APPLICATION OF MACHINE LEARNING METHODS TO IMAGER CLOUD PROPERTY ESTIMATION AND THE FEASIBILITY OF THEIR USE IN OPERATIONS AND CLIMATE DATA RECORDS

by

Charles H. White

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Atmospheric and Oceanic Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 2/7/2022

The dissertation is approved by the following members of the Final Oral Committee:

Andrew K. Heidinger, Senior GEO Scientist, NOAA/NESDIS
Steven A. Ackerman, Professor, Atmospheric and Oceanic Sciences
Grant W. Petty, Professor, Atmospheric and Oceanic Sciences
Daniel J. Vimont, Professor, Environment and Resources
Tristan S. L'Ecuyer, Professor, Atmospheric and Oceanic Sciences

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Estimates of cloud properties are critical to our understanding of weather and climate variability, but their estimation from satellite imagers is a nontrivial task. Machine learning (ML) approaches have recently gained popularity in earth science and remote sensing. This work explores the use of a kind of ML model, a neural network, for cloud detection and cloud-top pressure estimation from the Visible Infrared Imaging Radiometer Suite (VIIRS), Advanced Baseline Imager (ABI), and Moderate Resolution Imaging Spectroradiometer (MODIS). Several comparisons illustrate large improvement over current operational products which rely on more conventional statistical or physically-based approaches.

This increase in performance merits study into the interpretability of neural network cloud property models. A comparison of several modern interpretability frameworks for neural networks shows mixed results and implies that current tools may be insufficient for explaining neural network output in remote sensing tasks with multicollinear predictors. Nonetheless, we find some agreement on the importance of particular spectral features, spatial metrics, and numerical weather prediction output that could inform future algorithm development.

A key challenge in transitioning algorithms to satellite climate records is ensuring intersensor consistency. If this is not considered, then long-term analyses of clouds risk being affected by changes in observation platform which can be frequent in our longest satellite records. A method is proposed that simultaneously minimizes differences between imager predictions for matching observations and predictions with respect to a reference instrument. These results offer one pathway for ensuring the appropriateness of ML algorithms in the analysis of satellite climate records.

0 PREFACE

Machine Learning (ML) methods have rapidly gained popularity in the Earth science community including applications in weather prediction, climate data analysis, remote sensing, and other areas of research. This is due to many factors including the development of efficient techniques to train neural networks (Rumelhart et al., 1986), the development of convolutional neural networks (LeCun et al., 1995) the wide availability of large datasets (Deng et al., 2009 for example), and the increasing success of neural networks in processing image data (Krizhevsky et al., 2012). Given the relatively quick uptake of these approaches into key applications in Earth science (Düben et al., 2021), it is worth studying the subtleties of evaluation, interpretability, and generalization that accompany their implementation. This dissertation specifically focuses on the application of neural networks for the cloud detection and cloud-top pressure from spaceborne satellite imagers. While this work is almost exclusively framed around the specific task of imager cloud property estimation, the methods used can serve as a examples for developing ML-based models in the broader remote sensing and Earth science community.

In particular, I aim to explore three research questions:

1. Can machine learning improve cloud property estimation relative to conventional operational algorithms?

2. How do we address interpretability concerns from machine learning methods and to what extent can modern ML interpretability frameworks help us understand how these models outperform conventional approaches?

3. How do we improve the generalization capacity of these models and allow for their use across multiple heterogeneous sensors?

While these questions are posed as separate efforts in each chapter, their solutions are intimately connected and facilitate the transition of ML-based approaches to operations and cloud climate records. Question 1 motivates the investigation into interpretability and generalization since we show very substantial improvement for key conditions in cloud detection and cloud-top pressure estimation. Without first confirming that these approaches outperform modern conventional approaches, it is difficult to argue for their use due to concerns relating to interpretability, uncertainty estimation, and generalization.

Question 2 aims to understand how ML-based approaches improve upon modern conventional algorithms and informs future algorithm development and sensor design. This is a difficult task since certain ML approaches are often regarded as black boxes and it often proves difficult to understand how certain inputs are used and what statistical relationships are exploited in particular ML models. I attempt to characterize a sample of current tools that exist to understand neural network predictions. Addressing question 2 can give us insight into how to improve both conventional and ML approaches, understand the limitations of the ML models, and help us understand the physical relationships useful for identifying particular cloud properties.

A key use of cloud property estimates is the detection of long-term variability of clouds in satellite records of observations from multiple imagers. Question 3 specifically facilitates the use of the these approaches in cloud climate records. Ensuring consistency in ML approaches among multiple imagers directly contributes to our ability to assess changes in clouds over our current observational records. These efforts are key to ensuring that neural

networks developed in this work are useful in scientific applications relevant to climate.

In addressing these questions I illustrate that neural network-based approaches for cloud detection and cloud-top pressure estimation can significantly outperform modern operational methods for the Visible Infrared Imaging Radiometer Suite VIIRS) the Advanced Baseline Imager (ABI) and the Moderate Resolution Imaging Spectroradiometer (MODIS). In exploring interpretability frameworks for neural network models, I find that current methods for explaining predictions (even from relatively simple models) are unsatisfactory for Earth science remote sensing applications with multiple correlated features. However, there are some key takeaways about the importance of particular spectral channels, spatial metrics and NWP information that can be useful for future cloud-top pressure algorithm development. Finally, I develop a method to substantially improve the intersensor consistency of cloud-top pressure models developed for VIIRS and MODIS. This method is able to match or improve upon the intersensor consistency of an operational product, but with a roughly 40% reduction in error relative to a high quality reference instrument. This approach could be one way of facilitating the transition of ML algorithms into satellite climate records made up of multiple heterogeneous imagers. Furthermore we show how this approach can succeed even in scenarios where one imager lacks labeled data from a reference instrument.

Each chapter of this dissertation addresses each research question separately and represents an article that has already been published (Chapter 1; White et al., 2021), currently in peer-review (Chapter 2), or in preparation (Chapter 3) at the time of writing. While the introductions of each contain similar information, they are written specifically to motivate the objectives of each work. Thus, they are kept here in their entirety with only light editing. As a result, each chapter of this dissertation can be read and considered independently and

in any order.

1 EVALUATION OF VIIRS NEURAL NETWORK CLOUD DETECTION AGAINST CURRENT OPERATIONAL CLOUD MASKS

1.1 Introduction

Clouds serve many critical roles in the earth's weather and climate system, and are one of the largest sources of uncertainty 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 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 of training data 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; Håkansson et al., 2018; Sus et al., 2018; Wimmers et al., 2019; Marais et al., 2020).

In this analysis, we develop a neural network cloud mask (NNCM) 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 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 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 this work we demonstrate how a very simple semi-supervised learning approach can ameliorate this specific limitation.

There are several recent applications of ML in characterizing clouds from imager observations that use CALIOP 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. Håkansson 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 prediction (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. The Community Cloud retrieval for CLimate (CC4CL; Sus et al., 2018) also uses neural network based approaches for imager cloud detection. The CC4CL neural network models are trained with collocations between the Advanced Very-High Resolution Radiometer (AVHRR) and CALIOP. Adjustments are applied to shared MODIS and Advanced Along-Track Scanning Radiometer (AATSR) channels (accounting for differences in spectral response functions) to ensure the approaches generalize beyond AVHRR to those imagers as well. While the majority of these applications for cloud property estimates are relatively recent, there were successful implementations of ML approaches well before the launch of CALIOP using manually labeled scenes (Welch et al., 1992).

