Evaluation of VIIRS Neural Network Cloud Detection against Current Operational Cloud Masks

Charles H. White[1], Andrew K. Heidinger[2], and Steven A. Ackerman[1]

[1]University of Wisconsin - Madison, Department of Atmospheric and Oceanic Sciences, Madison, WI, USA
[2]NOAA/NESDIS/Center for Satellite Applications and Research, Madison, WI, USA

Correspondence: Charles H. White (cwhite25@wisc.edu)

Abstract. Cloud properties are critical to our understanding of weather and climate variability, but their estimation from satellite imagers is a nontrivial task. In this work, we aim to improve cloud detection which is the most fundamental cloud property. We use a neural network applied to Visible Infrared Imaging Radiometer Suite (VIIRS) measurements to determine whether an imager pixel is cloudy or cloud-free. The neural network is trained and evaluated using four years (2016-2019) of coincident

5   measurements between VIIRS and the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP). We successfully address the lack of sun glint in the collocation dataset with a simple semi-supervised learning approach. The results of the neural network are then compared with two operational cloud masks: the MODIS-VIIRS Continuity Cloud Mask (MVCM) and the NOAA Enterprise Cloud Mask (ECM).

We find that the neural network outperforms both operational cloud masks in most conditions examined with a few excep-

10   tions. The largest improvements we observe occur during the night over snow or ice covered surfaces in the high latitudes. In our analysis, we show that this improvement is not solely due to differences in optical depth-based definitions of a cloud between each mask. We also analyze the differences in true positive rate between day/night and land/water scenes as a function of optical depth. Such differences are a contributor to spatial artifacts in cloud masking and we find that the neural network is the most consistent in cloud detection with respect to optical depth across these conditions. A regional analysis over Greenland

15   illustrates the impact of such differences and shows that they can result in mean cloud fractions with very different spatial and temporal characteristics.

1   Introduction

Clouds serve many critical roles in the earth's weather and climate system, and are one of the largest sources of uncertainty

20   in future climate scenarios (Stocker et al., 2013). Determining their presence in current observational records is a fundamental first step in understanding their variability and impact. Polar-orbiting satellite imagers such as the Visible Infrared Imaging Radiometer Suite (VIIRS; Cao et al., 2013) offer frequent views of global cloud cover at high spatial resolution. However,

cloud detection from passive visible and infrared observations is a nontrivial problem. This is particularly true for clouds with low optical depths, and clouds above cold and visibly reflective surfaces (Ackerman et al., 2008; Holz et al., 2008). These
25   qualifications on imager cloud detection make it difficult to construct confident observational analyses of cloud variability from passive satellite instruments especially in the polar regions. As a result, many differences exist between cloud climate records made with different algorithms, or sensors with different capabilities (Stubenrauch et al., 2013).

Machine learning (ML) has become a popular tool for statistical modeling in earth sciences including the use of both supervised and unsupervised methods. Supervised ML methods in the earth sciences can require large amounts training data
30   often created from physically-based models, obtained from manual labeling, or observed from other instrument platforms. These approaches have been extensively used in characterizing the surface and atmosphere from remote sensing instruments. A sample of popular ML approaches (and their applications) used in satellite meteorology include naïve bayesian classifiers (Uddstrom et al., 1999; Heidinger et al., 2012; Cintineo et al., 2014; Bulgin et al., 2018), random forests (Kühnlein et al., 2014; Thampi et al., 2017; Wang et al., 2020), and neural networks (Minnis et al., 2016; Häkansson et al., 2018; Wimmers et al.,
35   2019; Marais et al., 2020).

In this analysis, we develop a neural network cloud mask that uses the moderate resolution channels from VIIRS to determine whether a given imager pixel contains a cloud or is cloud-free. We train the neural network using observations from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP; Winker et al., 2009). Observations from CALIOP are often used to validate cloud masks and cloud property estimates due to the instrument's ability to retrieve vertical profiles of the
40   atmosphere and characterize clouds with low optical depth. Additionally, its placement in the A-train constellation makes it a convenient reference for Moderate Resolution Imaging Spectroradiometer (MODIS) cloud property validation (Holz et al., 2008). The Suomi National Polar-orbiting Partnership (SNPP) VIIRS instrument, despite not being in the A-train constellation, makes spatially and temporally coincident observations with CALIOP roughly every two days. Thus, there is opportunity for matching observations between these two sensors with some limitations. One such limitation is that the range of atmospheric
45   and surface conditions sampled by CALIOP do not necessarily match that of SNPP-VIIRS. Conditions where collocations between these two sensors occur are even less representative, and do not contain instances of significant sun glint. In our training of a neural network cloud mask, we demonstrate how very a simple semi-supervised learning approach can ameliorate this specific limitation.

There are several recent applications of machine learning in characterizing clouds from imager observations that use CALIOP
50   as a source of labeled data. Perhaps most relevant is Wang et al. (2020) in which several random forest (RF) models are trained to identify the presence and phase of clouds from VIIRS observations under somewhat idealized conditions (spatially homogeneous and low aerosol optical depths). In such conditions the, RF models demonstrated improvements in cloud masking and cloud phase determination over current algorithms. Häkansson et al. (2018) uses CALIOP as a training source for estimating MODIS cloud-top heights with precomputed spatial features, MODIS brightness temperatures, and numerical weather pre-
55   diction (NWP) temperature profiles using a neural network. They additionally demonstrate the ability to accurately estimate cloud-top heights with channels only available on sensors such as the Advanced Very High Resolution Radiometer (AVHRR), and VIIRS. Similarly, Kox et al. (2014) trained a neural network with CALIOP to determine the presence of cirrus clouds

and estimate their optical depth and cloud-top height from SEVIRI observations. While the majority of these applications for cloud property estimates are relatively recent, there were successful implementations well before the launch of CALIOP using

60 manually labeled scenes (Welch et al., 1992).

Our approach aims to improve upon existing literature in several ways. Rather than relying on precomputed spectral, or textural features, we allow a neural network to learn relevant features from a local 3 pixel by 3 pixel image patch from all 16 moderate resolution VIIRS channels. This necessitates a relatively large neural network architecture in order to exploit the variability of these observations to discriminate cloudy from cloud-free scenes. We train the model without filtering CALIOP

65 collocations to encourage more reliable predictions under non-ideal conditions. Additionally, we specifically address issues caused by the lack of sun glint scenes in collocations between SNPP VIIRS and CALIOP. This specific implementation does not require surface temperatures, surface emissivity, the use of clear-sky radiative transfer modeling, snow cover or ice cover information. The only ancillary data used is a VIIRS-derived land/water mask in the level-1 geolocation product. This approach uses a single model for all surface types and solar illumination conditions and in some respects, greatly simplifies the processing

70 pipeline for imager cloud masking.

Current operational cloud masks for VIIRS include the NOAA Enterprise Cloud Mask (ECM; Heidinger et al., 2012; Heidinger et al., 2016), and the MODIS-VIIRS Continuity Cloud Mask (MVCM; Frey et al. 2019). The ECM algorithm was originally designed for AVHRR climate applications, and has since been extended to a wide range of geostationary and polar-orbiting imagers including VIIRS. The MVCM has heritage with the MODIS cloud mask (Ackerman et al., 2010), and has

75 been adjusted to only use channels available on both VIIRS and MODIS. Obtaining continuity in cloud detection between the two imagers is a specific goal of the MVCM.

In this analysis, we demonstrate that a neural network cloud mask outperforms both operational methods in detecting clouds identified by CALIOP. In particular, we note large improvements at night in the middle and high latitudes. Since cloud masks may have differing definitions of what substantiates a cloud, we evaluate the balanced accuracy of each approach after re-

80 moving clouds above an increasing lower optical depth threshold. The usefulness of the predicted probabilities as a proxy for uncertainties are assessed. We also show an example of how differences in cloud detection ability can result in vastly different spatial and temporal characteristics of regional mean cloud cover assessments in the polar regions.

2 Data

2.1 Collocation Methodology

85 The labeled data that is used to train and evaluate the performance of the neural network comes from version 4.2 of the 1 km CALIOP Cloud Layers product (Vaughan et al., 2009). A vertical profile is determined to be cloudy when the number of cloud layers is equal to or exceeds one. Otherwise the profile is assumed to be cloud-free. The CALIOP labels are set to zero for cloud-free observations, and one for cloudy observations. Other CALIOP information such as the cloud-top pressure and cloud feature type are used in the validation of the cloud masks. Cloud optical depth is obtained from the 5 km CALIOP Cloud

90 Layers product since it is unavailable at the 1 km resolution. There are difficulties in matching satellite imager measurements

with CALIOP. Many of these issues are discussed at length in Holz et al. (2008), and include differences in spatial footprint, viewing angle, the observation time between the two instruments, and the horizontal averaging applied within the CALIOP products to increase their signal to noise ratio.

Collocations between SNPP VIIRS and CALIOP are obtained by performing a nearest neighbors search between the 1 km CALIOP Cloud Layers product, and the 750 m (at nadir) VIIRS observations. A parallax correction is then applied to account for pixels with high altitude clouds that are observed at oblique viewing angles by VIIRS. The details of the parallax correction are identical to that of Holz et al. (2008). Collocations with times that differ by more than 2.5 minutes are removed. This is a particularly strict requirement relative to Heidinger et al. (2016) which uses a limit of 10 minutes and severely limits both the number of possible collocations between these instruments and the range of viewing conditions sampled. We make this choice because the time difference between observations is a critical factor in the representativeness of a CALIOP profile for a given imager pixel. This is particularly true for small clouds that occupy a horizontal area similar to or smaller than a single VIIRS pixel in environments with high wind speeds. Collocations are found for these instruments from January 2016 through December 2019. Some gaps in the collocation dataset exist and are primarily due to the availability of CALIOP data products. Following the recommendations from the CALIPSO team, we remove all CALIOP profiles that contain low-energy laser shots with 532 nm laser energies less than 80 mJ. This results in a relative sparsity of collocations over central South America after mid-2017. In total, roughly 29.5 million collocations were collected for this study with the above requirements.

2.2 Neural Network Inputs

The observations used as input into the neural network come from the moderate resolution channels (M1-M16) obtained from the NASA processing of SNPP VIIRS. An eight-category land/water mask included in the VNP03MOD geolocation product, which we reduce to a binary land/water mask is also used. Other inputs include the absolute value of latitude, solar zenith angle, and sun glint zenith angle. Sun glint zenith angle is the angle between the surface normal of the estimated specular point (the point of maximum sun glint) and atmospheric path viewed by VIIRS. The VIIRS/CrIS fusion channels (Weisz et al., 2017) are estimates of MODIS-like channels using coarse-resolution measurements from the Cross-track Infrared Sounder (CrIS) that are interpolated to match the moderate resolution channels of VIIRS. The VIIRS/CrIS fusion channels are used in a pseudo-labeling model for sun glint scenes (described later in section 3.1), but are not used in the final neural network model. For all of the channels used, a 3 pixel by 3 pixel array is extracted and is used to predict the cloudy or cloud-free label at the center pixel.

