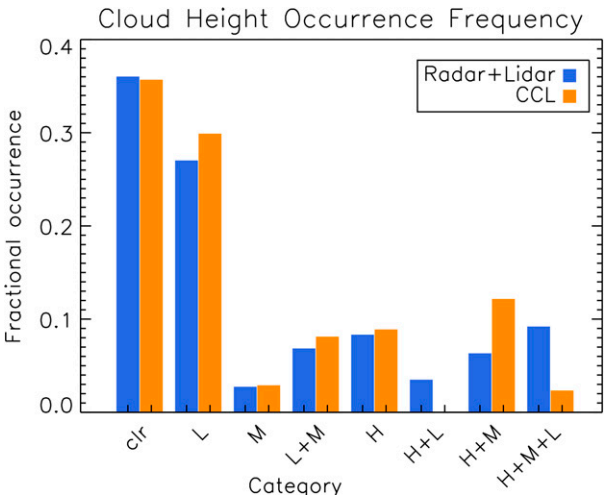

FIG. 1. Occurrence of clouds in the various CCL categories (horizontal axis), as observed by a combination of *CloudSat* radar and *CALIPSO* lidar (blue bars) that was matched in space and time to the ABI CCL designations (orange bars). This analysis was performed over the multimonth testing period described in section 2.

sensors (like ABI and VIIRS) are the gold standard for near-real-time cloud detection, it is clear from Fig. 1 that their ability to characterize the three-dimensional cloud scene is limited. Forecasters, therefore, will often rely on numerical weather prediction (NWP) or cloud proxies [like layer relative humidity (RH)] when determining whether low or midlevel cloud may be present in a scene that is obscured by higher cloud layers.

In this study, we hypothesize that supervised machine learning (ML) can be applied to this problem and used to infer the presence of low clouds in multilayered scenes directly from passive instruments. Training data for the models we develop consist of passive sensor data (*GOES-16* ABI) matched to active radar and lidar data (*CloudSat* and *CALIPSO*), supplemented by column relative humidity data. Specifically, we demonstrate the following:

- Relative humidity holds utility as a predictor for low cloud when used in an ML algorithm, and as such NWP input produces a substantial enough contribution to warrant inclusion in an otherwise observation-driven product.
- Random forests and neural networks can be successfully applied to this detection problem on a pixel-by-pixel basis.
- Both of these ML models outperform the current ABI CCL product in predicting the presence of low clouds, especially when multiple layers are present.
- These methods can be extended to work in nighttime mode and should be extendable to other sensors besides ABI.

This paper is organized as follows. Section 2 discusses the data used in the study. Section 3 describes the methodology, including the required preprocessing, and describes the ML approach as well as the motivation for using RH as an input feature. Section 4 shows the results, provides examples of the improved algorithm, and illustrates the implications of our

[1] https://www.goes-r.gov/products/opt2-cloud-layers-height.html.
[2] https://www.star.nesdis.noaa.gov/jpss/clouds.php.

proposed approach. Section 5 concludes the paper and outlines future work.

2. Data and analysis metrics

a. Data used in study

The data used in this study consist of space-and-time matched observations from 1) the ABI on *GOES-16*, 2) the radar and lidar on board *CloudSat* and *CALIPSO*, respectively, and 3) auxiliary data from an NWP model, as described below. Time periods of analysis and the process by which they are matched together are described in the following subsection.

1) *GOES-16*

The *GOES-16* ABI is positioned over the equator at approximately 75°W and provides continuous coverage over the Americas and the Atlantic. The ABI detects both reflected solar radiation and emitted infrared radiation in 16 bands whose centers vary from 0.47 to 13.3 μm. The ABI performs full disk scans at a time interval of 10–15 min, depending on operating mode, and the data are collected at spatial resolutions varying from 0.5 to 2 km as measured at the equator (Schmit et al. 2017). Here, we use level 2 data where all channels are averaged to the same 2 km grid. Additionally, we make use of the ground latitude and solar and sensor zenith angles.

For comparisons against the current operational approach, we obtain the existing CCL cloud classifications from local runs of CLAVR-x that are already matched in space and time to the 2 km ABI grid. Additionally, we make use of the existing CLAVR-x land–ocean mask.

2) *CLOUDSAT* AND *CALIPSO*

To evaluate low cloud presence, we use the active sensors on board two polar-orbiting satellites, *CloudSat* and *CALIPSO*. These sensors flew in formation [~1330 local time at ascending node (LTAN)] during the period of analysis for this study allowing joint observations of vertical cloud structure in overlapping regions of the atmosphere (Stephens et al. 2018; Mace et al. 2009). *CloudSat* carries the 94 GHz Cloud Profiling Radar (CPR) designed for the vertical profiling of clouds and hydrometeors (Stephens et al. 2008), and *CALIPSO* carries the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) instrument, which is a dual-wavelength polarization lidar providing high resolution vertical profiles of clouds and aerosols (Winker et al. 2007).

This study uses the 2B-GEOPROF (Marchand et al. 2008) and 2B-GEOPROF-lidar (Mace and Zhang 2014) products to provide a best estimate of actual low cloud presence in the 240 m resolution vertical bins provided in the products. To create a vertical cloud mask for a given radar profile, we start with the cloud mask from 2B-GEOPROF, considering cloud to be present in any radar bin where the mask has a value of 20 or greater (which represents clouds with a relatively low chance of being a false detection). This is supplemented with volume cloud fractions from 2B-GEOPROF-lidar, which are provided in bins matched to the radar footprint. Lidar-derived cloud is considered to be present in radar range bins containing 50% or greater lidar cloud coverage. In this way, cloud is flagged as occurring in those range bins where either the radar or lidar identifies a cloud return.

The final step is to convert this height-resolved cloud mask to one defined by the three broad pressure-based levels output by CCL (i.e., separated by the 631 and 350 hPa pressure levels); the *CloudSat* ECMWF-AUX product (Cronk and Partain 2017) provides for this translation. If there is cloud detected in any of the height bins with pressures exceeding 631 hPa, then the profile is considered to have low cloud. Further, midlevel clouds are clouds in the layers with pressures between 631 and 350 hPa, and high clouds have pressures below 350 hPa.

The advantage of this approach is that CPR and CALIOP together provide high-quality, complementary observations of the vertical structure of near-global cloudiness. For example, the lidar observes thin cirrus missed by the radar, whereas the radar penetrates deep into optically thick cloud layers, while the lidar attenuates in these situations. Even so, there are some notable limitations to this combination. First, extremely low-topped clouds (under 1 km) will be missed if heavily attenuating upper-level cloud is also present. This is because the CPR cannot reliably detect clouds in the bins comprising approximately the lowest 1 km above Earth's surface (e.g., Tanelli et al. 2008), and CALIOP will also miss these clouds if there is significant attenuation above the low-topped cloud. Second, the analysis is restricted to daytime-only since *CloudSat* suffered a battery anomaly in 2011 (Nayak 2012) that precluded operation of the radar at night. Finally, the analysis is subject to any biases introduced by the orbital characteristic of the A-Train satellites, namely, their ~1330 LTAN.

3) RELATIVE HUMIDITY AND SURFACE CHARACTERISTICS

This study uses RH, snow equivalent water depth, and ice fraction obtained from the operational 0.5°, 6 hourly Global Forecasting System (GFS) 12 h forecast data.[3] Although we recognize that more up-to-date surface-state information can be obtained directly from retrievals using sensors designed for this purpose, this study aims to produce a product suitable for operational use and instead utilizes information that is readily available for CLAVR-x ingest. To match the GFS variables in space and time to the ABI full disk grid, we temporally interpolate between two bounding 12 h forecasts (following the procedure used by CLAVR-x), and then apply bilinear spatial interpolation.

RH data from GFS are collected at 150, 650, 750, 850, 950, and 1000 hPa, and is adjusted to always be relative to liquid water. Using these data, we calculate the maximum RH between 650 and 1000 (RH_{max}). For the surface information, we combine the snow equivalent water depth and ice fraction to create a flag ($Flag_{snow/ice}$) that indicates whether snow or ice is present.

[3] https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs.

b. Data matching and preprocessing

The time periods selected for analysis in this study are driven by data availability, most notably that of *CloudSat*/*CALIPSO*, since these data are not processed in real time. The data are separated into "training" and "testing" subsets using continuous month-long periods separated by breaks of at least one month, in order to obtain a spread of seasonal coverage and avoid bias between training and testing. In developing our detection algorithm, the training dataset is used for model training and hyperparameter tuning, while the testing dataset is used only for evaluation of final model performance. Before applying the filtering described in the next paragraph, the full training period consists of 12.2 million radar profiles from October and December of 2018; as well as February, April, and June of 2019; and the full testing period consists of 9.4 million radar profiles from November of 2018 and March and May of 2019.