Our approach aims to improve upon existing literature in several ways. Significant

effort has gone into determining useful spectral characteristics in the development of past imager cloud masks. Still, it is possible that not all relevant variability is being exploited particularly that which involves three or more channels. 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 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 temperature, 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. Our approach uses a single model for all surface types and solar illumination conditions and in some respects, greatly simplifies the processing pipeline for imager cloud masking.

In this analysis, we demonstrate that a neural network cloud mask (NNCM) can outperform two operational VIIRS clouds masks 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 performance of each approach after removing 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.

1.2 Instruments and Data

VIIRS

VIIRS is a polar-orbiting visible, near-infrared, and infrared imager on board the S-NPP and NOAA-20 satellites. The swath width of VIIRS is roughly 3060 km allowing for at least twice daily views of any given ground location and more frequent views at higher latitudes. VIIRS altogether measures top-of-atmosphere radiation for 22 different channels. This is made up of five imaging channels (I-bands) with a nadir resolution of 375 m, and sixteen moderate resolution channels (M-bands) with a nadir resolution of 750 m (Table 1.1). VIIRS has an additional Day/Night Band (DNB) for nocturnal low-light applications. This work is focused entirely on the sixteen moderate resolution channels and does not include the use of the higher resolution I-bands or the DNB. Furthermore, we only consider VIIRS data from S-NPP which has an equatorial crossing time of 1:30 pm.

CALIOP

CALIOP is polar-orbiting lidar taking near-nadir observations on board the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite which also has an equatorial crossing time of roughly 1:30 pm. CALIOP measures at wavelengths of 1064 nm and 532 nm with a horizontal resolution of 333 m. The individual lidar footprints are aggregated in the creation of both the 1 km and 5 km CALIOP Cloud Layers products. CALIOP's ability to characterize optically thin cloud layers make it a suitable validation

source for imager cloud masking. While CALIOP, in many respects, is the more appropriate instrument for accurately estimating cloud properties (including cloud detection), its spatial sampling is extremely sparse relative to VIIRS and other imagers. This motivates our goal of extending CALIOP's cloud detection ability to passive imager measurements.

MVCM and ECM

Current operational cloud masks for VIIRS include the NOAA Enterprise Cloud Mask (ECM; Heidinger et al., 2012), and the Continuity MODIS-VIIRS Cloud Mask (MVCM; Frey et al. 2020). 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. This approach is based on several naive bayesian classifiers that are each trained specifically for different surface types. This approach is similarly trained using CALIOP collocations with VIIRS and makes probabilistic predictions of cloudy or cloud-free pixels. A key advantage of the ECM's naive bayesian approach is that certain predictors can be removed or turned off (such as visible channels during the night). Due to the simplicity of naive bayesian classifiers, the ECM is overall more interpretable than our proposed neural network.

The MVCM has heritage with the MODIS cloud mask (Ackerman et al., 2010), and has 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. The MVCM has a collection of cloud tests each with specified low-confidence and high-confidence thresholds used in a fuzzy-logic approach. The specific tests that are applied are determined by solar illumination and the surface type. The clear-sky confidence values

imparted by each applied test are combined to produce a preliminary overall clear-sky confidence value which can then be modified by clear-sky restoral tests. The MVCM's reliance on physically-based reasoning also make its predictions relatively interpretable compared to our neural network approach.

Collocation Methodology

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 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. (2012) 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 27.1 million collocations were collected for this study with the above requirements.

Neural Network Inputs

The observations used as input into the neural network come from the moderate resolution channels (M1-M16; Table 1.1) obtained from the NASA processing of SNPP VIIRS. All channels are either expressed as a reflectance or brightness temperature. In addition to the VIIRS channels we also include a binary land/water mask, solar zenith angle, sun glint zenith angle, and the absolute value of latitude. The binary land/water mask is created from an eight-category land/water mask included the in the VNP03MOD geolocation product which includes land, coastline, and various types of water surfaces. Our binary mask is created by grouping together all water surfaces as a single water category, and grouping

together land and coastline as a single land category. 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. For each of the twenty inputs, a 3 pixel by 3 pixel array is extracted and is used to predict the cloudy or cloud-free label at the center pixel.

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. A subset of the VIIRS/CrIS fusion channels without solar contributions (Table 1.2) are used in a pseudo-labeling model for sun glint scenes (described later in section 1.3), but these are not used in the final NNCM model. Table 1.3 summarizes which inputs are used for the NNCM, a neural network without pseudo-labeling, and the pseudo-labeling model.

Dataset Splitting

In statistical modeling it is important to ensure independence between the training, validation, and testing datasets. 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 Håkansson 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 collocations by year into our training set that consists of 14.3 million collocations from 2016

and 2018, a validation set consisting of 5.7 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 the training stage. The testing set is used to provide estimates of model performance which we will analyze in section 1.4 and is not seen by the model during the training or hyperparameter tuning stages.

The spatial and seasonal distribution of these collocations can be seen in Fig 1.1. There are slight differences in spatial sampling between the testing dataset and the validation and training datasets. We expect that this is due to a combination of the strict 2-minute time difference we require of the collocations and the exit of CALIPSO from the A-train in late 2018 (Braun et al., 2019). We select 2019 for our testing dataset since it provides the most spatially and temporally complete dataset. 2016 and 2018 are used in our training dataset since they offer the next largest number of collocations. We judged that 2017 was the least spatially and temporally representative hence its use only as a validation dataset for hyperparameter tuning and early stopping during training.

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 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 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. 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 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.

1.3 Methods

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 observations. 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 shortwave infrared 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, and the final NNCM model uses all VIIRS channels and no VIIRS/CrIS fusion channels. Stated simply, adding these pseudo-labels to the training dataset incentivizes the final NNCM model to match the predictions of an infrared-only model in areas with sun glint.

We first train a pseudo-labeling neural network model using only channels that are unaffected by sun glint. For VIIRS, these channels are M14, M15, and M16. In addition to these VIIRS channels, we also use a subset of the VIIRS/CrIS fusion estimates of MODIS-like channels (MODIS bands 27-36, Table 1.2) that are similarly unaffected by sun glint, the binary land/water mask and the absolute value of latitude. The VIIRS/CrIS channels are included in an effort to make up for the loss of the shortwave and shortwave infrared VIIRS bands (M1-M13). After training, the pseudo-labeling model is then used to make predictions for SNPP VIIRS scenes with sun glint of angles of less than 40 degrees over water. For this purpose, we select scenes from the fifteenth day of every month in 2018 (a year included in our training dataset). This is done to ensure even representation of seasons and combinations of sun glint angle and latitude. Of these predictions, roughly one million pseudo-labels are randomly sampled without replacement and added to the original training and validation datasets as if they were obtained from CALIOP. 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. Instead, 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.