2.3 Dataset Splitting

In statistical modeling, it is important to ensure independence between the training, validation, and testing datasets. As mentioned previously, the CALIOP Cloud Layer product's feature identification algorithm often relies on horizontal averaging to detect cloud layers of low optical depth. This averaging increases the signal to noise ratio and allows for more accurate identification of such features. As a result, clouds with low optical depth may have their attributes replicated across neighboring CALIOP profiles. As pointed out in Häkansson et al. (2018), separating imager and CALIOP collocations by random sampling

would result in three nearly identical datasets and would yield a model that greatly overfits. To avoid this, we stratify our col-
125 locations by year into our training set that consists of 16.5 million collocations from 2016 and 2018, a validation set consisting
of 5.9 million collocations from 2017, and our testing set consisting of 7.1 million collocations from 2019. The training set is
what is supplied to the model during the training stage. The validation dataset is used for hyperparameter tuning during model
development and early stopping during training stage. The testing set is used to provide estimates of model performance which
we will analyze in section 4 and is not seen by the model during the training and hyperparameter tuning stages.

130 ## 2.4 CALIOP Data Preprocessing

A common preprocessing step when training imager cloud masks with CALIOP observations is to filter the collocations using
several heuristics in order to infer when CALIOP cloud detection is unreliable or unrepresentative of the corresponding imager
pixel. Heidinger et al. (2012) filters AVHRR collocations so that only CALIOP observations where the 5 km along-track cloud
fraction is equal to 0% or 100% are included. Holz et al. (2008) only retained MODIS pixels where all collocated CALIOP
135 retrievals are identical. Wang et al. (2020) require that both the 1 km and 5 km CALIOP Cloud Layer products agree, that five
consecutive 1 km CALIOP profiles agree, and they additionally remove profiles with high aerosol optical depths. Many of these
filters achieve a similar result in requiring that CALIOP profiles, to a varying degree, are spatially homogeneous with regards
to the presence of clouds. This filtering is often applied to remove fractionally cloudy profiles, or profiles where the clouds may
have moved out of the corresponding imager pixel. Karlsson et al. (2020) employ an approach that filters AVHRR/CALIOP
140 collocations on the basis of cloud optical depth. This is done in an iterative fashion in order to determine the lower optical
depth threshold in which their cloud masking method can reliability detect clouds.

In our approach, we intentionally do not perform any of the above preprocessing steps to our training dataset. This is because
we include a substantial amount of spatial information in our neural network inputs. If such a spatial filter were applied to the
CALIOP data, then cloud edges and small clouds (often boundary-layer clouds) would rarely occur in our training dataset.
145 This would yield a large amount of bias in a model that accounts for any amount of spatial variability and could cause it to
generalize poorly. Alternatively, we apply a spatial filter to only our testing dataset to create a second filtered testing dataset
that we can evaluate our models against. This allows us to evaluate the performance of our cloud masking model against others
using only the most reliable CALIOP collocations without biasing any model that considers spatial variability. Additionally, we
can analyze the performance of our neural network approach in fractionally cloudy scenes using the unfiltered testing dataset
150 with the knowledge that these collocations may be overall less reliable. The specific filter we apply to our testing dataset
requires that five consecutive 1 km profiles agree. This spatial filter creates a filtered testing dataset of 5.9 million collocations
compared to the unfiltered testing dataset of 7.1 million collocations. In no way does this filter affect the training or validation
data.

3 Methods

3.1 Pseudo-Labeling Procedure

A general concern in using statistical models such as neural networks, is the ability for them to generalize to unseen data. One such scenario in this dataset is sun glint. Sun glint is the specular reflection of visible light usually over water surfaces which results in very large visible reflectivity for both cloudy and cloud-free scenes. In our dataset of VIIRS/CALIOP collocations, we never observe any substantial amount of sun glint. Thus, without accounting for sun glint, any statistical model will likely fail to make a reasonable assessment of cloud cover under these conditions. Often, this results in erroneously predicting cloud cover in sun glint regions due to their high visible reflectivity. In the ECM, sun glint is handled by turning off cloud tests that use visible and NIR channels with solar contributions. In the MVCM, this is handled by decision paths that use visible channels to detect clear-sky pixels specifically in sun glint regions. CLDPROP optical properties (which use the MVCM) also use a clear-sky restoral algorithm (Platnick et al., 2017) in an attempt to remove erroneously cloudy pixels, but it is not included in the MVCM output.

We aim to overcome this limitation by using a simple semi-supervised learning approach called pseudo-labeling (Lee, 2013). Pseudo-labeling is the approach of using a model to make predictions on unlabeled data, assuming that some or all of these predictions are correct, and adding these predictions to the original training dataset as if they were true labels. In our application, the pseudo-labeling model only uses VIIRS and VIIRS/CrIS fusion channels unaffected by sun glint, while the final model uses all VIIRS channels and no VIIRS/CrIS fusion channels.

We first train a neural network model using only channels that are unaffected by sun glint. For VIIRS, these channels are M14, M15, and M16 (central wavelengths of roughly 8.6 μm, 10.8 μm, and 12.0 μm respectively). In addition to these VIIRS channels, we also use some of the VIIRS/CrIS fusion estimates of MODIS-like channels (MODIS bands 27-36, with central wavelengths of 6.7 μm, 7.3 μm, 8.6 μm, 9.7 μm, 11.0 μm 12.0 μm 13.3 μm, 13.6 μm, 13.9 μm respectively). Similarly, these are only VIIRS/CrIS fusion channels that are unaffected by sun glint. The pseudo-labeling model is then used to make predictions for all SNPP VIIRS scenes for the 15th day of every month in 2018 that contain significant amounts of sun glint. Of these predictions, roughly one million pseudo-labels are randomly sampled without replacement and added to the original training and validation datasets. No pseudo-labels are added to the testing dataset.

The class probabilities for the pseudo-labeled examples are not required to be equal to 0 or 1. They are left unmodified in an effort to promote more reliable class probabilities in pixels affected by sun glint from the final neural network model. In section 4.5 we demonstrate the impact of training with these pseudo-labels on a single scene with a large amount of sun glint.

3.2 Neural Network Description and Training Details

We use a simple neural network model that consists of Fully Connected (FC) layers, Leaky Rectified Linear Unit activations (Leaky ReLU), Dropout (Srivastava et al., 2014) and a sigmoid activation as the last layer. The architecture of this model is described in Fig. 1. For each VIIRS pixel, a centered 3 pixel by 3 pixel image patch from all 20 inputs is passed to Layer Group

1 (LG1) of Fig. 1. All except the last FC layer are followed by Leaky ReLU activation and 2.5% Dropout. The last sigmoid activation bounds the output of the model between 0 (indicating cloud-free) and 1 (indicating cloudy).

The model in Fig. 1 is the result of a grid search over a fairly small set of hyperparameters. We tested several configurations by multiplying the number of units in all but the last FC layer by 0.25, 0.5, 1.0, and 2.0. We also tested dropout rates of 0%, 2.5%, 5%, and 10%, and Leaky ReLU vs. ReLU activations. This results in 32 model configurations which are each trained and evaluated three times. Two configurations with double the number of units in the FC layers reported slightly higher validation accuracies compared to that of Fig. 1 (a difference of 0.05%). However, we judged that the increase in prediction time was not worth the very small gains in performance. Across all model configurations, Leaky ReLU activation was better than ReLU. Dropout percentages larger than 2.5% only helped when models had a twice the number of units in the FC layers.

Data augmentation is a common method to artificially increase the diversity of examples in the training dataset (Shorten and Khoshgoftaar, 2019). This is often performed by creating plausible alternative views of training examples. Data augmentation methods have been critical in improving performance on widely-used computer vision benchmarks (Zhang et al., 2018, for example). In our case, we are limited by the chosen shape and nature of our input to the kinds of augmentations we can apply to our training dataset. For instance, we cannot reasonably scale, zoom, or translate (all common augmentations applied to images) a 3 pixel by 3 pixel image patch where the center values have special meaning. During training, we apply uniformly random 90 degree rotations (0, 90, 180, 270), horizontal flips, and vertical flips.

The neural network is trained to minimize binary cross-entropy, with a mini-batch size of 4,098 examples. All inputs are standardized by subtracting the mean and dividing by the standard deviation of each input calculated from the training dataset. The Adam optimizer is used with its suggested default parameters (Kingma and Ba, 2015), and we did not notice any substantial changes in the final model when other optimization algorithms were used. The learning rate is initially set to 5×10^{-3}. This value is selected using a learning rate range test (Smith, 2017). After each epoch, the model is evaluated on the validation set. The learning rate is reduced by a factor of 10 when the performance on the validation dataset does not improve for 3 epochs. This continues until a learning rate of 1×10^{-6} is reached. Training is stopped once the validation performance does not improve for 5 epochs. Both the final model, and the pseudo-labeling model are trained in the same way with the same set of hyperparameters. Although, since the input size is smaller, the pseudo-labeling model has fewer parameters in the first fully connected layer. Using the same set of hyperparameters is not necessarily ideal since the pseudo-labeling model may have a different set of optimal hyperparameters. We did not perform a separate hyperparameter grid search due to the large computational cost.

4 Results

4.1 Validation with CALIOP

When evaluating classification models many performance metrics need to be viewed in context of the class distribution. Otherwise, quantities such as accuracy (ACC) and true positive rate (TPR; equivalent to probability of detection) can be misleading. For example, a trivial binary classification model that predicts only the positive class achieves 0.9 ACC and 1.0 TPR in a dataset

with a positive/negative class distribution of 0.9 and 0.1 respectively. Thus, while metrics like ACC and TPR are useful, they
220 must be interpreted within the context of the mean cloud fraction.

We calculate the mean cloud fraction for all VIIRS/CALIOP collocations in our 2019 testing dataset over different surface
types for both day and night (Fig. 2). For each instance, a cloud fraction value is reported from CALIOP, the neural network
cloud mask, the MVCM and the ECM. Daytime cloud fractions include collocations where the solar zenith angle is less than 85
degrees. Land and water surface types are determined from the VIIRS level-1 geolocation data product. The presence of sea ice,
225 snow, and permanent snow (primarily Greenland and Antarctica) is determined from the National Snow and Ice Data Center
sea ice index included with the CALIOP Cloud Layer products. The cloud fraction estimates are not necessarily representative
of the true cloud fraction over these surface types since they only represent VIIRS/CALIOP collocations for 2019. Instead, we
use them to compare the relative tendencies of each cloud mask to generally overestimate or underestimate cloud cover for a
given surface type.

230 The neural network cloud fractions match closely to that of CALIOP with the exception of an underestimate of 7% over
nighttime permanent snow. In all other instances the neural network reports cloud fractions are within 3% of CALIOP. The
MVCM predicts smaller mean global cloud fraction compared to CALIOP. This seems to be due to a combination of slightly
overestimating cloud cover over daytime water, and underestimating cloud cover elsewhere. Of particular note are nighttime
snow scenes where MVCM underestimates by 17%, nighttime sea ice where it underestimates by 24%, and areas with per-
235 manent snow cover during the night where it underestimates by 30%. The ECM predicts roughly similar values to the neural
network with the exception of overestimating cloud cover during the night over sea ice by 12%.

$$TPR = \frac{TP}{P} \tag{1}$$

$$TNR = \frac{TN}{N} \tag{2}$$

$$ACC = \frac{TP + TN}{P + N} \tag{3}$$

240 $$BACC = \frac{TPR + TNR}{2} \tag{4}$$

In order to evaluate the performance of each cloud masking model, we calculate the balanced accuracy (BACC; Eq. 4) of
all cloud masks across each surface type examined in Fig. 2. BACC is the mean of the true positive rate (TPR; Eq. 1), and the
true negative rate (TNR; Eq. 2), where TP is the number of correctly identified clouds, P is the number of clouds, TN is the
number correctly identified of cloud-free scenes, and N is the number of cloud-free scenes. The advantage of using BACC over
245 ACC (Eq. 3) is that BACC accounts for class imbalance. One example of class imbalance is daytime sea ice scenes where the

mean CALIOP cloud fraction is 76%. A trivial model that predicts 100% cloud fraction would obtain 76% ACC, but only 50% BACC over daytime sea ice.