To match the *CloudSat*/*CALIPSO* data to the 2 km ABI full-disk grid, we loop through the radar profiles in a given time period and find the closest matching set of ABI (and CCL) observations. Since *CloudSat*/*CALIPSO* view near nadir while the ABI viewing angle increases with distance from the subsatellite point, a parallax correction is applied based on the top of the highest cloud observed in each profile by the radar/lidar. To be considered, the postcorrection match is required to occur within 7.5 km and 10 min of the time of the CPR profile. Since the radar footprint is approximately 1.4 km while the ABI box is at least 2 km on a side (and much larger at higher latitudes), several consecutive radar profiles may match to the same ABI information. This "mismatch" is purposefully retained to better represent the variability in ABI data for a given vertical cloud scenario. To prevent contamination by large parallax uncertainties at higher sensor zenith angles, we restrict data analysis to sensor zenith angles of less than approximately 70°, corresponding to the maximum angle at which CCL retrievals were performed. Furthermore, we consider only daylight pixels with a solar zenith angle not exceeding 82°; both of these thresholds were chosen for consistency with the version of CLAVR-x that we compare against. After applying these restrictions and any removing bad CCL retrievals from the testing data, the training and testing datasets have 6.8 and 4.8 million profiles, respectively.

c. Analysis metrics

A number of metrics of algorithm performance are used, defined in terms of standard contingency table nomenclature using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN); all vary between 0 and 1 unless otherwise noted:

- Probability of detection (PoD) is given by TP/(TP + FN); 1 is the best score.
- False alarm ratio (FAR) is given by FP/(TP + FP); 0 is the best score.
- Critical success index (CSI), or threat score, is a combination of PoD and FAR, and is given by TP/(TP + FN + FP); 1 is the best score.
- F1 score, often used in computer science, is closely related to the CSI and is given by TP/[TP + 0.5 × (FN + FP)]; 1 is the best score.
- Accuracy is defined as (TP + TN)/(TP + TN + FP + FN); 1 is the best score.
- Frequency bias is defined as (TP + FP)/(TP + FN); 1 is the best score, and it can take any value greater than or equal to zero.

3. Methodology

This section describes the models used for prediction of low cloud and how they are implemented using the training and validation datasets described above.

a. Relative humidity as a low cloud proxy

Before describing the ML models, we motivate the inclusion of information beyond what is observable by the ABI, specifically relative humidity. Cloud formation and maintenance requires a number of supporting conditions, including sufficient moisture, a source of uplift or diabatic cooling to produce supersaturation, and (in most cases) aerosol particles to serve as cloud condensation nuclei (Seinfeld and Pandis 2006). ABI observations alone cannot sense any of these parameters directly in cloud scenes, but RH is a parameter that addresses saturation in the column. We hypothesize that high levels of saturation, especially at low levels, may serve as a useful piece of auxiliary information in identifying the presence of low clouds, especially in situations where these clouds are completely obscured by higher clouds. Therefore, we explore the utility of RH as a proxy for low cloud.

While there is no guarantee that clouds will form if the RH is relatively high (e.g., 2 m RH is nearly always high over the tropical oceans), clouds will not form without local supersaturation. A few past studies have examined this relationship in terms of observed clouds (Walcek 1994; Tompkins 2003) or simulated clouds (Groisman et al. 2000), but none to our knowledge have extensively evaluated this relationship using the highly vertically resolved cloud mask that *CloudSat* and *CALIPSO* have provided on a near-global scale. Walcek (1994) find that RH is the strongest predictor of cloud cover among the various quantities they evaluated; they find a correlation coefficient of 0.6 globally, but the strength of this relationship varies by region and with height.

To provide additional moisture information to passive sensors to help predict low clouds, we investigate using low-level relative humidity as a proxy for cloud cover. The purpose of this is twofold: 1) to determine if RH is actually a useful predictor of cloud cover, and therefore should be included in our low cloud prediction scheme, and 2) to establish if RH thresholding is useful for predicting whether cloud is present or not. Using the full testing dataset, we examine the relationship between GFS RH thresholds (50%, 60%, 70%, 80%, and 90% with respect to liquid water) at various levels in the atmosphere and the presence of low cloud from the radar/lidar as defined earlier. In this experiment, we assign low

cloud any time the RH exceeds the given threshold on the given level(s), and we assess the predictive power of RH.

Predictive scores for RH thresholds at two levels are shown in Fig. 2. First, we choose 850 hPa (RH_{850}), as it lies between the surface and upper pressure bound of 631 hPa used as the cutoff between the low and midcloud levels in this study. Second, we choose the previously defined RH_{max}, because it approximately corresponds to the maximum RH in the low cloud layer. In both cases, the trade-off between using low and high thresholds is immediately apparent: using a low threshold of RH results in a relatively high PoD, but also a relatively high FAR. The plots are broken down between land and water surfaces, but note that since some latitude bands are dominated by water, the land results may have low statistical significance for those locations.

In terms of PoD and RH_{max}, it is not surprising that the specific threshold chosen has less effect over land than water; for example, PoD over water surfaces is nearly uniform for RH $\leq$ 80%. This is because, all else being equal, the lower atmosphere will be moistened more readily over water surfaces than land. It is also noteworthy that RH_{max} has more detection utility than RH_{850}, which is probably a reflection of the fact that validation is not being performed at a specific level, but over a broad layer that is similar to that given by the limits of RH_{max}. For FAR, specific values of RH make less difference to false alarms than does latitude: the tropics have more false alarms than higher latitudes. This is, to some degree, a function of midlatitude versus tropical dynamics; in the tropics, cloud-free moisture plumes occur more readily in subsidence regions, whereas midlatitude cloudiness is more driven by frontal dynamics.

Since it is a function of both PoD and FAR, CSI provides a useful summary of the utility of RH as a predictor of low-level cloudiness. It is apparent from Fig. 2 that CSI is largest at mid- and high latitudes and lowest in the tropics, for the reasons discussed above. Furthermore, the middle ranges of RH (i.e., between 60% and 80%) provide the highest CSIs on average. We note that the same analyses performed at 650 and 1000 hPa, (not shown) do not change these general conclusions. Given these results, it is apparent that RH has some predictive utility for low cloud, consistent with the findings of Walcek (1994), but it is not at all clear that there is a single threshold that should be used; the threshold would vary with geographic location and underlying surface type, and most likely with season as well. Therefore, our strategy is to provide RH and allow the ML algorithm to learn the best way to use RH information in terms of a larger low cloud detection scheme.

b. Machine learning approach

To identify low clouds in a scene on a pixel-by-pixel basis, we explore two machine learning approaches in this study. The first is the random forest (RF), a subset of decision tree methods, and the second is the artificial neural network (NN). Both models are trained on the training periods, and all results shown are obtained by running the trained models on the testing periods, with *CloudSat*/*CALIPSO* cloud boundaries as the truth.

Before describing the models in detail, we will describe the training data that are used as inputs to both models. Table 1 describes these training features (n_{feat} = 21), which includes data from all 16 ABI channels. Visible reflectances from channels 1 through 6 are normalized by the cosine of the solar zenith angle (Lee et al. 2021). In addition to the ABI data, we also provide three categories of auxiliary data. First, we provide saturation information via RH: RH_{max} provides information on low-level moisture content, and the RH at 150 hPa (RH_{150}) is used to provide information that may be relevant for high-cloud layers. Second, we provide surface information in terms of both a land mask flag ($Flag_{sfc}$) and a snow/ice flag ($Flag_{snow/ice}$). Finally, we provide a limited viewing angle dependency through latitude (lat), though we note that using sensor zenith angle instead results in virtually no change in algorithm performance. In section 4, the relative importance of these variables (and which can be potentially dropped with minimal performance decrease) will be discussed.

We now provide details on the ML models themselves. The RF classifier (Breiman 2001) is an ensemble methodology that uses a "forest" of decision trees to make a binary decision (low cloud present, or low cloud not present) based on training data. Decision trees can be visualized as a series of questions designed to split the data into increasingly distinct classes; RFs, in turn, are composed of multiple decision trees, where correlation between trees is reduced by restricting each tree to a randomly selected subset of the training data. Decision trees are gaining popularity for cloud type classification problems (e.g., Yu et al. 2021; Sedlar et al. 2021; Wang et al. 2020), and as one of the easier ML frameworks to interpret, they are a popular choice for physically explainable artificial intelligence.

We implement the RF model in Python using Scikit-learn (Pedregosa et al. 2011). Each leaf in each decision tree has a probability of low cloud associated with it that is determined from the distribution of samples in the training data; for a given input sample, low cloud is determined to be present if the average low cloud probability over all trees in the forest (P) is at least 0.5. Tunable hyperparameters include the maximum depth of trees, the minimum data points required for a split, the minimum number of samples per leaf node, and whether or not bootstrapping (sampling with replacement) is used. All experiments use 125 trees, which provides a balance between accuracy and training time, but we tuned the other parameters using a grid search with fivefold cross validation (Russell et al. 2010). The search space, which was set to optimize the F1 score, and optimized values, are listed in Table 2.