Before discussing the details of the NNCM, we train a naive model on only CALIOP

data ignoring the fact that sun glint scenes are not represented in order to better illustrate the purpose of pseudo-labeling. The neural network without pseudo-labels does not include solar zenith angle and sun glint zenith angle since these values for sun glint scenes are outside the range of values for these variables included in CALIOP collocations. The inputs to each model are summarized in Table 1.3.

In Fig. 1.2 we qualitatively compare the predictions of the NNCM (that is trained with pseudo-labels) to a neural network model that is not trained with these pseudo-labels. Without pseudo-labeling, 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 behavior is not surprising because sun glint is an out-of-domain prediction for the neural network without pseudo-labels. This issue is somewhat remedied by including pseudo-labels in training the NNCM (Fig. 1.2.d). Qualitatively, the ECM (Fig. 1.2.f) appears to be the least effected by sun glint and most able to correctly discriminate cloud-free from cloudy in the sun glint region. The MVCM (Fig. 1.2.e) over predicts cloud cover directly over the specular point, but captures small cloud variability surrounding it. The NNCM 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 lack of visible channels and 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.

To summarize, there are three neural network models trained in this work: (1) the NNCM,

(2) a neural network without pseudo-labels, and (3) the pseudo-labeling model. The NNCM is the approach we are proposing and evaluating. The neural network without pseudo-labels and the pseudo-labeling model are developed in support of the NNCM. The only purpose of the neural network without pseudo-labels is to illustrate the need for pseudo-labeling in Fig. 1.2. The purpose of the pseudo-labeling model is to provide training labels for the NNCM in sun glint scenes. Only the results from the NNCM are analyzed in Sections 1.4 and 1.5. In the following section we describe the details behind how the NNCM is trained.

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 Table 1.4. All except the last FC layer are followed by Leaky ReLU activation and 2.5% Dropout. Dropout is a neural network regularization technique where a fraction of the units in each layer are randomly ignored and helps prevent over-fitting. 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 Table 1.4 and through each layer group successively until the last sigmoid activation is reached. The last sigmoid activation bounds the output of the model between 0 (indicating cloud-free) and 1 (indicating cloudy).

The model in Table 1.4 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 with different randomly initialized weights. Two configurations with double the number of units in the FC layers reported slightly higher validation accuracies compared to that of Table 1.4 (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.

$$J = -(y \log \hat{y} + (1 - y) \log (1 - \hat{y})) \tag{1.1}$$

The neural network is trained to minimize binary cross-entropy, J (Eq. 1.1), where y is the label and $\hat{y}$ is the predicted probability. All inputs are scaled to have zero mean and unit variance with the means and standard deviations 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} with a mini-batch size of 4,098 examples. 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.

The development of the NNCM and the following analysis was performed using the TensorFlow (Abadi et al., 2016), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), and Matplotlib (Hunter, 2007) python libraries.

1.4 Results

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, Eq. 1.4) and true positive rate (TPR, Eq. 1.2; 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 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. 1.3). For each instance, a cloud fraction value is reported from CALIOP, the NNCM, 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, 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.

The NNCM 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 NNCM reports cloud fractions that 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 permanent snow

cover during the night where it underestimates by 30%. The ECM predicts roughly similar values to the NNCM with the exception of overestimating cloud cover during the night over sea ice by 12%.

$$TPR = \frac{TP}{P} \tag{1.2}$$

$$TNR = \frac{TN}{N} \tag{1.3}$$

$$ACC = \frac{TP + TN}{P + N} \tag{1.4}$$

$$BACC = \frac{TPR + TNR}{2} \tag{1.5}$$

In order to evaluate the performance of each cloud masking model, we calculate the balanced accuracy (BACC; Eq. 1.5) of all cloud masks across each surface type examined in Fig. 1.3. BACC is the mean of the true positive rate (TPR; Eq. 1.2), and the true negative rate (TNR; Eq. 1.3), 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 ACC (Eq. 1.4) 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 values are calculated for both the filtered (Table 1.5) and unfiltered (Table 1.6)

datasets. Table 1.5 represents the most reliable collocations, but this means that fractionally cloudy scenes, cloud edges, and boundary layer clouds are not well represented. The NNCM reports higher BACC over every surface type examined compared to both the ECM and MVCM for the both the filtered and unfiltered datasets. The most notable improvement from the NNCM occurs over sea ice, snow, and permanent snow during both day and night. McNemar's test (McNemar, 1947) is applied to the NNCM and the best operational model (either ECM or MVCM) for each category in both tables 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 0.001 in every comparison of the NNCM and the best operational model.

In a few cases, there are instances where one operational model has a higher TPR or TNR value than the NNCM for a particular surface type. We find that that when either the ECM or MVCM has a larger TPR value, it is often at the expense of a very low TNR value (and vice-versa for low TPR and high TNR). One notable example of this is nighttime sea-ice where the ECM has a TPR of 93.3% and a TNR of 36.6% in the analysis of the unfiltered data (Table 1.6). Another is nighttime permanent snow cover where the MVCM has a TPR of 43.6% and a TNR of 92.2%. The NNCM often has the most similar TPR and TNR values. However, this is not always the case. The largest TPR/TNR disparity for the NNCM is over nighttime water where it has a TPR of 93.6% and a TNR of 79.2%. This is a category where the MVCM has a smaller disparity between TPR and TNR, but still overall lower BACC than the NNCM. Generally when a model has a large disparity between TPR and TNR, that is an indicator of severely over-predicting one of the two classes.

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 as estimated from CALIOP (Fig. 1.4).

As expected, all cloud masks struggle with the identification of clouds that are optically thin and clouds that are close to the surface. The NNCM 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 NNCM has the highest TPR in the filtered dataset. This may indicate that the MVCM is better able to discriminate small clouds that are close to the surface. However, when these clouds are removed, the NNCM 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. 1.3. When considering optical depth, the NNCM 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 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. 1.4 and Tables 1.5 and 1.6 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 identified. The MVCM results in the TPR analysis of Fig. 1.4 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. 1.5). The cloud types are obtained from the 1 km CALIOP Cloud Layers product. Overall, the NNCM 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 for broken cumulus clouds implies that the NNCM has relatively worse performance in fractionally cloudy scenes compared to the MVCM. While these differences are fairly small, they may be indicative of a much larger difference in skill due to the relative unreliability of the unfiltered collocations. When examining the filtered dataset results for these same clouds, we see that the NNCM has the highest TPR. This suggests that the NNCM 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 NNCM 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. 1.3 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. 1.6). 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 NNCM 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 1.6 indicates that the NNCM 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 NNCM are not strictly due to the larger mean cloud fraction compared the MVCM. In addition to Tables 1.5 and 1.6, Fig. 1.6 implies that most of the improvement by the NNCM 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. 1.6) 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 NNCM 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. 1.7). McNemar's test is used to test the differences in model performance between the NNCM 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. 1.7.d, Fig. 1.7.f). Overall, the NNCM appears to be the least sensitive to latitude. Most large differences between the NNCM and the operational models occur over high latitude land. In particular, the NNCM 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 middle latitudes compared to the MVCM. The NNCM 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.