BACC is calculated for several surfaces separated for day and night (Table 1). The neural network cloud mask reports higher BACC over every surface type examined compared to both the ECM and MVCM. BACC is calculated on the filtered dataset so

250 Table 1 represents the most reliable VIIRS/CALIOP collocations. However, this means that fractionally cloudy scenes, cloud edges, and boundary layer clouds are not well represented here. The most notable improvement from the neural network occurs over sea ice, snow, and permanent snow during both day and night. McNemar's test (McNemar, 1947) is applied to the neural network and the best operational model (either ECM or MVCM) for each category in Table 1 with the null hypothesis that there is no difference in predictive performance between the two models. We reject the null hypothesis with a p-value less than

255 0.001 in every comparison of the neural network and the best operational model.

Cloud detection ability relies on many factors including the underlying surface and the characteristics of a given cloud. Clouds with low optical depth may have only a small impact on the top-of-atmosphere radiation observed by the imager. Similarly, clouds that are close to the surface, even if they are optically thick, may be difficult to identify due to low thermal contrast with the surface. We calculate the TPR for all collocations as a function of cloud-top pressure and cloud optical depth

260 as estimated from CALIOP (Fig. 3).

As expected, all cloud masks struggle with the identification of clouds that are optically thin and clouds that are close to the surface. The neural network has the largest TPR values across all cloud-top pressures and optical depths with a few exceptions. In the unfiltered dataset during the day, the MVCM has the highest TPR values for clouds with tops lower than 850 hPa. For the same cloud-top pressures, the neural network has the highest TPR in the filtered dataset. This may indicate that the MVCM

265 is better able to discriminate small clouds that are close to the surface. However, when these clouds are removed, the neural network detects a larger portion of the remaining clouds at all cloud-top pressures. During the night, the MVCM severely underestimates cloud cover for all cloud-top pressures lower than roughly 350 hPa. This is consistent with the overall lower mean cloud fraction for nighttime scenes reported in Fig. 2. When considering optical depth, the neural network consistently has a larger TPR for all values during the day and night for the filtered dataset. This is also true for the unfiltered dataset with

270 one exception where it is competitive with the MVCM at optical depths less than 0.2 during the day.

There are some differences between Fig. 3 and Table 1 that may seem unintuitive. For example, the ECM has much higher TPR during the night compared to the MVCM for all optical depths and all cloud-top pressures. However, its BACC values for all nighttime collocations is slightly less than that of the MVCM. In this case it is helpful to remember that BACC accounts for both clear and cloudy scenes, and weights each class equally. TPR only accounts for the proportion of clouds correctly

275 identified. The MVCM results in the TPR analysis of Fig. 3 appear to be to due to its tendency to underestimate cloud cover during the night over certain surfaces.

We also investigate the TPR of the three cloud masks as a function of cloud type (Fig. 4). The cloud types are obtained from the 1 km CALIOP Cloud Layers product. Overall, the neural network reports the highest TPR for most cloud types. One exception is the broken cumulus cloud type in the unfiltered dataset for which the MVCM has the highest TPR. This difference

280 for broken cumulus clouds implies that the neural network has relatively worse performance in fractionally cloudy scenes

compared to the MVCM. When examining the filtered dataset results for these same clouds, we see that the neural network has the highest TPR. This suggests that the neural network and the ECM are only better at detecting broken cumulus when they occupy a substantial horizontal area. When there is considerable fine-scale spatial variability, such as in the unfiltered dataset, these results suggest that the MVCM is the most likely to correctly detect a cloud. Besides the broken cumulus cloud type, the neural network has the highest TPR for both the filtered and unfiltered collocations. The largest differences are observed when comparing cloud masks for the transparent cloud types. Almost no differences are observed for deep convection which are likely optically thick and have high altitude cloud-tops.

As discussed previously, large TPR values do not necessarily indicate skilful models since they can be obtained by over predicting the positive class. The mean cloud fraction values from Fig. 2 offer some evidence that this is not the case for any of these cloud masks in most scenarios. To add additional context, we plot the receiver operating characteristic (ROC) curves under various geographic and solar illumination conditions (Fig. 5). The ROC curve of each cloud mask depicts the TPR and false positive rate (FPR) over a varying threshold applied to their class probabilities. The neural network and ECM both natively output cloud probabilities. The MVCM includes a clear-sky confidence estimate which we take the compliment of. An ideal model has a high TPR with very low FPR. A random classifier lies along the diagonal in the middle of a typical ROC plot where TPR is equal to FPR (not shown due to our choice of x and y axis limits).

Figure 5 indicates that the neural network can obtain higher TPR for any specified FPR in every scenario examined. This is true for both the filtered and unfiltered datasets. This result illustrates that the larger TPR values reported by the neural network are not strictly due to the larger mean cloud fraction compared the MVCM which predicts smaller mean global cloud fraction. In addition to Table 1, Fig. 5 implies that most of the improvement by the neural network comes from the high latitudes during the night, but small improvements can still be observed elsewhere. In every scenario the unfiltered results are worse than those of the filtered datasets. The largest discrepancy between the filtered and unfiltered datasets occurs in the low-latitudes over the ocean. This is likely due to the prevalence of small broken cumulus clouds that are mostly removed from the unfiltered dataset.

There are a few situations where the actual TPR and FPR of the models (marked by the colored circles in Fig. 5) are in unintuitive locations on the ROC curve. The ECM's FPR is larger than 40% for nighttime water scenes at the middle and high latitudes (not shown due to our choice of x-axis limits). We expect that this is related to the high mean cloud fraction over these regions measured by CALIOP. Given that the naïve Bayesian models behind the ECM require an initial guess, it is likely that the ECM is relying heavily on climatology in regions where cloud masking is difficult from infrared observations. Overall, it seems that the locations on the ROC curve of the actual TPR and FPR of the neural network are related to the mean cloud fraction of the different regions. This is particularly true for nighttime scenes, where statistical models may rely more heavily on the background mean cloud fraction. More cloudy regions such as middle and high latitude nighttime water (with cloud fractions of roughly 79%) have larger FPR. Conversely, nighttime low-latitude land (with a cloud fraction of 50%) has a much lower FPR. Applications that require specific TPR or FPR from a cloud mask could tune the thresholds applied to the cloud probabilities to reach their desired values indicated by the ROC curves.

Next, we examine the performance as a function of geographical region. The mean ACC on the filtered testing dataset is calculated on a 5 by 5 degree grid (Fig. 6). McNemar's test is used to test the differences in model performance between the

neural network and each operational model at every grid point. Only points with significant differences in model performance (p-values less than 0.001) are shown (Fig. 6.d, Fig. 6.f). Most large differences occur over high latitude land. In particular, the neural network shows large improvement (10-20% difference) over North America, Greenland, Northeastern Asia, and Antarctica over both the MVCM and ECM. Only small improvement (0-10% difference) is observed over the ocean at low and
320 middle latitudes compared to the MVCM. The neural network shows mixed results compared to the ECM in tropical ocean. A large contribution to the poor performance of the MVCM in the Arctic and Antarctic is likely due to the severe underestimation of cloud cover observed during the night at high latitudes.

All the of the previous analyses in this work rely heavily on an individual cloud mask's effective definition of cloud. A difficulty with comparing different clouds masks is that the definition of a cloud is somewhat subjective at low optical depths
325 and perhaps dependent the particular application. It is plausible that each cloud mask may be more effective at discriminating clouds around a certain optical depth threshold. Thus, a reasonable argument based on the reported global mean cloud fractions in Fig. 2, and the BACC values in Table 1, is that the MVCM, due to its lower global mean cloud fraction, may only be sensitive to clouds with slightly larger optical depths compared to the neural network model and ECM.

In order to further probe the differences in these cloud masks, we recalculate BACC after removing clouds below an increas-
330 ing lower optical depth threshold from our testing dataset (Fig. 7). The aim of this analysis is to understand how the optical depth of a cloud impacts its detectability by each approach, and identify if certain cloud masks perform better if we remove clouds with trivially low optical depths. Even if two cloud masks are developed around slightly different optical depth-based definitions of a cloud, we can reasonably expect their BACC values in Fig. 7 to converge when clouds with optical depths above both thresholds are removed.

335 As expected, when optically thin clouds are removed from our testing dataset, the BACC of all the cloud masks is improved. Consistent with Fig. 5, the filtered dataset has higher BACC for all scenarios. The neural network reports the highest BACC across all land/water, day/night, and latitude combinations examined with a few key exceptions. In low-latitude nighttime water scenes (Fig. 7.j), the ECM has larger BACC for every cloud optical depth threshold in the unfiltered dataset, but more similar values in the filtered dataset. In daytime land scenes at low latitudes (Fig. 7.a), the ECM has larger BACC values above an
340 optical depth threshold of roughly 0.4 for the unfiltered dataset, but has lower BACC values at most optical depths for the filtered dataset. The fact that the neural network BACC values are still equal to or larger than the other cloud masks for high optical depth clouds in most scenarios suggests the neural network is overall more skillful in cloud detection regardless of a reasonable optical-depth based definition of a cloud. Because of this, we can infer that improvement in BACC by the neural network in Table 1 is not solely due to discrepancies in the detection of optically-thin clouds.

345 It may be initially unintuitive why some of the curves in Fig. 7 vary so little with the removal of optically thin clouds. This is partially due to the choice of BACC as our performance metric, but it is also representative of the fact that cloud optical depth is not the only variable controlling the detectability of a cloud. Thermal contrast with the surface also plays a significant role. Often, this can be analysed by examining performance of a given cloud mask as a function of both optical depth and cloud-top height. However, this may be misleading where clouds in inversion layers may be warmer than the underlying surface. To
350 examine the approximate impact of thermal contrast with the surface, we calculate ACC as a function of the difference between

the VIIRS M15 measurement (10.8 μm) and the surface temperature obtained from the Global Forecasting System (GFS) analysis dataset (Fig. 8). We calculate ACC after the removal of clouds below two different cloud optical depth thresholds: 0.3, and 3.0. As expected, all cloud mask models perform well where the 10.8 μm measurement is significantly colder than the surface. Once the differences are greater than -16 K, the performance of all models decreases. Figure 8.b illustrates that

355   even for optically thick clouds, the performance of both operational models is largely dependent on thermal contrast with the surface. The neural network appears to be more robust to scenes where the 10.8 μm measurement is similar to or warmer than the surface. This is surprising given that the neural network is not supplied with any information about surface characteristics other than latitude and whether it is viewing a land or water surface.