In addition to the RF, we use a fully connected classification neural network (Burkov 2019; Géron 2019; Chollet 2018) to predict low cloud presence. We implement the NN model in Python using TensorFlow on a Google Colab GPU. For the NN, we scale all the input features to values between 0 and 1 using unity-based normalization. Since the model predictions are a binary prediction of the presence of low cloud, we use the binary cross-entropy loss with a final softmax output layer. Early stopping is not used during training. To find the optimal

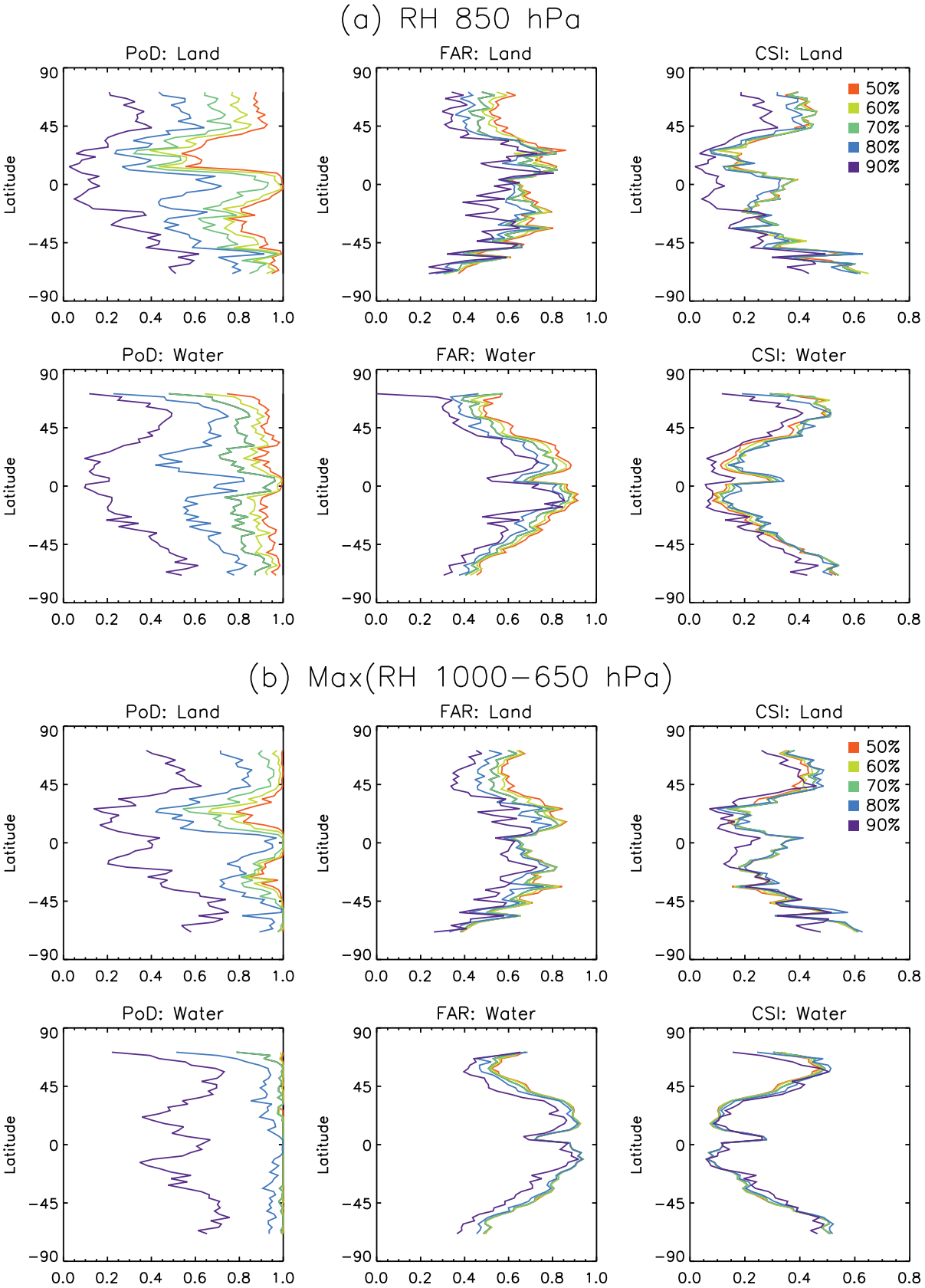

FIG. 2. (a) (left) PoD, (center) FAR, and (right) CSI for low cloud as predicted by 850 hPa RH exceeding various values (given by colors) for (top) land and (bottom) water surfaces (as labeled). Time period of analysis is that of the full testing dataset. (b) As in (a), but for the maximum RH between 1000 and 650 hPa, as defined in the text.

TABLE 1. Variables used in the baseline random forest and neural network experiment.

Variable	Description and units	Notes
REFL01 through REFL06	ABI channel 1–6 visible reflectances	Normalized by cos(solar zenith angle)
TB07–TB16	ABI channel 7–16 brightness temperature (K)	—
RH_{max}	Maximum RH between 650 and 1000 hPa (%)	—
RH_{150}	RH at 150 hPa (%)	—
Lat	Latitude (°N)	—
$Flag_{sfc}$	0 indicates land or mix, 1 = water surface	Set to 1 where CLAVR-x *land_class* is 0, 5, 6, 7
$Flag_{snow/ice}$	0 indicates snow/ice free land surface, 1 = snow/ice present	Set to 1 if either GFS snow depth or ice fraction are >0

model architecture and hyperparameter setup, we perform a guided search using Hyperopt (Bergstra et al. 2013). The search is conducted using the tree of Parzen estimators (TPE) algorithm, an optimization approach that sequentially constructs models to approximate the performance based on historical measurements (Bergstra et al. 2011). We perform 100 model searches on the training data using threefold cross validation and select the model with the highest validation balanced accuracy. The NN parameters, search space, and optimized values are shown in Table 2.

4. Results and discussion

a. Algorithm performance

Following training, low cloud occurrence is predicted for all points in the testing dataset. Figure 3 shows the PoD of low cloud as a function of cloud types derived from the CLAVR-x algorithm. For reference, the frequency of occurrence of each cloud type, and the frequency of occurrence of low cloud in each category according to the radar/lidar observations, are shown in Table 3. The "all" category includes all scenes without regard to cloud type (including points that CLAVR-x identifies as clear sky), and "all cloud" includes all such cases except clear-sky and CLAVR-x's "probably clear" category. Different color bars represent the different algorithms; note that the special case of the CCL + RF combination will be addressed toward the end of this section.

Results for the RF and NN are remarkably similar, with neither method showing a consistent advantage. Overall, the PoD of low cloud increases from 0.685 with the original CCL algorithm, to 0.815 (0.807) for RF (NN) (Fig. 3, "all"). The increases are most pronounced in three categories: overshooting tops, overlapping, and cirrus. Low cloud occurring under cirrus shows the most robust improvements, with PoD increasing from 0.183 for CCL to 0.686 (0.684) for RF (NN). There is little change in the supercooled water category, and decreased performance in the water cloud and fog categories, since these categories are dominated by single-layer low cloud, which CCL already captures quite well.

Corresponding values of FAR are shown in Fig. 4. Across all categories, the FAR decreases from the CCL-only value of 0.210 to 0.147 (0.137) for RF (NN); however, the categories that show the greatest improvement in PoD come with the

TABLE 2. Hyperparameters that are tuned and their optimized values.

Variable	Search space	Optimized value
	Random forest	
Maximum features to consider per split	$\sqrt{n_{feat}} - 2$, $\sqrt{n_{feat}}$, $\sqrt{n_{feat}} + 2$	$\sqrt{n_{feat}}$
Maximum tree depth	10, 20, 30, … , 110, 130	30
Minimum number of samples per split	2, 6, 10, 14	10
Minimum number of samples at each leaf node	2, 4, 6	6
Bootstrapping (selection with replacement?)	True, false	True
	Neural network	
Number of epochs	30, 45, 60, 75, 100, 120	100
Batch size	10 000, 50 000, 100 000, 150 000	10 000
Optimizer	Adam, Nadam	Adam
Learning rate	0.0001–0.1	0.001 24
First moment exponential decay rate (β_1)	0.8–1.0	0.892
Second moment exponential decay rate (β_2)	0.9–1.0	0.925
Number of fully connected layers	2–6	3
Activation type (per layer)	Hyperbolic tangent (tanh), rectified linear unit (relu)	[tanh, relu, relu]
Dropout rate (per layer)	0–0.4	[0., 0.081, 0.018]
Number of hidden units (per layer)	8–200	[37, 77, 71]

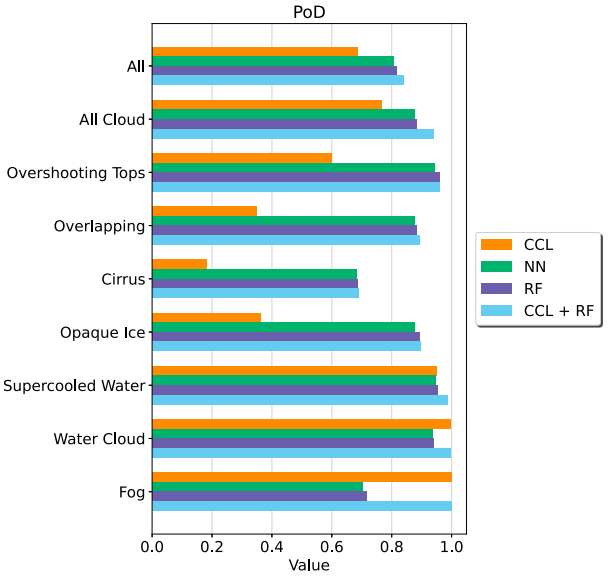

FIG. 3. Probability of detection of low cloud as a function of CLAVR-x cloud type (vertical axis) for CCL, RF, NN, and combined CCL + RF. The "all" category includes all scenes without regard to cloud type (including points that CLAVR-x identifies as clear sky), and "all cloud" includes all such cases except clear-sky and CLAVR-x's "probably clear" category.

trade-off of higher FAR values. For example, FAR in the cirrus category increases from 0.114 for CCL to 0.219 (0.206) for RF (NN). We argue that this increase is not particularly surprising, given the way that CCL assigns cloud base and therefore layer designations. The only way that CCL can detect low cloud in a scene with upper-level cirrus is by extending that cloud layer downward until the base drops below 631 hPa. This would require exceptionally thick cirrus. By contrast, RF and NN do not require this. Given the ambiguities of detecting low cloud signal in a cirrus scene, it is not surprising that increased detection comes with an increase in false alarms. By contrast, for the easier-to-retrieve supercooled water and water cloud categories, FAR is significantly lower for RF and NN than it is for CCL. CSI, which combines FAR and POD, is shown in Fig. 5. It is noteworthy that the CSI score improves across the board relative to CCL (except for fog, where there is a nearly identical result for RF and a slight drop for NN).