Similarly, we calculate the mean BACC on the same grid in Fig. 1.8 using the filtered testing dataset. The BACC values are somewhat noisier since areas with extremely high cloud fraction depend largely on the correct identification of a few cloud-free CALIOP profiles. An example of this is over the Southern Ocean, where the ECM has a large disparity between ACC (Fig. 1.7.e) and BACC (1.8.e). A slight tendency to overestimate cloud cover for this region yields very large differences to the NNCM (Fig. 1.8.f). Besides this example and some areas where the MVCM improves upon the NNCM in the Southern Ocean, the results are largely similar to those of Fig. 1.7.

All 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 and perhaps depends on 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. 1.3, and the BACC values in Tables 1.5 and 1.6, 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 NNCM and ECM.

In order to further probe the differences in these cloud masks, we recalculate BACC after removing clouds below an increasing lower optical depth threshold from our testing dataset (Fig. 1.9). 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 to converge when clouds with optical depths above both thresholds are

removed. As expected, when optically thin clouds are removed from our testing dataset, the BACC of all the cloud masks is improved. Consistent with Fig. 1.6, the filtered dataset has higher BACC for all scenarios. The NNCM 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. 1.9.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. 1.9.a), the ECM has larger BACC values above an 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 NNCM BACC values are still equal to or larger than the other cloud masks for high optical depth clouds in most scenarios suggests the NNCM 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 improvements in BACC by the NNCM in Tables 1.5 and 1.6 are not solely due to discrepancies in the detection of optically-thin clouds.

It may be initially unintuitive why some of the curves in Fig. 1.9 vary so little with the removal of optically thin clouds. This is partially due to the choice of BACC as our primary 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 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 Global Forecasting System (GFS) twelve-hour forecasts made every six hours (Fig. 1.10). These surface temperatures are matched to VIIRS observations by linearly interpolating in space and time from the preceding and subsequent GFS forecasts. Given the spatial and temporal resolution of the GFS products, these should only be interpreted as very rough estimates of the surface temperature. The differences are calculated after the removal of clouds below two different cloud optical depth thresholds: 0.3, and 3.0. As expected, all cloud masks perform well where the 10.8 μm measurement is significantly colder than the surface. The performance of all models decreases as the VIIRS 10.8μm brightness temperatures become more similar to or larger than the surface temperature. Figure 1.10.b illustrates that even for optically thick clouds, the performance of both operational models is largely dependent on thermal contrast with the surface. The NNCM 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 NNCM is not supplied with any information about surface characteristics other than latitude and whether it is viewing a land or water surface.

Uncertainty Assessment

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 of the original training dataset. With this

significant caveat in mind, we calculate the ACC with respect to the cloud probabilities of the NNCM and ECM, as well as the clear-sky confidence from the MVCM (Fig. 1.11). 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 NNCM appears to be the best calibrated with the ACC on the unfiltered collocations closely following 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 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.

Cloud Detection Consistency

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 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 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 examining TPR differences between these conditions at fixed cloud optical depths could help reveal 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 difference in TPR across daytime, nighttime, land, and water for three latitude bands (Fig. 1.12). An important consideration for Fig. 1.12 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 NNCM has TPR differences less than or equal to 5% for all scenarios examined except for the difference between 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. 1.12.d (low latitude nighttime water and nighttime land), the ECM is the only mask with differences greater than 5%. In Fig. 1.12.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. 1.12. In a few cases (Fig. 1.12.a, 1.12.b, 1.12.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 high latitudes (Fig. 1.12.i, 1.12.l) where differences are larger than 10% for clouds with optical depths greater than 1.0.

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 NNCM and the MVCM. We focus this analysis on Greenland because it is one of the worst performing regions for both masks. We process every S-NPP VIIRS scene in 2019 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 NNCM 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. 1.13.a, 1.13.b).

Consistent with the CALIOP validation, we observe large differences over the Greenland land mass (Fig. 1.13.c). The NNCM 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 of 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 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 to 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 as a function of time (Fig. 1.13.d). To account for some of the differences in sampling, this average only includes grid points from the NNCM 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 NNCM to follow the CALIOP 1 km or 5 km products closely. When calculating the mean cloud fraction, individual values on the regular latitude/longitude grid are weighted to account for differences in surface area between locations.

The largest differences occur in northern hemisphere winter, with better agreement between the MVCM and NNCM 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. 1.13.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 fraction. Despite differences in temporal sampling, the NNCM shows somewhat similar variability to both CALIOP products. Overall, the NNCM 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 NNCM 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 two cloud masks, the NNCM appears to give a more realistic assessment of cloud cover variability in this analysis and more closely aligns with that of CALIOP.

1.5 Discussion

There are few common themes in much of the analysis done in section 1.4. The BACC calculated over global averages of a few surface types suggests that the NNCM 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 NNCM 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. The NNCM's improvement at higher latitudes raises some questions on its dependence on latitude particularly since it is the only model that uses this information directly in its inputs. To test this dependency, we retrained and evaluated the NNCM after removing latitude, solar zenith angle, sun glint angle, and the land/water mask. The largest change in BACC was a decrease of -0.5% over nighttime water, and all other surfaces changed by less than 0.2%. Considering these results, it is probable that the NNCM depends on latitudinal mean cloudiness in some capacity over water (similar to the ECM over the Southern Ocean). However, it is difficult to assess how this information is utilized and whether it is serving a purpose similar to that of a climatological first guess, or if it is changing the usage of other

observations.

Despite training using an unfiltered dataset that contains fractionally cloudy pixels identified by CALIOP, the NNCM 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. Broken cloudiness is a consistent problem in using CALIOP as a reference. 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 NNCM. 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 NNCM uncertainty estimates can be interpreted. Reported uncertainties by the NNCM 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 associated with regions

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

There are many areas for improvement in the NNCM approach. For instance, we included all 16 moderate resolution channels in our algorithm. It is plausible that some channels 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 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 property algorithms trained with CALIOP.

Despite the large increase in BACC made by our NNCM 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 NNCM 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 fusion channels represent spectral regions not covered in the native VIIRS channels such as those with significant carbon-dioxide (MODIS bands 33-36) , water vapor (MODIS bands 27 and 28), and ozone (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.

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 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 estimated by CALIOP collocations. The relatively coarse-resolution of the ancillary data might cause issues around boundaries of surface types or around large horizontal gradients in surface temperature. This mischaracterization of the surface condition could result in errors in cloud detection if a given model is highly dependent on this information. This is potentially one of the explanations for the disparity in performance in instances of low thermal contrast with the surface. 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. 1.12 we conclude that the NNCM is the most consistent

in identifying clouds across various geographical, solar illumination, and surface conditions while controlling for cloud optical depth. There are several reasons why the NNCM 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 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 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 surface and illumination conditions could potentially contribute to very different spatial and temporal variability. Because of this, we argue that TPR differences over varying surface and illumination conditions could be useful metrics for identifying such issues in cloud mask development and assessment. We suspect that this is a particularly important consideration for the use of cloud masking approaches in climate records. For example, annual sea ice loss or trends in seasonal snow cover could produce erroneous trends in cloud cover if a given cloud mask's TPR differs significantly to that of ice-free ocean or snow-free land.