4.2 Uncertainty Assessment

360   Class probabilities produced by machine learning models are often used to obtain uncertainty estimates. While these values are typically not the same as true uncertainties, they can be useful for interpreting model output. For binary classification models, an approximation for uncertainty can be usually obtained by examining the distance from the decision threshold. These uncertainty estimates are generally unreliable when predictions are made on inputs that are outside the distribution the original training dataset. With this significant caveat in mind, we calculate the ACC with respect to the cloud probabilities of

365   the neural network and ECM, as well as the clear-sky confidence from the MVCM (Fig. 9). A model with a cloud probability threshold of 0.5 is perfectly calibrated if its predictions lie along the line where $ACC = \min(\hat{y}, 1-\hat{y})$ where $\hat{y}$ is the scalar predicted cloud probability. The MVCM appears to follow a different convention with a decision threshold of 0.95 since that is where the minimum accuracy is reached with respect to the MVCM clear-sky confidence.

Overall, the neural network appears to be the best calibrated with the ACC on the unfiltered collocations closely following

370   the expected values from a perfectly calibrated model. It is slightly over-confident when predicting cloud probabilities for clear-sky cases in the range of 0.1 to 0.4. The ECM appears to be overconfident for the majority of cloud probability values. The assessment of MVCM accuracy as a function of clear-sky confidence is somewhat noisy, but could be attributed to the extremely low number of values in the calculated intervals. Despite the minor differences, all cloud masks examined here have accuracies that vary in an intuitive way with their predicted cloudy or clear-sky probability values. The differences among them

375   can be mostly attributed to how well their class probabilities correspond to a particular level of accuracy. As a result, we expect that these values can be used to convey the relative uncertainty in estimating which imager pixels the CALIOP cloud products might determine to be cloudy. However, it remains to be demonstrated if accurate uncertainties in predicting CALIOP cloud detection translate well to accurate uncertainties outside of CALIOP collocations.

4.3 Cloud Detection Consistency

380   Evidenced by much of the previous analysis, the detectability of a cloudy pixel by a cloud masking algorithm can depend on a number of factors including surface characteristics, solar illumination, cloud optical depth, cloud-top height, thermal contrast with the surface, and the algorithm itself. The variation of BACC, ACC, TPR, and FPR across these conditions suggest that clouds of a fixed optical depth may be more likely detected over certain surface conditions or time of day. This is potentially

problematic and conducive to spatial and temporal artifacts in cloud amount analyses. Consider for example, a cloud of fixed
385   low optical depth advected sequentially over a cold land surface, a relatively warm ocean surface, and sea ice. Regardless of the
overall accuracy of a cloud mask or effective definition of a cloudy scene, an algorithm with a varying TPR over these surface
types could produce spatial artifacts related to these surfaces. Considering that solar illumination may change during this time
further complicates this example and could produce unrealistic cloud amount variability through time. In many scenarios, this
is unavoidable due to the limitations of the satellite instrument. However, we argue that a desirable quality of a cloud mask is
390   consistency in TPR across varying surface types and solar illumination conditions, and that, ideally, cloud detection should be
dependent on characteristics of the cloud and not characteristics of the surface or solar illumination. We expect that additionally
optimizing for minimal TPR differences between these conditions at fixed cloud optical depths could help eliminate artificial
spatial and temporal variability in cloud amount analyses.

To investigate this concern, we calculate the TPR for clouds above an increasing optical depth threshold. Then, we find the
395   difference in TPR across daytime, nighttime, land, and water for three latitude bands (Fig. 10). An important consideration for
Fig. 10 is that a cloud mask can have low accuracy, but also low TPR differences if it makes consistent predictions with respect
to cloud optical depth across the conditions examined.

In general, as the lower optical depth threshold increases, TPR differences decrease for all cloud masks with a few exceptions.
The neural network has TPR differences less than or equal to 5% for all scenarios examined except for the difference between
400   nighttime water and nighttime land, and the difference between daytime land and nighttime land at the high latitudes. In both
instances, the differences converge to less than 5% at optical depths greater than 1. All cloud masks struggle with consistency
at high latitudes and for optically thin clouds.

The ECM shows strong consistency in TPR between daytime and nighttime water at all latitudes for both datasets. However,
it struggles in many other scenarios. In Fig. 10.d (low latitude nighttime water and nighttime land), the ECM is the only mask
405   with differences greater than 5%. In Fig. 10.f (high latitude nighttime water – nighttime land) the ECM has the largest TPR
difference observed of roughly 28% for optically thin clouds.

The MVCM has the largest TPR differences in nine out of the twelve scenarios examined in Fig. 10. In a few cases (Fig. 10.a,
10.b, 10.g) the large TPR differences converge to zero at larger optical depths. However, in other cases, the large differences
remain even for optically thick clouds. This is especially true for daytime/nighttime consistency over both land and water at
410   high latitudes (Fig. 10.i, 10.l) where differences are larger than 10% for clouds with optical depths greater than 1.0.

4.4   Regional Analysis

In order to give some context to the largest differences we have observed when validating with CALIOP collocations, we
perform a limited regional analysis comparing the neural network cloud mask and the MVCM. We focus this analysis on
Greenland because it is one of the worst performing regions for both masks. We process every SNPP-VIIRS scene in 2019
415   where the nadir VIIRS ground track comes within the bounding box of latitudes 60N to 80N and longitudes 70W to 20W. This
results in a total of 4,412 six-minute VIIRS scenes. Due to the large amount of scenes, we additionally subsample every fifth

pixel from every fifth scanline. For the neural network and the MVCM we calculate the mean cloud fraction for the region 58N to 84N, and 80W to 10W using a grid size of 0.5 degrees latitude and 1 degree longitude (Fig. 11.a, 11.b).

420 Consistent with the CALIOP validation, we observe large differences over the Greenland land mass (Fig. 11.c). The neural network predicts 10-25% higher cloud fraction over Greenland varying with exact location. Differences over the ocean to the southeast of Greenland are negative and fairly small. However, the ocean to the north and west of Greenland have large positive differences similar to those over Greenland itself. Based on the spatial characteristics of the mean MVCM cloud fraction over the ocean, we hypothesize that these differences may be a result sea ice cover. A similar result was found previously in Liu et al. (2010), where MODIS cloud detection errors related to the presence of sea ice were suggested to contribute to large errors

425 in cloud fraction trends.

Focused regional comparisons between imagers and CALIOP can be difficult due to the relative sparsity of CALIOP observations in small geographical regions. A domain-wide averaged cloud fraction comparison between the two imager cloud masks and CALIOP is subject a large amount of error due to the differences in spatial sampling and observation times. We calculate a domain-wide average of cloud fraction for CALIOP and the two cloud masks and plot the 31-day moving average

430 as a function of time (Fig. 11.d). To account for some of the differences in sampling, this average only includes grid points from the neural network and MVCM for which CALIOP has sampled on the same day. This effectively removes the impact of differences in spatial sampling, but ignores differences in temporal sampling. Thus, we should still not expect either the MVCM or the neural network to follow the CALIOP 1 km or 5 km products extremely closely. The averages across space are weighted by the cosine of latitude expressed in radians.

435 The largest differences occur in northern hemisphere winter, with better agreement between the MVCM and neural network occurring during northern hemisphere summer. This suggests that the MVCM's tendency to underestimate cloud cover during conditions with no solar illumination heavily contributes to the spatial differences observed in Fig. 11.c. Similarly, the magnitude of the seasonal cycle in the MVCM is likely exaggerated due to variation of solar zenith angle throughout the year. Both cloud masks also show very different shapes to the seasonal cycle even when ignoring the overall differences in mean cloud

440 fraction. Despite differences in temporal sampling, the neural network shows somewhat similar variability to both CALIOP products. Overall, the neural network shows mean cloud fractions more similar to the 5 km CALIOP product despite being trained with labels from the 1 km product. This is not a surprising result since the neural network is a statistical algorithm and is incentivized to predict the majority class (cloudy) in uncertain conditions when both classes are given equal weight. The 5 km CALIOP product likely has a larger mean cloud fraction due to its ability to detect clouds with low optical depths. Of the

445 two cloud masks, the neural network appears to give a more realistic assessment of cloud cover variability in this analysis and more closely aligns with that of CALIOP.

4.5 Sun Glint Example

As mentioned in Section 3.1 the VIIRS/CALIOP collocations used in this analysis do not contain scenes with a significant amount of sun glint. Figure 12 illustrates the impact of pseudo-labeling sun glint scenes with VIIRS infrared and VIIRS/CrIS

450 fusion channels. Without pseudo-labeling (Fig. 12.c), the high visible reflectivity causes the neural network model to over

predict cloud cover in these regions. Even areas far away from the specular point with only marginal sun glint are significantly impacted. This is somewhat remedied by including pseudo-labels in training the neural network (Fig. 12.d). Qualitatively, the ECM (Fig. 12.f) appears to be the least effected by sun glint and most able to correctly discriminate clear from cloudy in the sun glint region. The MVCM (Fig. 12.e) over predicts cloud cover directly over the specular point, but captures small cloud variability surrounding it. The neural network with pseudo-labels makes relatively realistic predictions compared to without pseudo-labeling. However, it does not capture small cloud variability around the specular point to the same degree as the ECM. The pseudo-labeling model likely has low skill in such conditions due to the low contrast between a low-level fractionally cloudy pixel and the background. There appears to be little disagreement between the cloud masks for the larger, more reflective, and colder cloud features.

5   Discussion

There are few common themes in much of the analysis done in section 4. The BACC calculated over global averages of a few surface types suggests that the neural network is better at discriminating cloudy from cloud-free scenes in most scenarios. Further analysis shows that a large majority of this improvement comes from collocations located at the middle and high latitudes. According to the CALIOP collocations, the ECM and neural network cloud masks appear relatively comparable over low-latitude land and ocean with the MVCM trailing slightly behind both in this region. The ECM appears slightly more capable of identifying low-level small clouds in the unfiltered dataset in low-latitude nighttime scenes over water. Despite training using an unfiltered dataset that contains fractionally cloudy pixels identified by CALIOP, the neural network cloud mask still struggles in fractionally cloudy scenes. This is likely due to a combination of noisy labels from CALIOP in these conditions and the low contrast with the underlying and surrounding surface. A consistent problem in using CALIOP as a reference is broken cloudiness. These clouds pose a significant challenge to cloud masking in general, but are particularly difficult to handle when the corresponding CALIOP profile is not fully representative of its collocated imager pixel. Future efforts to provide a high-quality, fine-resolution, globally-distributed cloud labels could prove extremely useful to solve these issues.

Our choice of training on an unfiltered collocation dataset was made to avoid any bias with regards to the spatial characteristics of cloud cover. We expect that filtering out spatially variable clouds from the training dataset would result in an even worse characterization of small clouds by the neural network. Despite training on a relatively unreliable collection of CALIOP collocations, we report much higher BACC for the vast majority of scenarios, especially in homogeneously cloudy scenes represented by the filtered testing dataset.