The geographic distribution of these metrics for RF, NN, and CCL is shown in Fig. 6. It is first noteworthy that the results for RF and NN are not only similar in bulk, but in their geographic distribution as well. There are no notable scenarios where one outperforms the other. We will therefore address these together as the ML method. Overall, the ML method of low cloud detection increases PoD and decreases FAR, resulting in improved CSI score in most geographic regions. The increased detection ability is most notable in the middle to high latitudes of both hemispheres, and is reflected by significant CSI increases in these regions. In the eastern portions of the South Pacific where stratus are common, it is no surprise that CCL performs well, and ML provides less improvement here. However, the adjacent South American continent has significant detection increases with the ML method.

b. Assessment of cross sections

To provide another perspective on algorithm performance, we examine cloud vertical cross sections for selected profiles in Fig. 7. Although there are millions of such profiles that could be shown, we focus on cirrus and opaque ice (two of the CLAVR-x cloud types from Table 3 that showed marked improvement in low cloud detection relative to CCL), and choose profiles 1000 through 2000 of the testing dataset for each cloud type. The reasoning for this is that a width of approximately 1000 profiles is easily viewable by eye (representing a horizontal distance of ~1100 km). For objectivity, we would have started with the first 1000 profiles of each cloud type, but for cirrus, these profiles contained relatively little low cloud, so the starting point was advanced by 1000 for both cloud types.

For each cloud type, the cross sections show equivalent radar reflectivity from the *CloudSat* CPR (displayed in rainbow colors), supplemented by locations where the CALIOP instrument on *CALIPSO* observed clouds but the CPR did not (purple colors). Each of the plots consist of sequential profiles in time from left to right and contain one CLAVR-x cloud type. Since cloud type changes from profile to profile (and since the CPR and CALIOP pass into or out of the *GOES-16* field of view regularly), this means that there are

TABLE 3. Cloud types from CLAVR-x, their frequency of occurrence, and the frequency of occurrence of low clouds observed by the radar/lidar in each category. "All" and "all cloud" are defined in the text.

ABI/CLAVR-x cloud type	Occurrence fraction	Low cloud occurrence from radar/lidar
All	1.000	0.418
All cloud	0.643	0.090
Overshooting tops	0.004	0.070
Overlapping	0.076	0.088
Cirrus	0.130	0.088
Opaque ice	0.056	0.082
Supercooled water	0.137	0.094
Water cloud	0.161	0.090
Fog	0.078	0.090

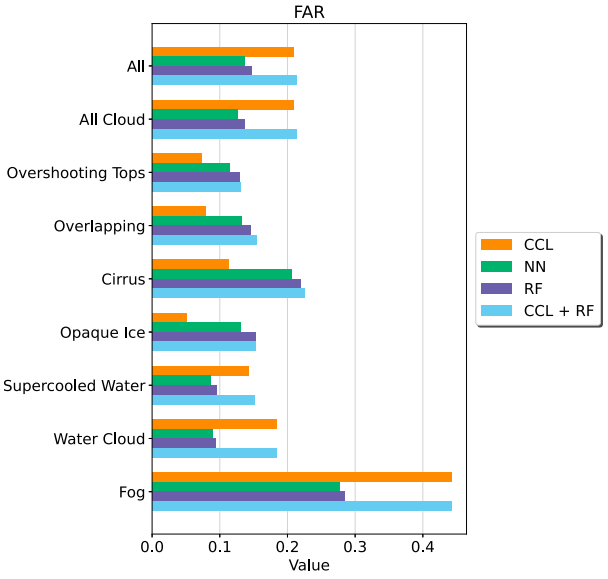

FIG. 4. As in Fig. 3, but for false alarm ratio.

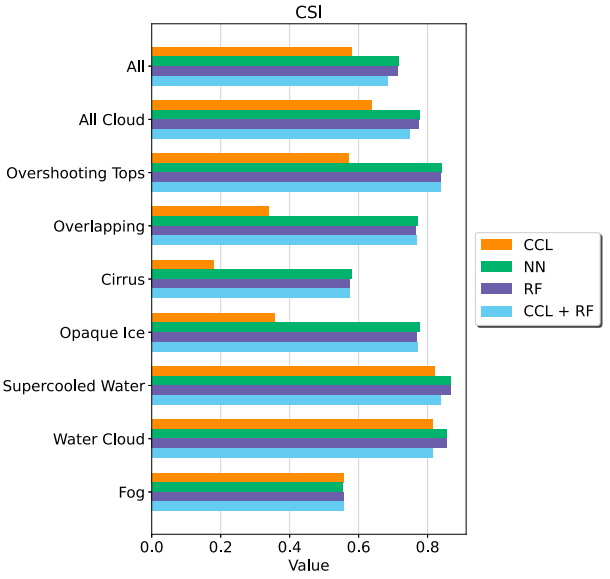

FIG. 5. As in Fig. 3, but for critical success index.

"breaks" in these plots (represented by gray dashed lines in the top row) where adjacently plotted profiles are actually nonadjacent in time.

The left panels of Fig. 7 are for scenes that CLAVR-x identifies as cirrus. Figure 7c adds a black background where CCL identified cloud in one of three pressure-based bins for H, M, and L layers. It is noteworthy that while the high (H) and middle (M) categories do in fact largely occur in regions where the radar/lidar identify cloud, CCL predicts very little low cloud occurrence, as evidenced by the lack of cloud identification in the lowest 5 km (no black shading) even when the radar/lidar show low clouds. This is especially true for the radar-identified low clouds near profile 1200, and the lidar identified low clouds (the thin purple lines, representing the top of these cloud layers, below which the lidar is completely attenuated) near profile 1350, and between profiles 1800 and 2000.

Figures 7e and 7g retain the same H and M predictions for CCL, but supplement CCL's low (L) layer cloud prediction with a black background when RH_{max} exceeds 60% and 80%, respectively. As discussed in section 3a, RH can be a useful predictor of low cloud occurrence, but the threshold required varies widely. In fact, RH_{max} exceeding 60% is obviously too low, as this would predict low cloud throughout the entire scene. Setting RH_{max} to 80% produces a much better result, albeit with significant false alarms in profiles less than about 1400.

A much better low cloud prediction is produced by our ML-based RF model, as depicted in Fig. 7i, where the CCL prediction is supplemented by positive RF predictions (hereafter referred to as CCL + RF; since NN performance is so similar, only RF is shown here). As discussed earlier, RF uses RH_{max} as one of its inputs, but does not require specification of a threshold. In this panel, it is clear that RF largely captures the profiles where low cloud is present, while producing relatively little cloud in regions where it is not present in the observations.