We note several potential caveats in the assessment of the NNCM 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 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 NNCM to estimate the class probabilities produced by the pseudo-labeling model, and not the predicted class labels themselves. This approach appears to be successful in preventing severe over-clouding of sun glint regions, but it can only be expected to perform as well as a model that uses infrared observations exclusively. There are very specific conditions in which the two operational masks outperform the NNCM and it may be possible to use MVCM or ECM predictions as pseudo-labels to address deficiencies in the NNCM if these conditions can be identified without the use of CALIOP. We have not evaluated how the NNCM performs specifically in cloud-free scenes with high aerosol loading in this analysis. We expect that the ability for CALIOP to distinguish cloud from aerosol layers could add an another layer of difficulty in addition to the ability of VIIRS observations to distinguish these features.

One source of bias in this assessment is our choice of using the 1 km CALIOP Cloud Layers products in the vast majority of our comparisons. It is possible that some optically thin clouds that are detected in the 5 km CALIOP product but are missed in the 1 km product could be correctly identified by the imager cloud masks. This is plausible in conditions such as daytime low-latitude ocean where a thin cirrus cloud has large thermal contrast with the surface. We have not investigated this specific concern in this work due to the difficulty of ensuring mutually representative collocations between the 5 km CALIOP product and

the 750 m observations. It is possible that the slight overestimation in daytime mean cloud fraction by the MVCM (Fig. reffig:fig03) could be due to the detection of clouds missed by the 1 km CALIOP product. For purely statistical approaches, like the NNCM, it is difficult to separate this possibility from that of over-predicting cloud fraction simply because cloudy scenes are more common than cloud-free.

1.6 Conclusions

In this work, we examine the performance of a neural network cloud mask (NNCM) 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 NNCM. However, the NNCM 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 NNCM. 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 NNCM 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 NNCM 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 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 NNCM.

Band	Spectral Range (μm)	Units
M1	0.400 - 0.421	Refl.
M2	0.436 - 0.451	Refl.
M3	0.477 - 0.496	Refl.
M4	0.541 - 0.561	Refl.
M5	0.662 - 0.680	Refl.
M6	0.738 - 0.752	Refl.
M7	0.843 - 0.881	Refl.
M8	1.225 - 1.252	Refl.
M9	1.368 - 1.383	Refl.
M10	1.571 - 1.631	Refl.
M11	2.234 - 2.280	Refl.
M12	3.598 - 3.791	BT [K]
M13	3.987 - 4.145	BT [K]
M14	8.407 - 8.748	BT [K]
M15	10.234 - 11.248	BT [K]
M16	11.405 - 12.322	BT [K]

Table 1.1: The band, spectral range, and units of all sixteen moderate resolution VIIRS channels. Each channel is expressed as a reflectivity (Refl.), or a brightness temperature (BT).

VIIRS/CrIS Fusion Channel	Spectral Range of MODIS Equivalent Channel (μm)
MODIS 27	6.535 – 6.895
MODIS 28	7.175 – 7.475
MODIS 29	8.400 – 8.700
MODIS 30	9.580 – 9.880
MODIS 31	10.780 – 11.280
MODIS 32	11.770 – 12.270
MODIS 33	13.185 – 13.485
MODIS 34	13.485 – 13.785
MODIS 35	13.785 – 14.085
MODIS 36	14.085 – 14.385

Table 1.2: The VIIRS/CrIS fusion channels used in the pseudo-labeling model. All channels are expressed as brightness temperatures.

Inputs	NNCM	Neural network without pseudo-labels	Pseudo-labeling model
M1-M13	X	X	
M14-M16	X	X	X
MODIS 27 - MODIS 36			X
| Latitude |	X	X	X
Solar Zenith Angle	X		
Sun Glint Angle	X		
Land/Water Mask	X	X	X

Table 1.3: Summary of the inputs included in the three neural networks used in this work. See the main text for description of each model.

Layer Group (LG)	Layer Type	Input Size	Output Size
LG1	FC(200), Leaky ReLU, Dropout(2.5%)	180 (3x3x20)	200
LG2	FC(200), Leaky ReLU, Dropout(2.5%)	200	200
LG3	FC(100), Leaky ReLU, Dropout(2.5%)	200	100
LG4	FC(50), Leaky ReLU, Dropout(2.5%)	100	50
LG5	FC(25), Leaky ReLU, Dropout(2.5%)	50	25
LG6	FC(1), Sigmoid	25	1

Table 1.4: The architecture of the NNCM. LG refers to Layer Group and is used to describe the collection of layers in each row. 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 which dropout is applied.

	NNCM			ECM			MVCM			Cloud Fraction	Number (Million)
	BACC	TPR	TNR	BACC	TPR	TNR	BACC	TPR	TNR		
Day Global	0.968	0.982	0.954	0.938	0.957	0.918	0.910	0.941	0.879	0.662	2.96
Night Global	0.934	0.960	0.908	0.849	0.927	0.772	0.876	0.853	0.900	0.721	2.91
Day Water	0.969	0.985	0.952	0.940	0.977	0.902	0.909	0.966	0.852	0.735	1.99
Night Water	0.932	0.976	0.888	0.842	0.969	0.715	0.893	0.899	0.887	0.803	1.99
Day Land	0.965	0.974	0.956	0.917	0.898	0.936	0.887	0.866	0.908	0.512	0.97
Night Land	0.916	0.906	0.927	0.808	0.791	0.825	0.808	0.705	0.912	0.542	0.91
Day Sea Ice	0.966	0.966	0.966	0.883	0.962	0.804	0.879	0.859	0.899	0.775	0.29
Night Sea Ice	0.895	0.932	0.859	0.661	0.944	0.379	0.790	0.663	0.917	0.757	0.31
Day Permanent Snow	0.961	0.964	0.959	0.885	0.840	0.929	0.822	0.739	0.905	0.421	0.30
Night Permanent Snow	0.863	0.832	0.895	0.701	0.671	0.731	0.694	0.461	0.927	0.578	0.36
Day Snow Land	0.954	0.961	0.947	0.855	0.859	0.852	0.864	0.825	0.903	0.631	0.16
Night Snow Land	0.920	0.927	0.913	0.758	0.827	0.688	0.778	0.675	0.880	0.617	0.19

Table 1.5: BACC, TPR, and TNR 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 because sea ice collocations are also counted in the water category, and the two snow categories are also counted in the land category. Cloud fraction is calculated from the CALIOP collocations.

	NNCM			ECM			MVCM			Cloud Fraction	Number (Million)
	BACC	TPR	TNR	BACC	TPR	TNR	BACC	TPR	TNR		
Day Global	0.905	0.934	0.877	0.879	0.906	0.853	0.851	0.902	0.801	0.635	3.63
Night Global	0.879	0.920	0.838	0.808	0.889	0.726	0.830	0.816	0.843	0.687	3.46
Day Water	0.900	0.937	0.863	0.876	0.930	0.822	0.842	0.935	0.749	0.691	2.48
Night Water	0.864	0.936	0.792	0.796	0.930	0.663	0.832	0.860	0.804	0.747	2.45
Day Land	0.910	0.925	0.895	0.865	0.835	0.895	0.839	0.807	0.871	0.515	1.16
Night Land	0.884	0.870	0.899	0.782	0.754	0.810	0.783	0.671	0.895	0.542	1.01
Day Sea Ice	0.931	0.941	0.922	0.851	0.944	0.759	0.852	0.832	0.872	0.757	0.31
Night Sea Ice	0.870	0.906	0.834	0.650	0.933	0.366	0.772	0.640	0.903	0.741	0.33
Day Permanent Snow	0.930	0.928	0.932	0.854	0.790	0.917	0.795	0.692	0.899	0.430	0.32
Night Permanent Snow	0.836	0.797	0.875	0.684	0.646	0.722	0.679	0.436	0.922	0.577	0.40
Day Snow Land	0.905	0.920	0.891	0.818	0.815	0.820	0.827	0.779	0.875	0.619	0.19
Night Snow Land	0.887	0.890	0.885	0.737	0.797	0.678	0.756	0.641	0.870	0.610	0.21

Table 1.6: Same as Table 5, but all metrics are computed for the unfiltered collocations.