It should also be noted that the decision to use CALIOP as a reference and the lack of filtering applied to the training dataset affects how the neural network uncertainty estimates can be interpreted. Reported uncertainties by the neural network should not be purely attributed to the ability of the model to detect clouds based on spectral variability alone. Since we include neighboring pixels in the inputs, spatial variation in VIIRS channels is also a contributor. Additionally, these uncertainty estimates are also a function of how representative CALIOP profiles typically are of a given pixel. This suggests that uncertainties associ-

ated with regions of broken clouds are elevated due to the difficulty of obtaining mutually representative collocations between
485   CALIOP and VIIRS.

There are many areas for improvement in the neural network approach. For instance, we included all 16 moderate reso-
lution channels in our algorithm. It is plausible that some channels that are not especially useful in cloud detection, or the
useful information they provide to the task is redundant among other channels. Pruning inputs to the model could ultimately
speed up processing and could reduce the likelihood of over-fitting. Future work could investigate the benefit of including the
490   375 m I-band measurements from VIIRS. We did not include I-band measurements, since obtaining these observations more
than doubled the processing time for creating the collocation dataset, training the model, and making predictions. Sub-pixel
information from the I-band measurements could likely help identify small cloud features. However, we expect that the poor
representation of small clouds by the CALIOP/VIIRS collocations would severely limit the usefulness of their incorporation.
Further work is needed to in order to properly assess how I-band measurements could be used to maximize their value in cloud
495   property algorithms trained with CALIOP.

Despite the large increase in BACC made by our neural network approach, there is still room for improvement particularly
during the night. One potential solution might be the incorporation of VIIRS/CrIS fusion channels into the inputs of the final
neural network model. Similar to the usage of I-band measurements, this may increase the prediction time. However, the
spectral regions covered by the I-bands are already well-represented in the moderate resolution channels. The VIIRS/CrIS
500   fusion channels represent spectral regions not covered in the native VIIRS channels such as those with significant CO_2 (
MODIS bands 33-36) , H_2O (MODIS bands 27 and 28), and O_3 (MODIS band 30) absorption. Thus, the increase in cloud
detection accuracy may be worth the trade-off of increased prediction time associated with their inclusion. However, an added
difficulty is that the fusion channel estimates are made from relatively coarse resolution CrIS channels. This could negatively
impact cloud detection for fractionally cloudy pixels to an even greater degree.

505   Our approach currently includes very little ancillary data: only a VIIRS-derived binary land/water mask. The MVCM uses
several, including surface temperatures, sea ice, snow cover, and Normalized Difference Vegetation Index maps. The ECM
also includes surface temperatures, sea ice, snow cover, tropopause temperatures, and clear-sky estimates of many channels
using radiative transfer models. Anecdotally, we notice that some spatial artifacts we have observed in the two operational
cloud masks appear to be related to the relatively coarse resolution of the ancillary datasets. Early experiments with the neural
510   network lead us to believe that including surface temperature increased the frequency of spatial artifacts in its output. This
motivated our decision to initially not include information such as surface temperatures in our approach even though it lead to
substantial increases in cloud detection performance. We leave it to future work to investigate how to include coarse-resolution
ancillary data in the neural network without increasing the prevalence of spatial artifacts in cloud masking output.

For all scenarios examined in Fig. 10 we conclude that the neural network is the most consistent in identifying clouds across
515   various geographical, solar illumination, and surface conditions while controlling for cloud optical depth. There are several
reasons why the neural network model might be successful in this regard. The ECM and MVCM both apply different tests
based on surface condition and solar zenith angle. The ECM, for example, is a collection of naïve Bayesian models trained
for different surface types. This a very intuitive approach, but in practice requires partitioning collocation datasets according

to surface type and reduces the number of collocations that can be used for training each model. Similarly the MVCM uses

520   different decision pathways and restricts or requires usage of certain inputs accordingly. We hypothesize that training a only single model (rather than multiple), and instead providing the land/water mask and solar-zenith angle as inputs has contributed to its consistency in cloud detection under these varying conditions.

In one of the worst performing regions for all three cloud masks, we observe very substantial differences in mean cloud fraction for 2019 across both space and time. These results demonstrate how differences in TPR of a cloud mask over varying

525   surface and illumination conditions could potentially contribute to very different spatial and temporal variability. Because of this, we argue that minimizing TPR differences over varying surface and illumination conditions could be a useful objective in cloud mask development in addition to maximizing accuracy or other related performance metrics.

We note several potential caveats in the assessment of the neural network cloud mask in addition to issues with fractional cloudiness. One clear limitation with using CALIOP as a source for labels is the relatively narrow range of sensor viewing

530   angle and solar illumination combinations. We examined one specific example of this in sun glint and have limited, but not completely removed, its adverse impact on cloud detection using pseudo-labeling. One disadvantage of the pseudo-labeling approach, is that the associated uncertainty estimates lose much of their meaning in domains where we exclusively train on pseudo-labels. We have attempted to limit the impact of this issue by training the neural network to estimate the class probabilities produced by the pseudo-labeling model, and not the predicted class labels themselves. This approach appears to

535   be successful in preventing severe overclouding of sun-glint regions, but reasonably, it can only be expected to perform as well as a model that uses infrared observations exclusively. Additionally, we have not evaluated how the neural network performs specifically in cloud-free scenes with high aerosol loading. We expect that this could depend largely on the ability for CALIOP to distinguish cloud from aerosol layers.

6   Conclusions

540   In this work, we examine the performance of a neural network cloud mask for VIIRS that is trained with coincident CALIOP observations and compared it with two operational cloud masks. Both the MVCM and ECM appear to be slightly better at identifying small broken clouds than the neural network. However, the neural network outperforms both operational masks in most other conditions. We observe particularly large improvement at the middle and high latitudes during the night where the operational masks missed substantial fractions of optically-thick clouds that were correctly identified by the neural network.

545   We have ruled out the possibility that the improvement is due to disagreements in each approach's effective definition of a cloud. Furthermore, we find that uncertainty estimates from the neural network are well calibrated and appropriately represent the ability to estimate cloudy or cloud-free labels from CALIOP. When examining differences in true positive rate, we find that the neural network is the most consistent in identifying clouds of a fixed optical depth when considering day/night and land/water conditions. A regional analysis over Greenland for 2019 confirms that such differences could contribute to vastly

550   different assessments of the spatial and temporal variability of cloud cover over certain regions. Some issues with the global

representativeness of VIIRS/CALIOP collocations are successfully mitigated with a simple semi-supervised learning approach, but more work is needed in improving detection of fractionally cloudy pixels by the neural network.

Author contributions. C.W. developed the neural network, completed the analysis, and wrote the manuscript. A.H. and S.A. supervised the project, and contributed to the review and editing of the paper. All authors contributed to the initial conceptualization and planning of this

555 work.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to acknowledge Denis Botambekov's help in running the Enterprise Cloud Mask. Additionally we wish to thank the UW Madison SSEC/CIMSS Atmosphere SIPS team sponsored under NASA contracts NNG15HZ38C and 80GSFC20C097 for providing software tools for easily obtaining VIIRS data. This work was supported by the JPSS program under NOAA

560 CA NA15NES4320001. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision.

References

Ackerman, S., Richard, F., Kathleen, S., Yinghui, L., Liam, G., Bryan, B., and Paul, M.: Discriminating Clear-Sky from Cloud with MODIS, Algorithm Theoretical Basis Document (MOD35) - Version 6.1, Tech. rep., NASA, https://atmosphere-imager.gsfc.nasa.gov/sites/default/
565   files/ModAtmo/MOD35_ATBD_Collection6_1.pdf, 2010.

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., and McGill, M.: Cloud Detection with MODIS. Part II: Validation, J. Atmos. Ocean. Technol., 25, 1073–1086, https://doi.org/10.1175/2007JTECHA1053.1, http://journals.ametsoc.org/jtech/article-pdf/25/7/1073/3333780/2007jtecha1053_1.pdf, 2008.

Bulgin, C. E., Mittaz, J. P., Embury, O., Eastwood, S., and Merchant, C. J.: Bayesian Cloud Detection for 37 Years of Advanced Very
570   High Resolution Radiometer (AVHRR) Global Area Coverage (GAC) Data, Remote Sens., 10, 97, https://doi.org/10.3390/rs10010097, http://www.mdpi.com/2072-4292/10/1/97, 2018.

Cao, C., De Luccia, F. J., Xiong, X., Wolfe, R., and Weng, F.: Early On-Orbit Performance of the Visible Infrared Imaging Radiometer Suite Onboard the Suomi National Polar-Orbiting Partnership (S-NPP) Satellite, IEEE Trans. Geosci. Remote Sens., 52, 1142–1156, https://doi.org/10.1109/TGRS.2013.2247768, 2013.

575 Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., and Lindsey, D. T.: An Empirical Model for Assessing the Severe Weather Potential of Developing Convection, Weather Forecast., 29, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1, http://journals.ametsoc.org/waf/article-pdf/29/3/639/4655021/waf-d-13-00113_1.pdf, 2014.

Frey, R., Ackerman, S., Holz, R., and Dutcher, S.: The Continuity MODIS-VIIRS Cloud Mask (MVCM) User's Guide Based on NASA MODIS Cloud Mask (MOD35, MYD35) Reprocessed Data-Version 1 Product User's Guide Version 1.0, https://ladsweb.modaps.eosdis.
580   nasa.gov/missions-and-measurements/viirs/MODIS_VIIRS_Cloud-Mask_UG_04162020.pdf, 2019.

Håkansson, N., Adok, C., Thoss, A., Scheirer, R., and Hörnquist, S.: Neural network cloud top pressure and height for MODIS, Atmos. Meas. Tech., 11, 3177–3196, https://doi.org/10.5194/amt-11-3177-2018, https://amt.copernicus.org/articles/11/3177/2018/, 2018.

Heidinger, A., Botambekov, D., and Walther, A.: A Naïve Bayesian Cloud Mask delivered to NOAA Enterprise - Version 1.2, Tech. rep., NOAA NESDIS Center for Satellite Applications and Research, https://www.star.nesdis.noaa.gov/goesr/documents/ATBDs/Enterprise/
585   ATBD_Enterprise_Cloud_Mask_v1.2_Oct2016.pdf, 2016.

Heidinger, A. K., Evan, A. T., Foster, M. J., and Walther, A.: 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, 2012.

Holz, R. E., Ackerman, S. A., Nagle, F. W., Frey, R., Dutcher, S., Kuehn, R. E., Vaughan, M. A., and Baum, B.: Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP, J. Geophys. Res.
590   Atmos., 114, https://doi.org/10.1029/2008JD009837, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008JD009837https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JD009837https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2008JD009837, 2008.

Karlsson, K. G., Johansson, E., Håkansson, N., Sedlar, J., and Eliasson, S.: Probabilistic Cloud Masking for the Generation of CM SAF Cloud Climate Data Records from AVHRR and SEVIRI Sensors, Remote Sens., 12, 713, https://doi.org/10.3390/rs12040713, https://www.mdpi.
595   com/2072-4292/12/4/713, 2020.

Kingma, D. P. and Ba, J. L.: Adam: A Method for Stochastic Optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., International Conference on Learning Representations, ICLR, https://arxiv.org/abs/1412.6980v9, 2015.

Kox, S., Bugliaro, L., and Ostler, A.: Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing, Atmos. Meas. Tech., 7, 3233–3246, https://doi.org/10.5194/amt-7-3233-2014, https://amt.copernicus.org/articles/7/3233/2014/, 2014.