The right panels of Fig. 7 show the same progression of results, but this time for the CLAVR-x opaque ice category. Many of the segments featured in this cross section are both optically and geometrically deep, especially between profiles 1200 and 2000. The bright band is clearly visible in the equivalent radar reflectivity, with reflectivity decreasing toward the surface (a sign of strong attenuation for the millimeter wavelength CPR), revealing these clouds to contain significant stratiform precipitation. These deep segments are mixed with relatively thick cirrus, generally without underlying cloud. As in the cirrus case, Fig. 7d reveals that the original CCL misses most of the low cloud occurrence for these cases, due to blocking by the higher ice clouds. Using RH thresholds (Figs. 7f,h) has little utility in these cases, as both the 60% and 80% thresholds are met even in clear areas. By comparison, the RF solution, Fig. 7j, presents a better balance between detections and false alarms. It correctly detects the low cloud starting at profile 1000, the group of low clouds near profile 1100, and the heavily precipitating areas between profiles 1200 and 2000, while only missing the low clouds around profile 1200 and with minimal false alarms in the brief break of low cloud just prior to profile 1600 and again just prior to profile 1900.

c. Algorithm explainability: Feature importance

The RF and NN implementations have proven to be near equals in terms of the solutions they provide. One factor that can be considered in choosing an ML model is the ability to "look under the hood" (e.g., McGovern et al. 2019; Vilone and Longo 2020), and this might give a slight edge to the RF model since it is possible to view the individual decisions trees

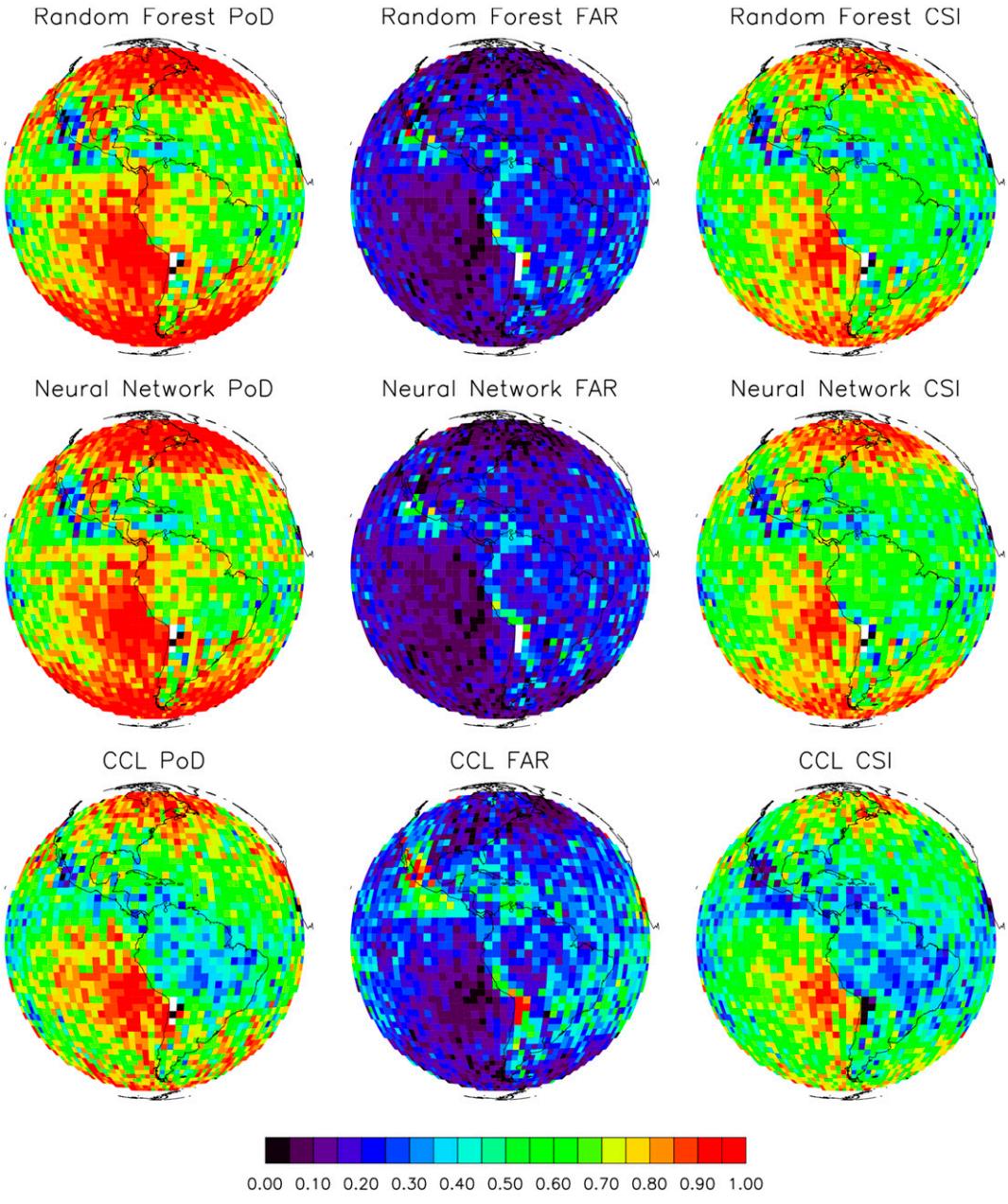

FIG. 6. Maps of (left) PoD, (center) FAR, and (right) CSI for (top) RF, (middle) NN, and (bottom) CCL.

and analyze which input features were most heavily used in the decision-making process, and whether these features were most influential near the base of the trees versus in the branches. A particularly useful way to evaluate the importance of the features used by a model is by calculation of the permutation importance (Breiman 2001). To do so, each feature is randomly shuffled, one at a time, and the performance of the algorithm is reevaluated in terms of some scoring metric. This has the advantage of retaining the distribution of a given feature, while preventing the feature from contributing to the final model in a meaningful way.

We calculated the permutation feature importance of the trained RF model using the mean accuracy and show the results in Fig. 8. The most important information for the trained model originated in the 0.47 μm visible channel, followed by the 2.2 μm "cloud particle size" near-infrared channel, the 1.37 μm "cirrus band," and then the low-level RH given by RH_{max}. There are numerous pathways by

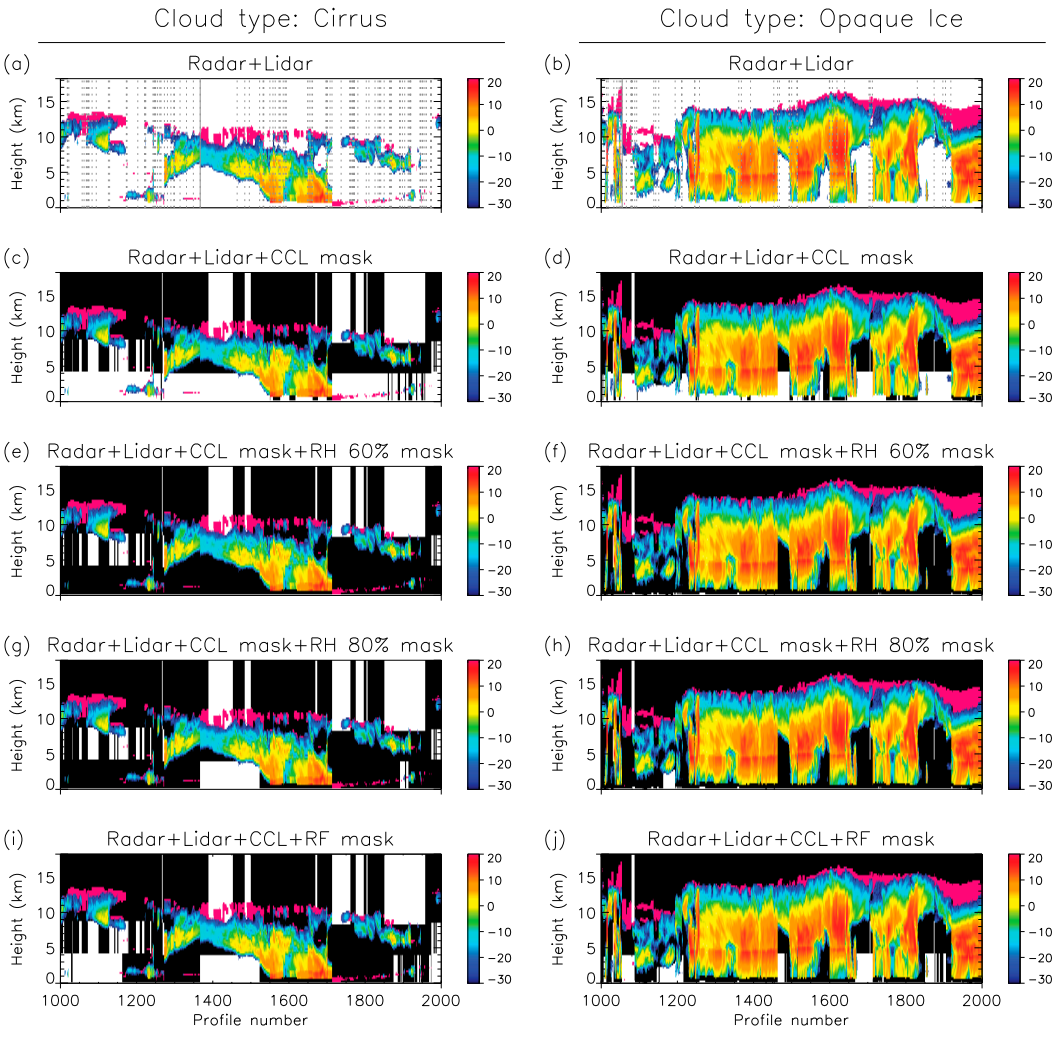

FIG. 7. Example radar and lidar profiles for CLAVR-x cloud types (left) "cirrus" and (right) "opaque ice." Rainbow colors are equivalent radar reflectivity (dBZ) and purple shading is lidar cloud mask where there is no radar return. (a),(b) The radar and lidar profile, with vertical gray dashed lines indicating a break between consecutive profiles. (c),(d) Cloud-layer detection by the CCL algorithm with a black background; recall that detection applies across a broad H, M, or L layer. The remaining plots supplement CCL cloud detection below 631 hPa with (e),(f) $RH_{max} > 60\%$, (g),(h) $RH_{max} > 80\%$, and (i),(j) the new RF prediction.

which the trained RF model might use the information from these bands, but one physical scenario stands out: the 0.47 μm radiance is related to the column-integrated cloud water, the 1.37 μm band is sensitive to high clouds, and the 2.2 μm band includes information that can be used to discriminate between large ice particles and small water drops. It is therefore possible that these bands are used by the algorithm to discriminate scenes containing thin high clouds with underlying low clouds, from similar scenes without low clouds. For example, a scene with thin cirrus overlying thick water cloud might have relatively high radiance in all three of these bands, but the same scene without the lower cloud would feature only an elevated 1.37 μm radiance (and RH_{max} would possibly be low as well). By contrast, the least important features used by the trained RF model are those that characterize surface type. They were initially included in hopes that they would assist the algorithm in identifying between bright surfaces due to cloudiness and those due to snow or ice. It appears that this is of minimal importance to the trained algorithm. It is notable that five of the six visible bands appear before any infrared band in terms of permutation-based feature importance.