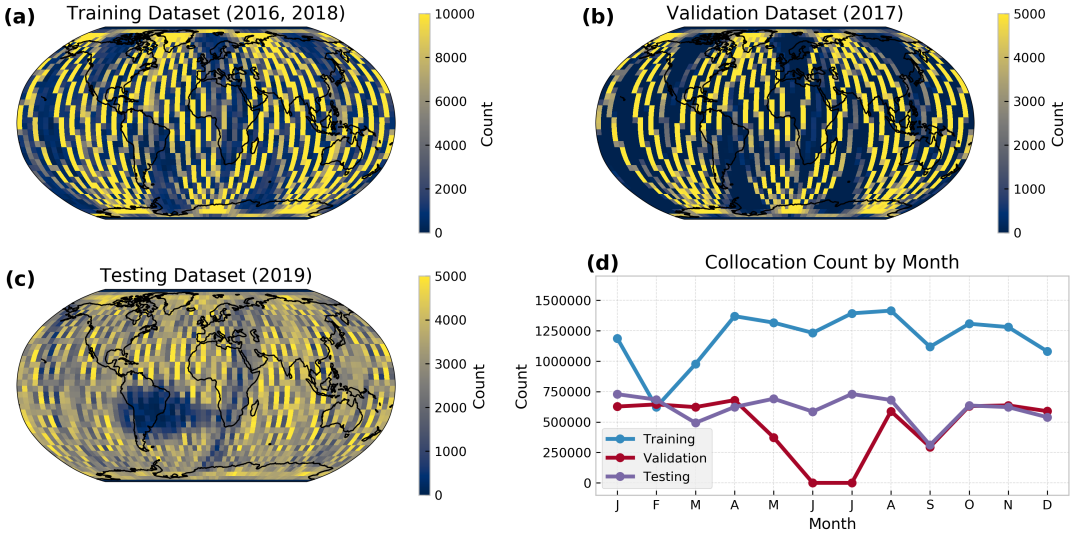

Figure 1.1: Spatial distribution of the unfiltered S-NPP VIIRS/CALIOP collocations for the (a) training, (b) validation, and (c) testing datasets. Panel (d) indicates the seasonal distribution of collocations for each unfiltered dataset. Note the difference in color bar limits between (a), (b), and (c).

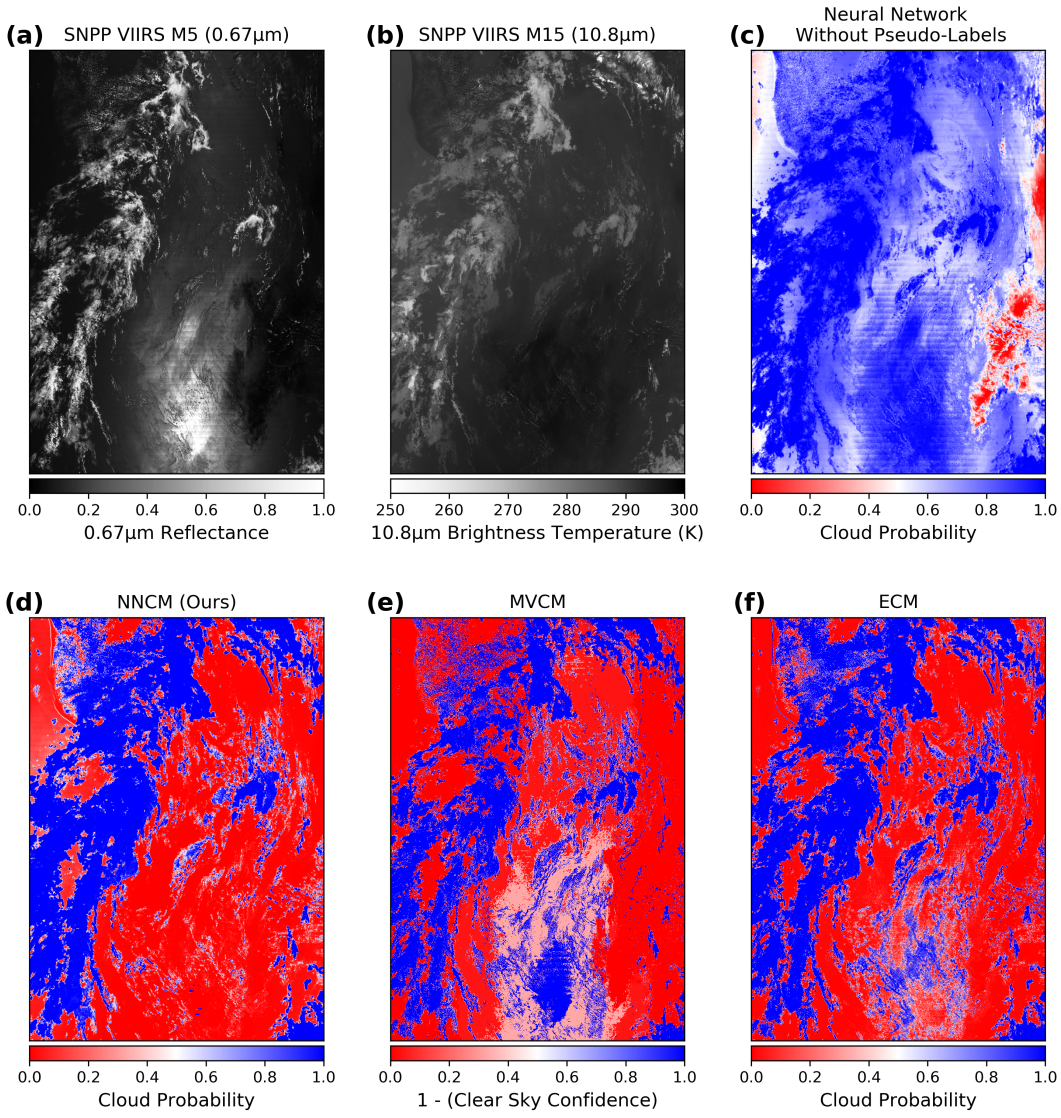

Figure 1.2: Comparison of the neural network cloud mask without pseudo-labels (c),the NNCM (d), the MVCM (e), and the ECM (f). Also shown are band M5 with a central wavelength of roughly 0.67µm (a) and band M15 with a central wavelength of roughly 10.8µm (b).