600 Kühnlein, M., Appelhans, T., Thies, B., and Nauß, T.: Precipitation Estimates from MSG SEVIRI Daytime, Nighttime, and Twilight Data with Random Forests, J. Appl. Meteorol. Climatol., 53, 2457–2480, https://doi.org/10.1175/JAMC-D-14-0082.1, http://journals.ametsoc.org/jamc/article-pdf/53/11/2457/3577888/jamc-d-14-0082_1.pdf, 2014.

Lee, D.-H.: Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, ICML 2013 Work. Challenges Represent. Learn., pp. 1–6, 2013.

605 Liu, Y., Ackerman, S. A., Maddux, B. C., Key, J. R., and Frey, R. A.: Errors in Cloud Detection over the Arctic Using a Satellite Imager and Implications for Observing Feedback Mechanisms, J. Clim., 23, 1894–1907, https://doi.org/10.1175/2009JCLI3386.1, http://journals.ametsoc.org/jcli/article-pdf/23/7/1894/3971933/2009jcli3386_1.pdf, 2010.

Marais, W., Holz, R., Reid, J., and Willett, R.: Leveraging spatial textures, through machine learning, to identify aerosol and distinct cloud types from multispectral observations, Atmos. Meas. Tech., 13, 1–35, https://doi.org/10.5194/amt-2020-74, https://amt.copernicus.org/articles/13/5459/2020/, 2020.

610 McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika, 12, 153–157, https://doi.org/10.1007/BF02295996, https://link.springer.com/article/10.1007/BF02295996, 1947.

Minnis, P., Hong, G., Sun-Mack, S., Smith, W. L., Chen, Y., and Miller, S. D.: Estimating nocturnal opaque ice cloud optical depth from MODIS multispectral infrared radiances using a neural network method, J. Geophys. Res., 121, 4907–4932, https://doi.org/10.1002/2015JD024456, https://onlinelibrary.wiley.com/doi/abs/10.1002/2015JD024456, 2016.

615 Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., and Riedi, J.: 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, /pmc/articles/PMC5896565/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5896565/, 2017.

620 Shorten, C. and Khoshgoftaar, T. M.: A survey on Image Data Augmentation for Deep Learning, J. Big Data, 6, 1–48, https://doi.org/10.1186/s40537-019-0197-0, https://link.springer.com/articles/10.1186/s40537-019-0197-0, 2019.

Smith, L. N.: Cyclical Learning Rates for Training Neural Networks, in: Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017, pp. 464–472, Institute of Electrical and Electronics Engineers Inc., https://doi.org/10.1109/WACV.2017.58, http://arxiv.org/abs/1506.01186, 2017.

625 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15, 1929–1958, http://jmlr.org/papers/v15/srivastava14a.html, 2014.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P.: AR5 Climate Change 2013: The Physical Science Basis — IPCC, https://www.ipcc.ch/report/ar5/wg1/, 2013.

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Hei-
630 dinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G.: Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel, Bull. Am. Meteorol. Soc., 94, 1031–1049, https://doi.org/10.1175/BAMS-D-12-00117.1, http://journals.ametsoc.org/bams/article-pdf/94/7/1031/3740632/bams-d-12-00117_1.pdf, 2013.

Thampi, B. V., Wong, T., Lukashin, C., and Loeb, N. G.: Determination of CERES TOA Fluxes Using Machine Learning Algorithms. Part I: Classification and Retrieval of CERES Cloudy and Clear Scenes, J. Atmos. Ocean. Technol., 34, 2329–2345, https://doi.org/https://doi.org/10.1175/JTECH-D-16-0183.1, 2017.

Uddstrom, M. J., Gray, W. R., Murphy, R., Oien, N. A., and Murray, T.: A Bayesian Cloud Mask for Sea Surface Temperature Retrieval, J. Atmos. Ocean. Technol., 16, 117–132, https://doi.org/10.1175/1520-0426(1999)016<0117:ABCMFS>2.0.CO;2, http://journals.ametsoc.org/jtech/article-pdf/16/1/117/3302781/1520-0426, 1999.

Vaughan, M. A., Powell, K. A., Kuehn, R. E., Young, S. A., Winker, D. M., Hostetler, C. A., Hunt, W. H., Liu, Z., Mcgill, M. J., and Getzewich, B. J.: Fully automated detection of cloud and aerosol layers in the CALIPSO lidar measurements, J. Atmos. Ocean. Technol., 26, 2034–2050, https://doi.org/10.1175/2009JTECHA1228.1, http://journals.ametsoc.org/jtech/article-pdf/26/10/2034/3606994/2009jtecha1228_1.pdf, 2009.

Wang, C., Platnick, S., Meyer, K., Zhang, Z., and Zhou, Y.: 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, https://amt.copernicus.org/articles/13/2257/2020/, 2020.

Weisz, E., Baum, B. A., and Menzel, W. P.: Fusion of satellite-based imager and sounder data to construct supplementary high spatial resolution narrowband IR radiances, J. Appl. Remote Sens., 11, 1, https://doi.org/10.1117/1.jrs.11.036022, 2017.

Welch, R. M., Sengupta, S. K., Goroch, A. K., Rabindra, P., Rangaraj, N., and Navar, M. S.: Polar Cloud and Surface Classification Using AVHRR Imagery: An Intercomparison of Methods, J. Appl. Meteorol., 31, 405–420, https://doi.org/10.1175/1520-0450(1992)031<0405:PCASCU>2.0.CO;2, 1992.

Wimmers, A., Velden, C., and Cossuth, J. H.: Using Deep Learning to Estimate Tropical Cyclone Intensity from Satellite Passive Microwave Imagery, Mon. Weather Rev., 147, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1, http://journals.ametsoc.org/mwr/article-pdf/147/6/2261/4881287/mwr-d-18-0391_1.pdf, 2019.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., Hunt, W. H., and Young, S. A.: Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms, J. Atmos. Ocean. Technol., 26, 2310–2323, https://doi.org/10.1175/2009JTECHA1281.1, http://journals.ametsoc.org/jtech/article-pdf/26/11/2310/3335867/2009jtecha1281_1.pdf, 2009.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D.: MixUp: Beyond empirical risk minimization, in: 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., https://arxiv.org/abs/1710.09412, 2018.

	Neural Network BACC		ECM BACC		MVCM BACC		Collocation Count (Million)	
	Day	Night	Day	Night	Day	Night	Day	Night
All	0.968	0.934	0.938	0.849	0.910	0.876	2.96	2.91
Water	0.969	0.932	0.940	0.842	0.909	0.893	1.99	1.99
Land	0.965	0.916	0.917	0.808	0.868	0.808	0.97	0.91
Sea Ice	0.966	0.895	0.883	0.661	0.879	0.790	0.29	0.31
Permanent Snow	0.961	0.863	0.885	0.701	0.822	0.694	0.30	0.36
Snow	0.954	0.920	0.855	0.758	0.864	0.778	0.16	0.19