A shortcoming of permutation feature importance is that it is difficult to interpret when significant correlation is present between input features. This is because, when permutating only one feature at a time as we did to create Fig. 8, a feature

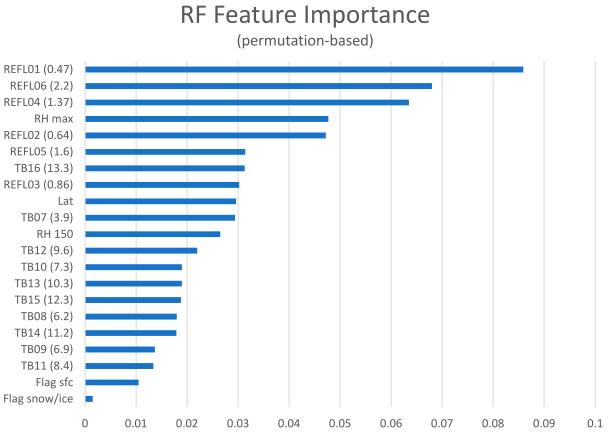

FIG. 8. Permutation-based feature importance for the RF model, shown as the difference in mean accuracy caused by the shuffling of each feature. See Table 1 for variable names; numbers in parenthesis are band centers in micrometers.

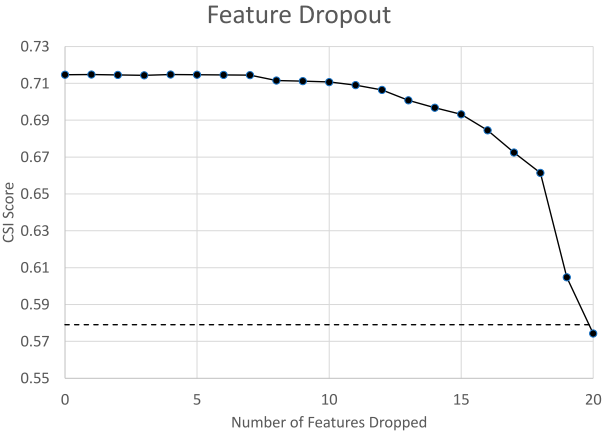

FIG. 9. CSI score for low cloud detection of RF model as features are successively dropped. Features are dropped in order from least important to most important as given in Fig. 8. The dashed line represents CCL.

may be deemed unimportant because (i) it is redundant, i.e., its information is important but already contained in the remaining features; (ii) it is irrelevant, i.e., its information would not be helpful even if none of the other features are included; or (iii) a combination of the two. Disentangling these cases for each feature is fairly complex as it requires testing the effect of simultaneous permutation (or other removal) of a large number of combinations of features. Common methods include sequential forward/backward selection (McGovern et al. 2019) and causal discovery (Ebert-Uphoff and Deng 2012). However, here we take a simpler approach only requiring few tests that nevertheless provides powerful insights. Namely, we successively drop features and train new RF models, starting from the least important feature and proceeding until only the most important feature remains. The CSI score for low cloud detection as a function of number of features dropped is shown in Fig. 9. Consistent with the above findings, all of the infrared-only channels can be dropped with only minimal impact on model performance (corresponding to a CSI decrease from 0.715 to 0.697 with 14 features dropped). However, as more important features are dropped, performance quickly deteriorates. For example, the effect of dropping any one feature from the sixteenth onward is greater than the cumulative effect of removing the 12 least important features simultaneously.

d. Algorithm implementation and variations

There are several ways that this algorithm can be implemented on near-real-time data. First, for operational implementation, Fig. 9 suggests that the algorithm can operate with a much smaller feature subset, which would lessen calculation time; the ideal algorithm would seek a balance between calculation time and required accuracy. Next, we note that the ML-based low cloud determination stands on its own, requiring only the satellite and GFS inputs described in Table 1. Since it uses a completely different approach than the CLAVR-x CCL product, it stands to reason that it can be combined with CCL by supplementing the CCL-based low cloud mask with the RF-based mask. In doing so, we do not allow clear pixels to be reclassified by RF as cloudy, for consistency with the original method. This combination methodology (CCL + RF) was already demonstrated in Figs. 7i and 7j. The overall performance is shown statistically by the light blue bars in Figs. 3–5. In general, it allows detection capabilities similar to RF, but at the expense of increased false alarms relative to RF (and dominated by values for CCL alone for those cloud types where CCL already has a large number of false alarms).

An example implementation of the RF + CCL combination on a full disk *GOES-16* ABI image is shown in Fig. 10. A large number of clouds previously classified as H + M have been reclassified as H + M + L. Two notable examples are the midlatitude cyclone over central North America, and the frontal system over the Southern Ocean in the lower left of the image. In other locations, M pixels have been reclassified as L + M. The pink colors mark the appearance of the new H + L category, which was not possible with CCL alone.

Although the focus of this work is daytime low cloud identification, the general methodology can be extended to other inputs, targets, and times of day. To demonstrate this, we created RF models with slightly different inputs and conducted some additional "toy experiments." The results of these experiments, in addition to those of the main work (the "full" algorithm), are shown on the performance diagram in Fig. 11. This diagram shows success ratio (1 − FAR) versus PoD, with CSI in curved contours and frequency bias as diagonal lines.

The results from the full algorithm are shown as orange dots; each orange dot represents a result corresponding to a

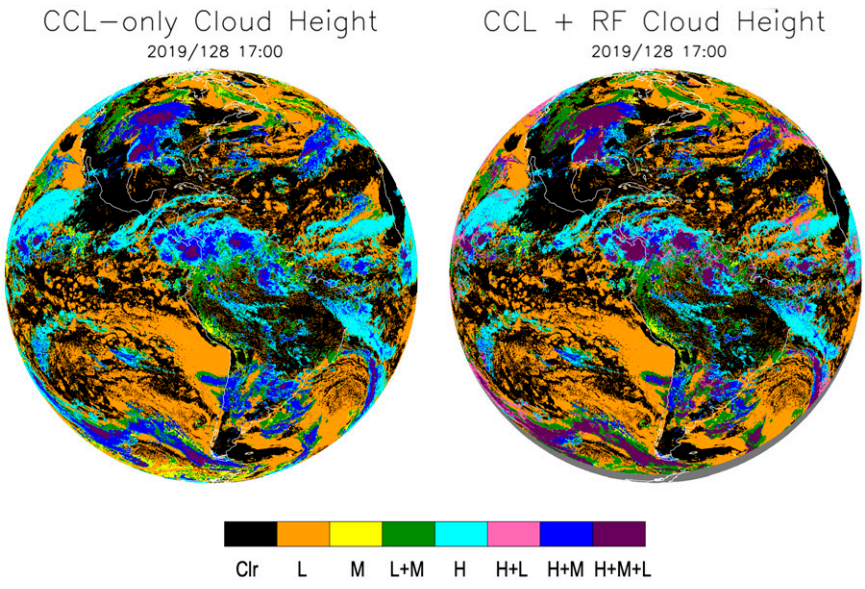

FIG. 10. Spatial comparison of low cloud detection between (left) CCL and (right) CCL + RF from full-disk ABI scan at 1700 UTC 8 May 2019.

probability of low cloud (P) between 0.1 and 0.9. As discussed earlier, a value of $P = 0.5$ has been used in all results until present; that point is represented by the middle orange dot that is closest to the diagonal line designating a frequency bias of unity. The spread demonstrates that different thresholds can be adopted depending on whether one's goal is to maximize detections or minimize false alarms.

Next, we note that the results presented in this work are applicable to daylit portions of Earth, and it has been demonstrated that the infrared channels have little influence on the trained model when the visible channels are present. Therefore, we might ask: can a similar model be trained excluding the visible channels only, and still provide useful information? Perhaps surprisingly, given the previous statement, the answer is yes. To test the feasibility, we remove ABI channels 1–7 from the training (the 3.9 μm channel is also excluded, as it responds to both reflected sunlight and emitted thermal radiation). The resulting model performs as shown by the solid blue dots (labeled "nighttime") in Fig. 11. Though the overall CSI decreases from 0.715 for the full algorithm to 0.656 for the IR-only algorithm (for $P = 0.5$), this still represents an improvement over CCL, whose CSI is 0.579 and which has a frequency bias farther from unity. Therefore, we find that an accompanying nighttime algorithm can be implemented with only a minor decrease in detection and corresponding modest increase in false alarms.