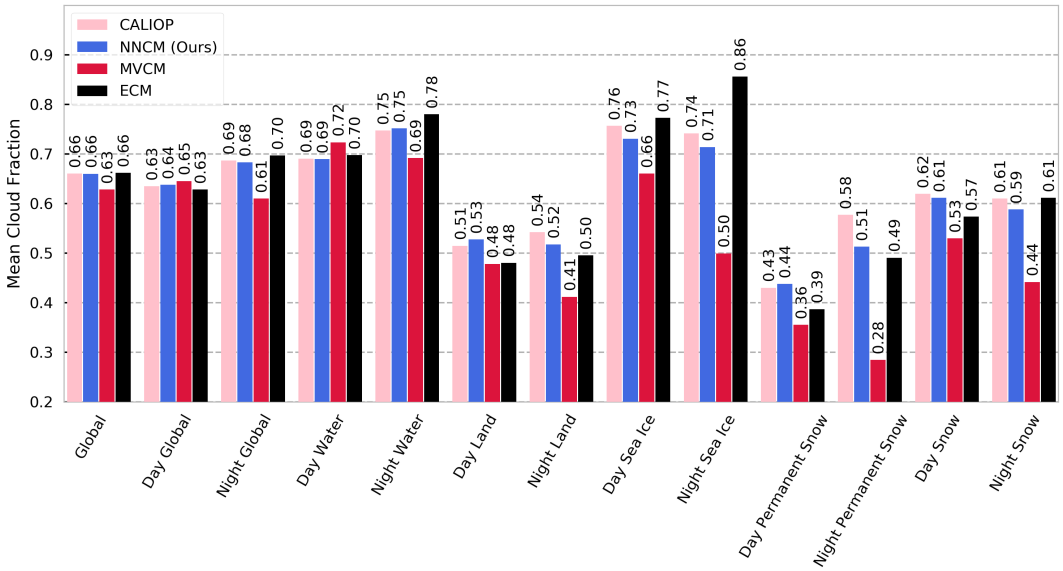

Figure 1.3: Mean cloud fraction for the 2019 unfiltered testing dataset. Each bar grouping from left to right shows the value from the CALIOP 1 km product, the NNCM, MVCM, and ECM. Time of day and surface categorizations are described in the main text.

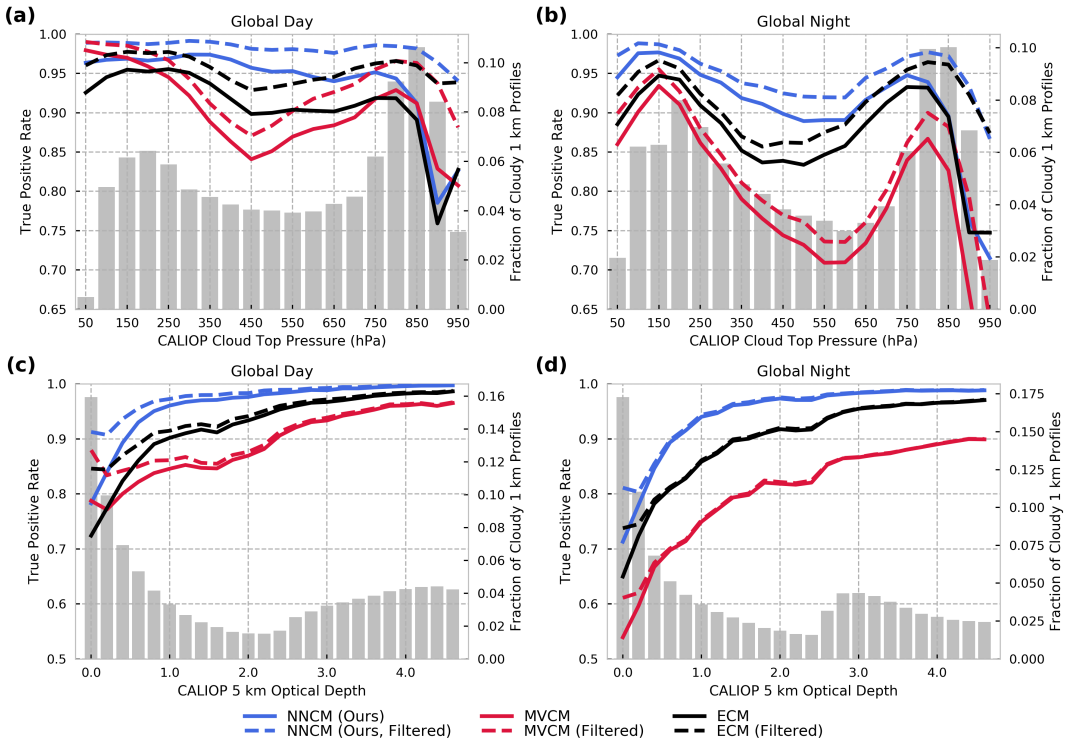

Figure 1.4: True positive rate (TPR) calculated as function of cloud-top pressure (a,b) and optical depth (c,d) for daytime and nighttime collocations respectively. The grey bars represent the fraction of cloudy 1 km CALIOP profiles. Only profiles with non-zero optical depths are included in (c) and (d).

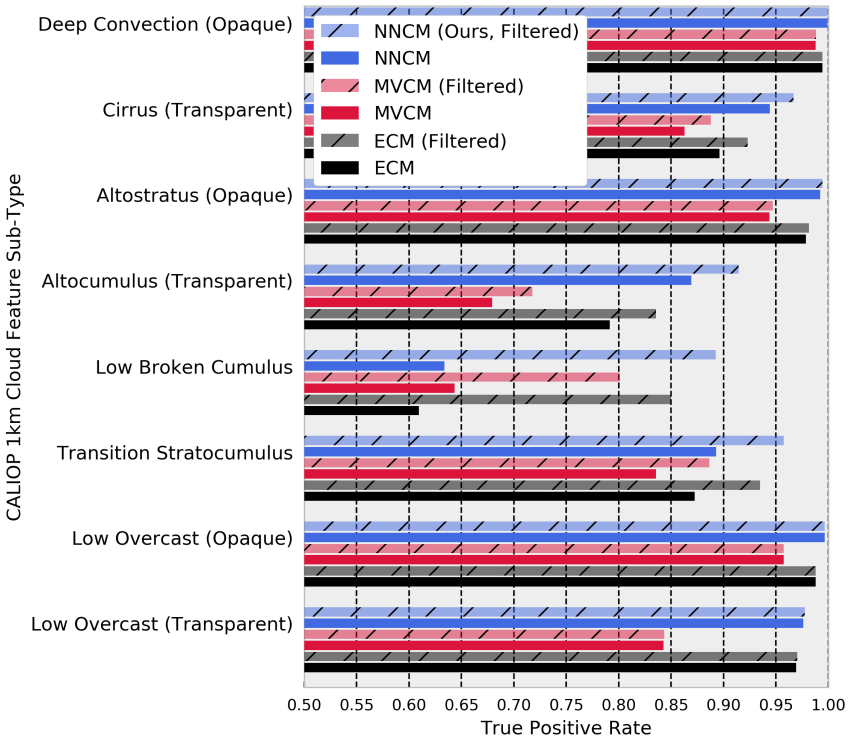

Figure 1.5: The True Positive Rate (TPR) for various CALIOP cloud-feature types from the 1 km CALIOP Cloud Layers product. The order shown in the legend indicates the ordering of the bars in each grouping.