Table 1. Balanced Accuracy (BACC) calculated for each cloud mask over different surfaces during day and night for the filtered dataset. Collocation counts do not sum to the count listed in the "All" row. Sea Ice collocations are also counted in the water category, and the two snow categories are also counted in the land category
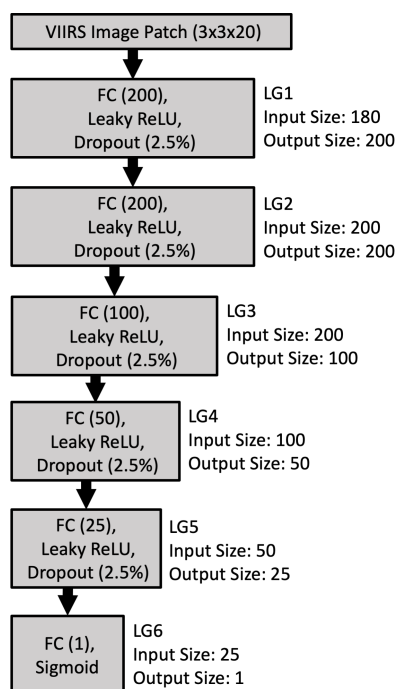
Figure 1. Schematic of the neural network model used. LG refers to Layer Group and is used to describe the collection of layers in each box. FC(x) refers to the fully connected layers where x is the number of units in each layer. Similarly, Dropout(x) refers to the fraction of inputs to which dropout is applied.
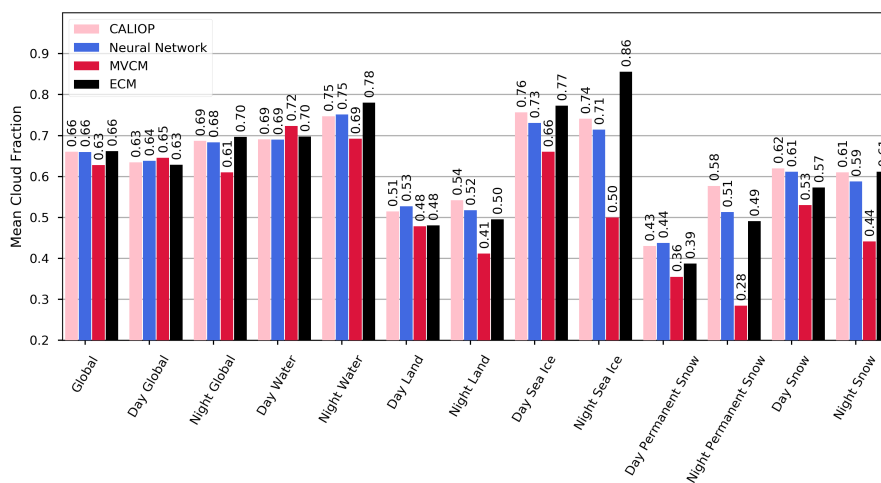
Figure 2. Mean cloud fraction for the 2019 unfiltered testing dataset. Each bar grouping from left to right shows the value from the CALIOP 1 km product, the neural network, MVCM, and ECM. Time of day and surface categorizations are described in the main text.
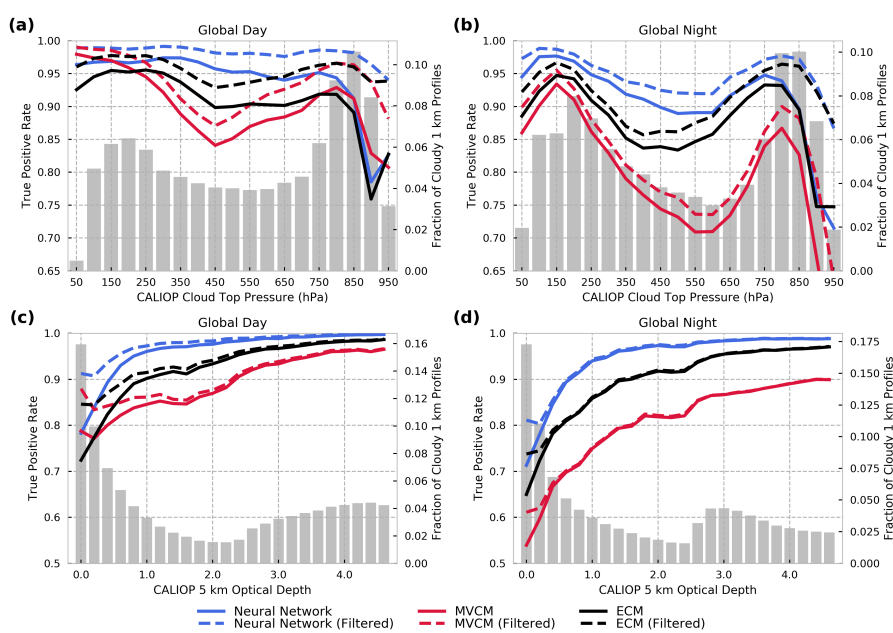
Figure 3. True positive rate (TPR) calculated as function of cloud-top pressure (a,b) and optical depth (c,d) for daytime and nighttime collocations respectively. The grey bars represent the fraction of cloudy 1 km CALIOP profiles. Only profiles with non-zero optical depths are included in (c) and (d).
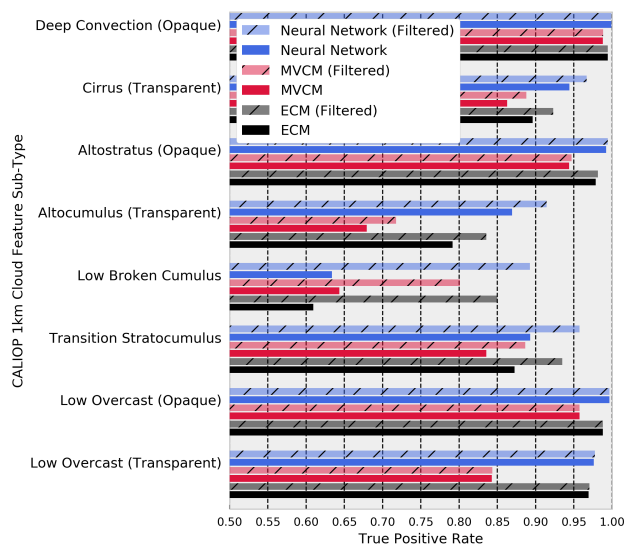
Figure 4. The True Positive Rate (TPR) for various CALIOP cloud-feature types from the 1km CALIOP Cloud Layers product. The order shown in the legend indicates the ordering of the bars in each grouping.
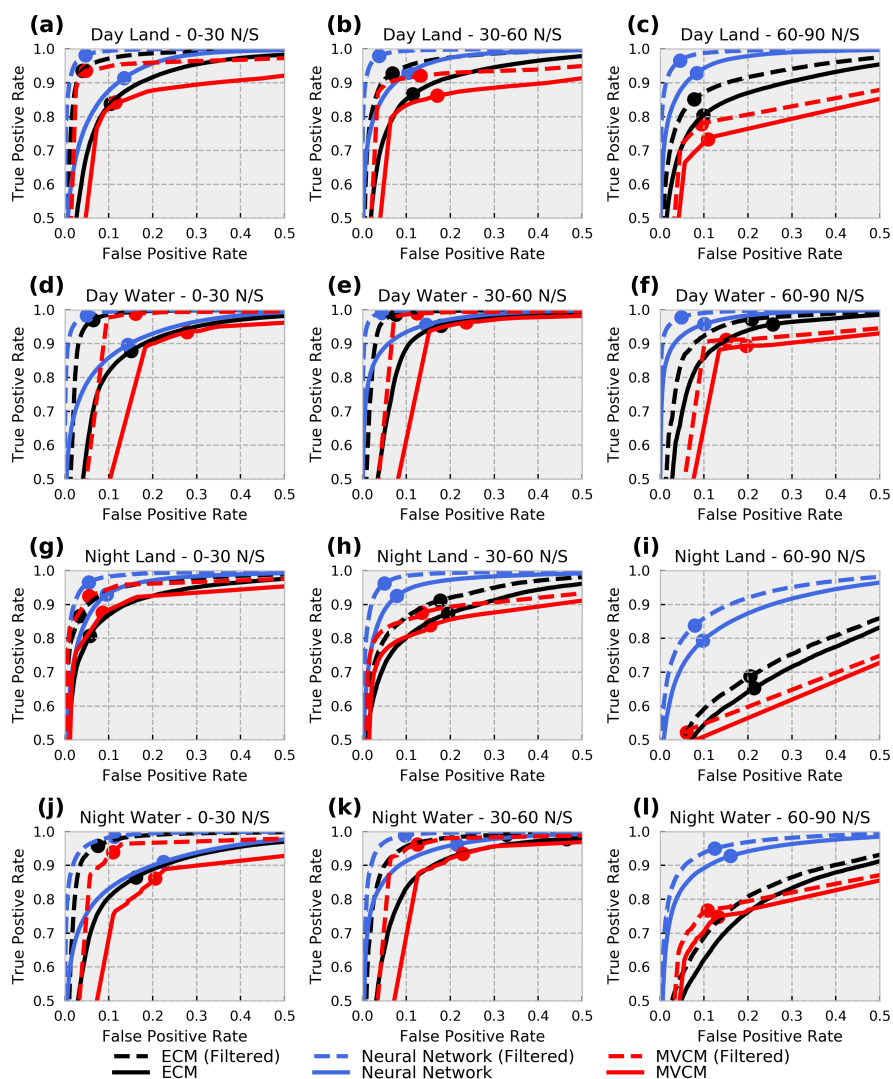
Figure 5. Receiver operating characteristic (ROC) curves for all three cloud masking models. The text above each subplot indicates the subset of collocations for which the curves are plotted. Note that the x and y axis limits are somewhat atypical for ROC curve plots and are chosen here to emphasize the differences between the masks and different datasets. The TPR and FPR for the model using the standard threshold of 0.5 for the neural network and ECM, as well as the integer cloud mask for MVCM are also shown with similarly colored circles.
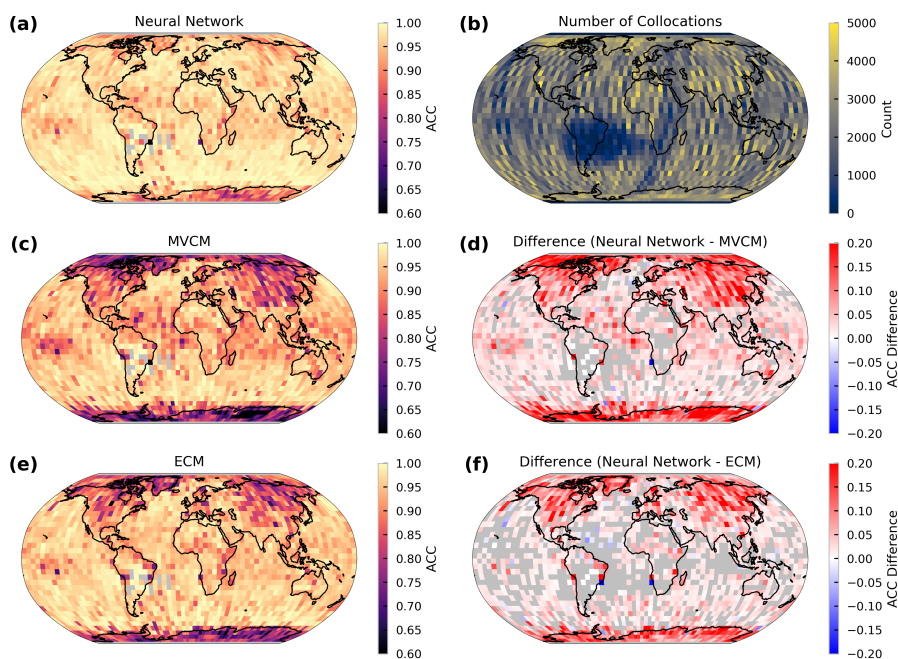
Figure 6. Geographic comparison of the ACC between the three cloud masks on the filtered testing dataset. Each grid cell is 5 degrees latitude by 5 degrees longitude. The gap in coverage over South America is due to the removal of low-energy laser shots from the CALIOP datasets. Cells with less than 100 collocations are not shown in (a) or (c)-(f). Differences are only shown where determined significant by McNemar's test with p-values less than 0.001.