Oftentimes ML model performance can be enhanced by filtering the input data in ways that we expect may be useful for the problem at hand. For example, we can train our model with ABI channel ratios and differences that we suspect may be relevant for low cloud detection. We created three additional "toy" experiments to investigate such combinations. Experiment *R16 + 3* uses all inputs as the full model, plus 1) the 10.3–12.3 μm brightness temperature difference, which is commonly used in cloud classification algorithms (e.g., Purbantoro et al. 2018); 2) the 3.9–11.2 μm brightness temperature difference, often used for fog and low cloud detection; and 3) the 1.37 to 0.64 μm reflectance ratio. Since 1.37 μm is in a water vapor absorption band, its sensitivity to clouds will depend on their height, with relatively low sensitivity to lower-level clouds compared to 0.64 μm in a moist atmosphere. Experiment *R16 + 5* adds two more features, the 4) 6.9–7.3 and 5) 7.3–13.3 μm brightness temperature differences. The 6.9 and 7.3 μm channels are in water vapor absorption bands, with peak weighting functions in the middle and lower troposphere, respectively; the difference becomes small when a relatively dry mid- to upper troposphere overlies optically thick clouds extending up to the midlevels (Hirose et al. 2019). Finally, experiment *Rmixed* is the same as *R16 + 5* except that ABI channels that form differences or ratios are excluded from also serving as independent inputs. Results of these three experiments are shown in Fig. 11, again for $P = 0.5$. Remarkably, the overall performance of each of these experiments is nearly identical to the full algorithm. We therefore find that, for our application, providing the ABI channel data directly performs approximately the same as when we also (or alternately) providing some key ratios and differences.

It is also noteworthy that the full model can be used for mid- or high-level cloud detection (or any other combination of vertical levels) by simply adjusting the cloud height level used as "truth" during training. With this simple

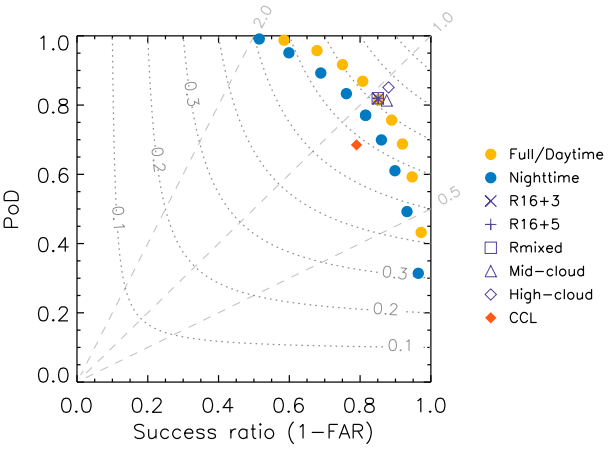

FIG. 11. Performance diagram summarizing performance of RF models discussed in the text as evaluated on the testing dataset. The solid dots show performance of the full (daytime) and nighttime algorithms for probabilities of low cloud (P) varying from 0.1 to 0.9 in 0.1 increments (the "topmost" dots represent 0.1, and the "bottommost" represent 0.9). CCL shown by red diamond; other symbols show performance of various models for $P = 0.5$. CSI is displayed as dotted curved contours, and frequency bias as dashed lines.

change, the resulting performance for mid- (M) and high-level (H) cloud is shown in Fig. 11 for $P = 0.5$. For both cases, CSI is increased relative to the low cloud algorithm (though with a degraded frequency bias for midlevel cloud). It is not surprising that the algorithm would perform better at higher levels in the atmosphere, since the probability of obscuration by upper-level clouds is decreased, and recent works have already established the general utility of ML-based decision tree and neural network methodologies for cloud masking (e.g., White et al. 2021; McCandless and Jiménez 2020; Kilpatrick et al. 2019; Hollstein et al. 2016).

Finally, although this work has focused on ABI, it has been demonstrated that using subsets of the full set of channels still results in a low cloud retrieval that improves upon the baseline CCL in multilayer scenes. It should therefore be possible to apply the same methodology to other sensors, like VIIRS, MODIS, or other ABI-like sensors in geostationary orbit, by either training the algorithm with the specific channels applicable to those instruments, or by using similar channels. Most ABI bands share overlap with the bands on VIIRS, but there are some notable exceptions in the infrared: bands 8–10 (6.2–7.3 μm water vapor bands), 12 (9.6 μm ozone band), 14 (11.2 μm infrared window band), and 16 (13.3 μm CO_2 band). None of these bands appear in the top half of most significant features in Fig. 8, with the exception of band 16. However, band 16 is highly correlated with band 15, and removing band 16 while retaining 15 results in virtually no change in performance (CSI decrease of 0.0001).

5. Conclusions

The detection of low-level clouds in the atmosphere has proved particularly challenging from satellite-based sensors when multiple cloud layers are present. The machine learning methods described in this work have shown skill in detecting low cloud in multilayer scenes. Inputs include visible reflectances, infrared brightness temperatures, estimates of relative humidity at multiple levels in the atmosphere, and information about the underlying surface. These methods can be used in a "stand-alone" mode, or they can be combined with the existing CCL representation of three-dimensional cloudiness, providing significant improvement in the overall characterization of vertical cloud structure.

It has been demonstrated that these machine learning methods can be applied in either daytime or nighttime mode and are likely to be applicable to a variety of other sensors besides ABI. They are also applicable to other levels in the atmosphere besides just low clouds, although these clouds were the focus of this work due to their operational forecasting significance. As such, one could envision that a complete three-dimensional retrieval, independent of the existing CCL algorithm, could be developed using the principals outlined in this work. Such a retrieval would likely need to be more than a simple "stacking" of the ML-based methods applied independently to multiple levels, as cloudiness at one vertical level is likely to be highly correlated with cloudiness in adjacent levels. In particular, it would be beneficial for such an algorithm to learn and predict the cloud mask for the entire column, rather than for just one level in the column. This use of multioutput learning is a fairly straightforward extension of the method described in this paper, and would expand prediction capabilities beyond just low clouds. Finally, we note that by using regression instead of classification models, it should be possible for the ML algorithm to predict cloud fraction in these vertical levels, instead of a simple cloud mask.

The CCL products for ABI and VIIRS are currently displayed in the Satellite Loop Interactive Data Explorer in Real Time (SLIDER; Micke 2018) web application.[4] As SLIDER serves as something of a test bed for products in transition to operations, we plan to display the machine learning–based enhancement described in this paper as an experimental product in the future.

Acknowledgments. This work was funded by the GOES-R Program Office under Grant NA19OAR4320073. The *CloudSat* products were obtained from the *CloudSat* Data Processing Center (http://www.cloudsat.cira.colostate.edu). The GFS data are available at https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs.

Data availability statement. The training and testing data used in this study are available in the long-term archive at https://doi.org/10.7910/DVN/LPXYBL. They are available in self-documented netCDF format.

[4] https://rammb-slider.cira.colostate.edu.

REFERENCES

Andersen, H., and J. Cermak, 2018: First fully diurnal fog and low cloud satellite detection reveals life cycle in the Namib. *Atmos. Meas. Tech.*, **11**, 5461–5470, https://doi.org/10.5194/amt-11-5461-2018.

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl, 2011: Algorithms for hyper-parameter optimization. *24th Int. Conf. on Neural Information Processing Systems*, Granada, Spain, NeurIPS, 2546–2554, https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

——, D. Yamins, and D. Cox, 2013: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. 30th Int. Conf. on Machine Learning*, Atlanta, GA, PMLR, 115–123, http://proceedings.mlr.press/v28/bergstra13.html.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Burkov, A., 2019: *The Hundred-Page Machine Learning Book.* Burkov, 141 pp.

Chollet, F., 2018: *Deep Learning With Python.* Manning, 361 pp.

Cronk, H., and P. Partain, 2017: CloudSat ECMWF-AUX auxiliary data product process description and interface control document. Cooperative Institute for Research in the Atmosphere Doc., 11 pp., https://www.cloudsat.cira.colostate.edu/cloudsat-static/info/dl/ecmwf-aux/ECMWF-AUX_PDICD.P_R04.20070718.pdf.

Ebert-Uphoff, I., and Y. Deng, 2012: Causal discovery for climate research using graphical models. *J. Climate*, **25**, 5648–5665, https://doi.org/10.1175/JCLI-D-11-00387.1.

Géron, A., 2019: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* 2nd ed. O'Reilly, 819 pp.

Groisman, P. Y., R. S. Bradley, and B. Sun, 2000: The relationship of cloud cover to near-surface temperature and humidity: Comparison of GCM simulations with empirical data. *J. Climate*, **13**, 1858–1878, https://doi.org/10.1175/1520-0442(2000)013<1858:TROCCT>2.0.CO;2.

Heidinger, A., and W. C. Straka III, 2013: Algorithm theoretical basis document: ABI cloud mask. NOAA/NESDIS Center for Satellite Applications and Research Tech. Rep., 106 pp., http://www.star.nesdis.noaa.gov/goesr/docs/ATBD/Cloud_Mask.pdf.

Hillger, D., and Coauthors, 2013: First-light imagery from *Suomi NPP* VIIRS. *Bull. Amer. Meteor. Soc.*, **94**, 1019–1029, https://doi.org/10.1175/BAMS-D-12-00097.1.

Hirose, H., S. Shige, M. K. Yamamoto, and A. Higuchi, 2019: High temporal rainfall estimations from Himawari-8 multiband observations using the random-forest machine-learning method. *J. Meteor. Soc. Japan*, **97**, 689–710, https://doi.org/10.2151/jmsj.2019-040.

Hollstein, A., K. Segl, L. Guanter, M. Brell, and M. Enesco, 2016: Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sens.*, **8**, 666, https://doi.org/10.3390/rs8080666.

Kilpatrick, K. A., G. Podestá, E. Williams, S. Walsh, and P. J. Minnett, 2019: Alternating decision trees for cloud masking in MODIS and VIIRS NASA sea surface temperature products. *J. Atmos. Oceanic Technol.*, **36**, 387–407, https://doi.org/10.1175/JTECH-D-18-0103.1.

Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff, 2021: Applying machine learning methods to detect convection using Geostationary Operational Environmental Satellite-16 (GOES-16) Advanced Baseline Imager (ABI) data. *Atmos. Meas. Tech.*, **14**, 2699–2716, https://doi.org/10.5194/amt-14-2699-2021.

Leinonen, J., A. Guillaume, and T. Yuan, 2019: Reconstruction of cloud vertical structure with a generative adversarial network. *Geophys. Res. Lett.*, **46**, 7035–7044, https://doi.org/10.1029/2019GL082532.

Mace, G. G., and Q. Zhang, 2014: The CloudSat radar-lidar geometrical profile product (RL-GeoProf): Updates, improvements, and selected results. *J. Geophys. Res. Atmos.*, **119**, 9441–9462, https://doi.org/10.1002/2013JD021374.

——, ——, M. Vaughan, R. Marchand, G. Stephens, C. Trepte, and D. Winker, 2009: A description of hydrometeor layer occurrence statistics derived from the first year of merged CloudSat and CALIPSO data. *J. Geophys. Res.*, **114**, D00A26, https://doi.org/10.1029/2007JD009755.

Marchand, R., G. G. Mace, T. Ackerman, and G. Stephens, 2008: Hydrometeor detection using *CloudSat*—An Earth-orbiting 94-GHz cloud radar. *J. Atmos. Oceanic Technol.*, **25**, 519–533, https://doi.org/10.1175/2007JTECHA1006.1.

McCandless, T., and P. A. Jiménez, 2020: Examining the potential of a random forest derived cloud mask from GOES-R satellites to improve solar irradiance forecasting. *Energies*, **13**, 1671, https://doi.org/10.3390/en13071671.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Micke, K., 2018: Every pixel of *GOES-17* imagery at your fingertips. *Bull. Amer. Meteor. Soc.*, **99**, 2217–2219, https://doi.org/10.1175/BAMS-D-17-0272.1.

Nayak, M., 2012: CloudSat anomaly recovery and operational lessons learned. *SpaceOps 2012 Conf.*, Stockholm, Sweden, American Institute of Aeronautics and Astronautics.

Noh, Y.-J., and Coauthors, 2017: Cloud-base height estimation from VIIRS. Part II: A statistical algorithm based on A-Train satellite data. *J. Atmos. Oceanic Technol.*, **34**, 585–598, https://doi.org/10.1175/JTECH-D-16-0110.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, https://www.jmlr.org/papers/v12/pedregosa11a.html.

Platnick, S., and Coauthors, 2017: The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua. *IEEE Trans. Geosci. Remote Sens.*, **55**, 502–525, https://doi.org/10.1109/TGRS.2016.2610522.

Purbantoro, B., J. Aminuddin, N. Manago, K. Toyoshima, N. Lagrosas, J. T. S. Sumantyo, and H. Kuze, 2018: Comparison of cloud type classification with split window algorithm based on different infrared band combinations of Himawari-8 satellite. *Adv. Remote Sens.*, **7**, 218–234, https://doi.org/10.4236/ars.2018.73015.

Qin, Y., A. D. L. Steven, T. Schroeder, T. R. McVicar, J. Huang, M. Cope, and S. Zhou, 2019: Cloud cover in the Australian region: Development and validation of a cloud masking, classification and optical depth retrieval algorithm for the Advanced Himawari Imager. *Front. Environ. Sci.*, **7**, 20, https://doi.org/10.3389/fenvs.2019.00020.

Russell, S. J., P. Norvig, and E. Davis, 2010: *Artificial Intelligence: A Modern Approach.* 3rd ed. Prentice Hall, 1132 pp.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI

on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

Sedlar, J., L. D. Riihimaki, K. Lantz, and D. D. Turner, 2021: Development of a random forest cloud regime classification model based on surface radiation and cloud products. *J. Appl. Meteor. Climatol.*, **60**, 477–491, https://doi.org/10.1175/JAMC-D-20-0153.1.

Seinfeld, J. H., and S. N. Pandis, 2006: *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change.* 2nd ed. Wiley, 1203 pp.

Shang, H., and Coauthors, 2018: Diurnal cycle and seasonal variation of cloud cover over the Tibetan Plateau as determined from Himawari-8 new-generation geostationary satellite data. *Sci. Rep.*, **8**, 1105, https://doi.org/10.1038/s41598-018-19431-w.

Stephens, G. L., and Coauthors, 2008: CloudSat mission: Performance and early science after the first year of operation. *J. Geophys. Res.*, **113**, D00A18, https://doi.org/10.1029/2008JD009982.

——, D. Winker, J. Pelon, C. Trepte, D. Vane, C. Yuhas, T. L'Ecuyer, and M. Lebsock, 2018: *CloudSat* and *CALIPSO* within the A-Train: Ten years of actively observing the Earth system. *Bull. Amer. Meteor. Soc.*, **99**, 569–581, https://doi.org/10.1175/BAMS-D-16-0324.1.

Stöckli, R., J. S. Bojanowski, V. O. John, A. Duguay-Tetzlaff, Q. Bourgeois, J. Schulz, and R. Hollmann, 2019: Cloud detection with historical geostationary satellite sensors for climate applications. *Remote Sens.*, **11**, 1052, https://doi.org/10.3390/rs11091052.

Tanelli, S., S. L. Durden, E. Im, K. S. Pak, D. G. Reinke, P. Partain, J. M. Haynes, and R. T. Marchand, 2008: CloudSat's cloud profiling radar after two years in orbit: Performance, calibration, and processing. *IEEE Trans. Geosci. Remote Sens.*, **46**, 3560–3573, https://doi.org/10.1109/TGRS.2008.2002030.

Tompkins, A. M., 2003: Impact of temperature and humidity variability on cloud cover assessed using aircraft data. *Quart. J. Roy. Meteor. Soc.*, **129**, 2151–2170, https://doi.org/10.1256/qj.02.190.

Vilone, G., and L. Longo, 2020: Explainable artificial intelligence: A systematic review. arXiv, https://arxiv.org/abs/2006.00093.

Walcek, C. J., 1994: Cloud cover and its relationship to relative humidity during a springtime midlatitude cyclone. *Mon. Wea. Rev.*, **122**, 1021–1035, https://doi.org/10.1175/1520-0493(1994)122<1021:CCAIRT>2.0.CO;2.

Wang, C., S. Platnick, K. Meyer, Z. Zhang, and Y. Zhou, 2020: A machine-learning-based cloud detection and thermodynamic-phase classification algorithm using passive spectral observations. *Atmos. Meas. Tech.*, **13**, 2257–2277, https://doi.org/10.5194/amt-13-2257-2020.

Wang, T., E. J. Fetzer, S. Wong, B. H. Kahn, and Q. Yue, 2016: Validation of MODIS cloud mask and multilayer flag using CloudSat-CALIPSO cloud profiles and a cross-reference of their cloud classifications. *J. Geophys. Res. Atmos.*, **121**, 11 620–11 635, https://doi.org/10.1002/2016JD025239.

White, C. H., A. K. Heidinger, and S. A. Ackerman, 2021: Evaluation of Visible Infrared Imaging Radiometer Suite (VIIRS) neural network cloud detection against current operational cloud masks. *Atmos. Meas. Tech.*, **14**, 3371–3394, https://doi.org/10.5194/amt-14-3371-2021.

Wind, G., S. Platnick, M. D. King, P. A. Hubanks, M. J. Pavolonis, A. K. Heidinger, P. Yang, and B. A. Baum, 2010: Multilayer cloud detection with the MODIS near-infrared water vapor absorption band. *J. Appl. Meteor. Climatol.*, **49**, 2315–2333, https://doi.org/10.1175/2010JAMC2364.1.

Winker, D. M., W. H. Hunt, and M. J. McGill, 2007: Initial performance assessment of CALIOP. *Geophys. Res. Lett.*, **34**, L19803, https://doi.org/10.1029/2007GL030135.

Yu, Z., S. Ma, D. Han, G. Li, D. Gao, and W. Yan, 2021: A cloud classification method based on random forest for FY-4A. *Int. J. Remote Sens.*, **42**, 3353–3379, https://doi.org/10.1080/01431161.2020.1871098.