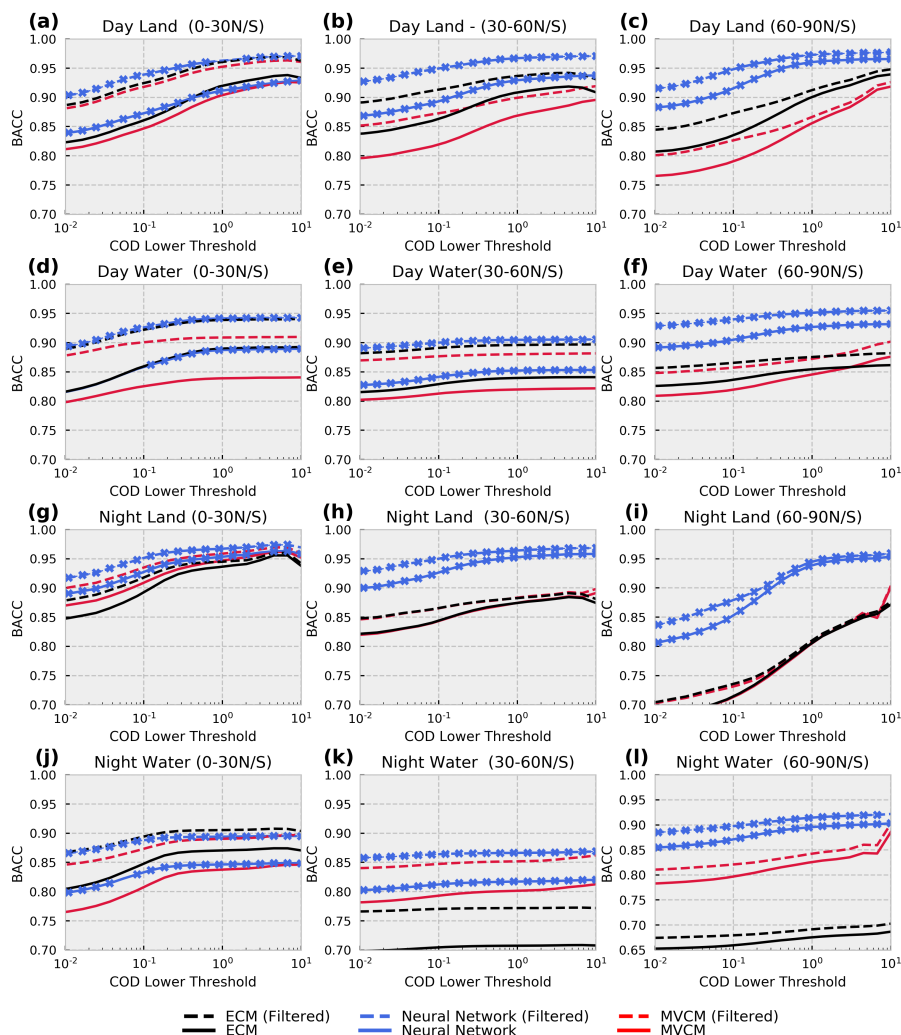
Figure 7. Balanced Accuracy (BACC) recalculated after removing clouds below a certain cloud optical depth (COD) threshold. Tick marks on the neural network lines indicate significant differences in performance between the neural network and the best operational model using McNemar's test with p-values less than 0.001. Note that the y-axis limits are different for (l) compared to the other subplots.
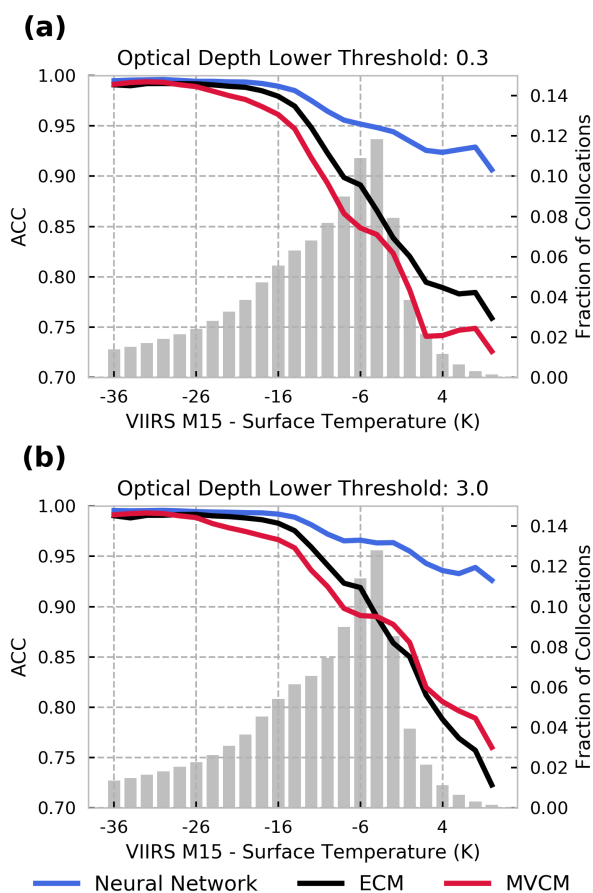
Figure 8. ACC calculated as a function of thermal contrast with the surface approximated by the difference between VIIRS M15 (10.8 μm) and surface temperature in Kelvin. Each subplot represents a set of collocations consisting of clear-sky scenes and cloudy scenes with optical depths greater than 0.3 (a) and 3.0 (b).
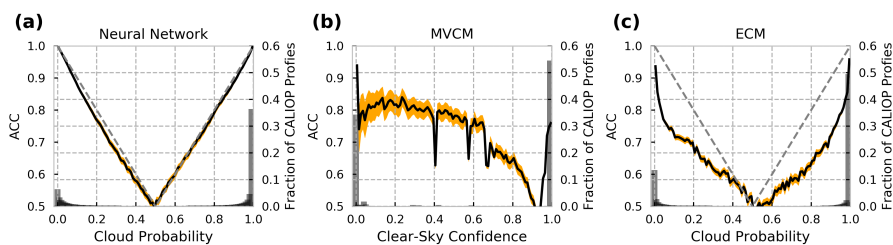
Figure 9. Uncertainty assessments for (a) the neural network cloud mask model (b) the MVCM, and (c) the ECM. ACC values (left y-axis) for cloud probability and clear sky confidence values are calculated for bins of size 0.01. For (a) and (c) a perfectly-calibrated model is plotted with the grey dashed line (see main text). Orange shading indicates the 99.9% confidence interval. Grey bars indicate the fraction of collocations falling within each bin of width 0.01.
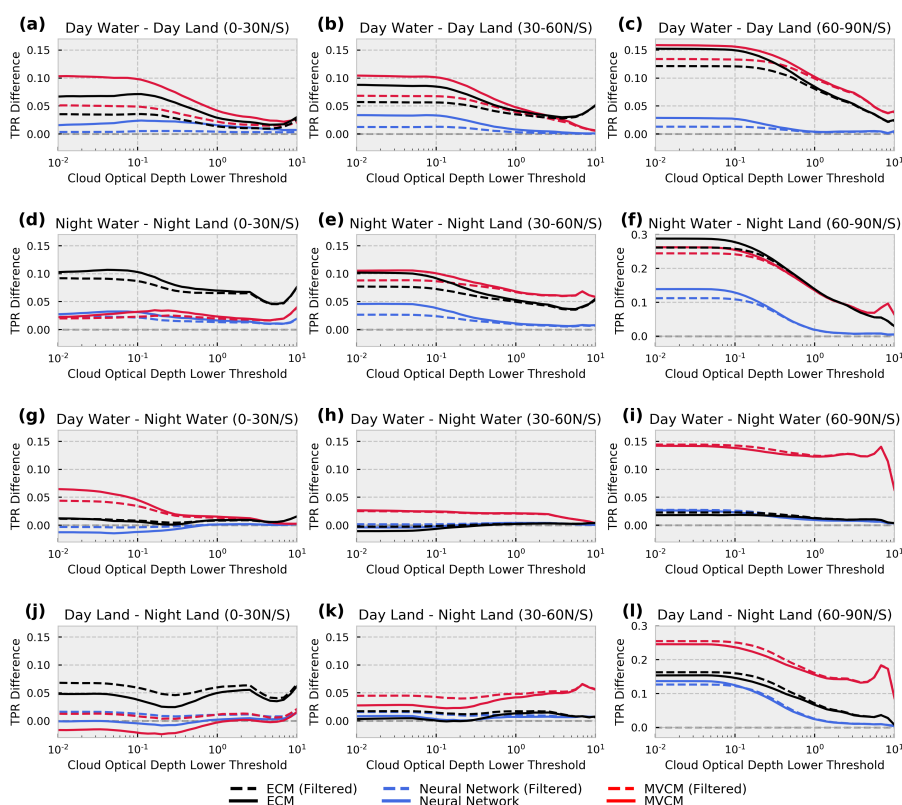
Figure 10. TPR differences over combinations of land/water and day/night conditions. The specific TPR difference and latitude is labeled at the top of each subplot. Note that the y-axis limits are different for (f) and (l).
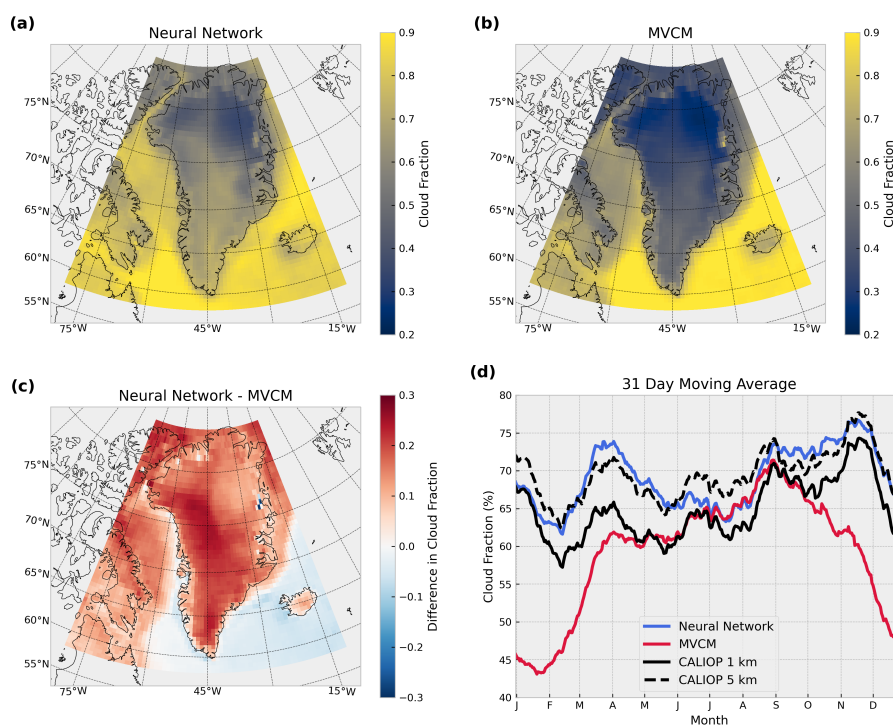
Figure 11. Regional analysis of cloud fraction over Greenland. (a) and (b) illustrate the mean cloud fraction for the neural network and the MVCM for all selected VIIRS scenes in 2019. (c) is the difference between (a) and (b). (d) is the domain-wide 31-day moving average of grid points spatially matched with CALIOP (see main text for details).
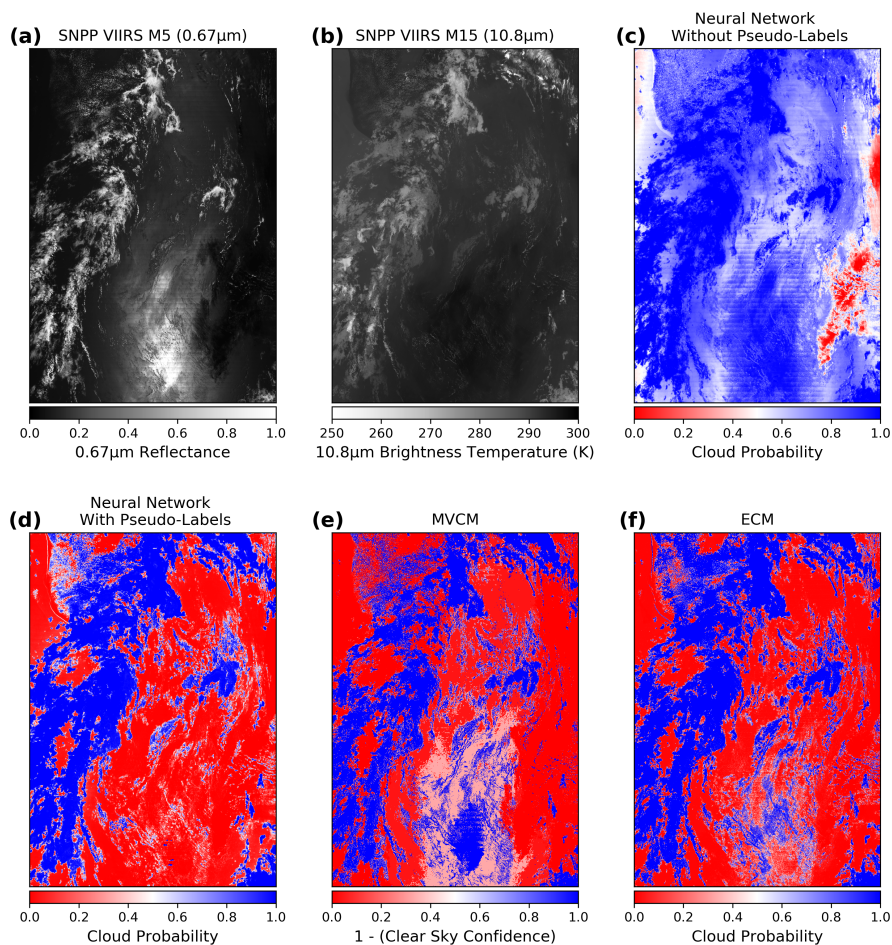
Figure 12. Comparison of the neural network cloud mask without pseudo-labels (c), with pseudo-labels (d), the MVCM (e), and the ECM (f). Also shown are band M5 with a central wavelength of roughly 0.67μm (a) and band M15 with a central wavelength of roughly 10.8μm (b).