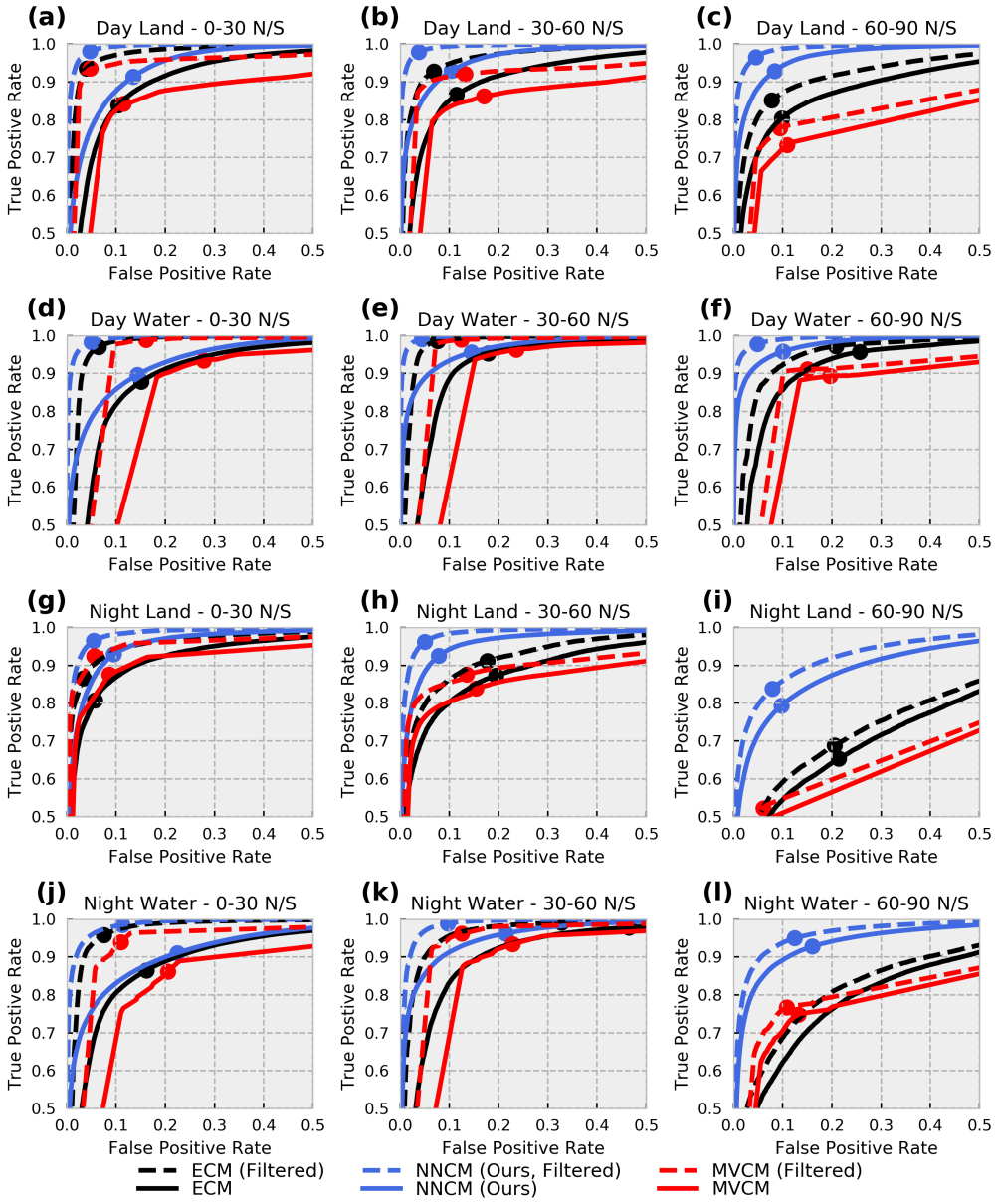

Figure 1.6: Receiver operating characteristic (ROC) curves for all three cloud masks. The text above each subplot indicates the subset of collocations for which the curves are plotted. Note that the x and y axis limits are somewhat atypical for ROC curve plots and are chosen here to emphasize the differences 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.

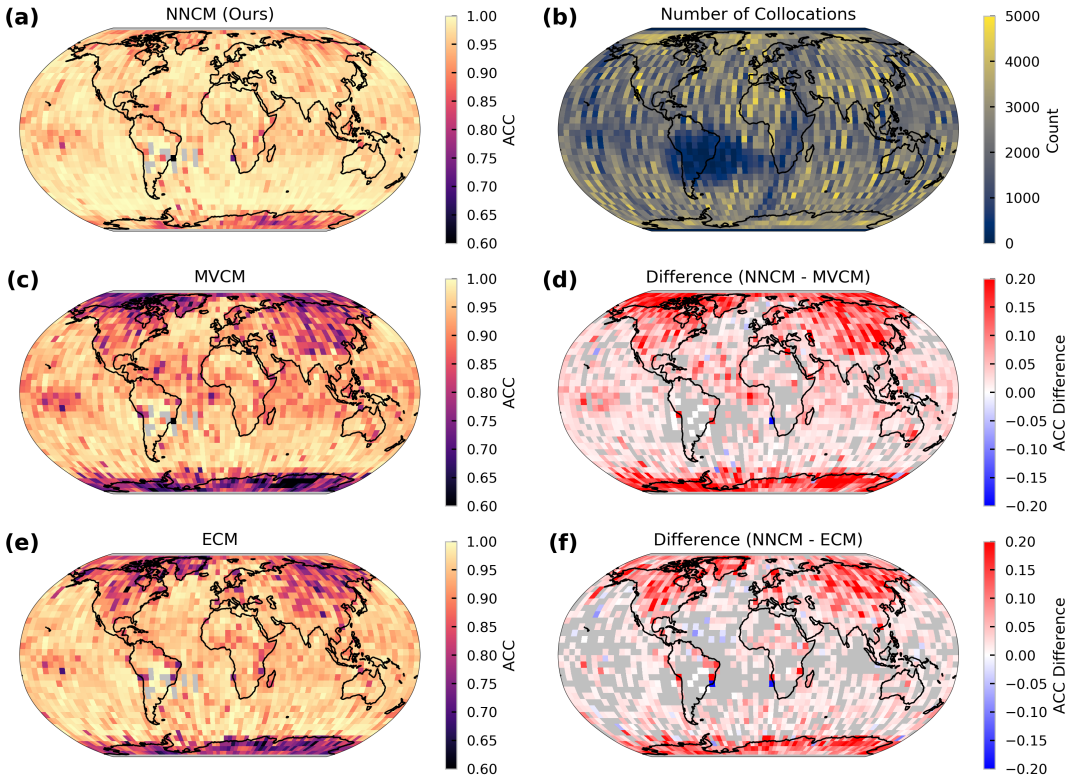

Figure 1.7: Geographic comparison of the ACC between the three cloud masks on the filtered testing dataset. Each grid cell is 5 degrees latitude by 5 degrees longitude. The gap in coverage over South America is due to the removal of low-energy laser shots from the CALIOP datasets. Cells with less than 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.

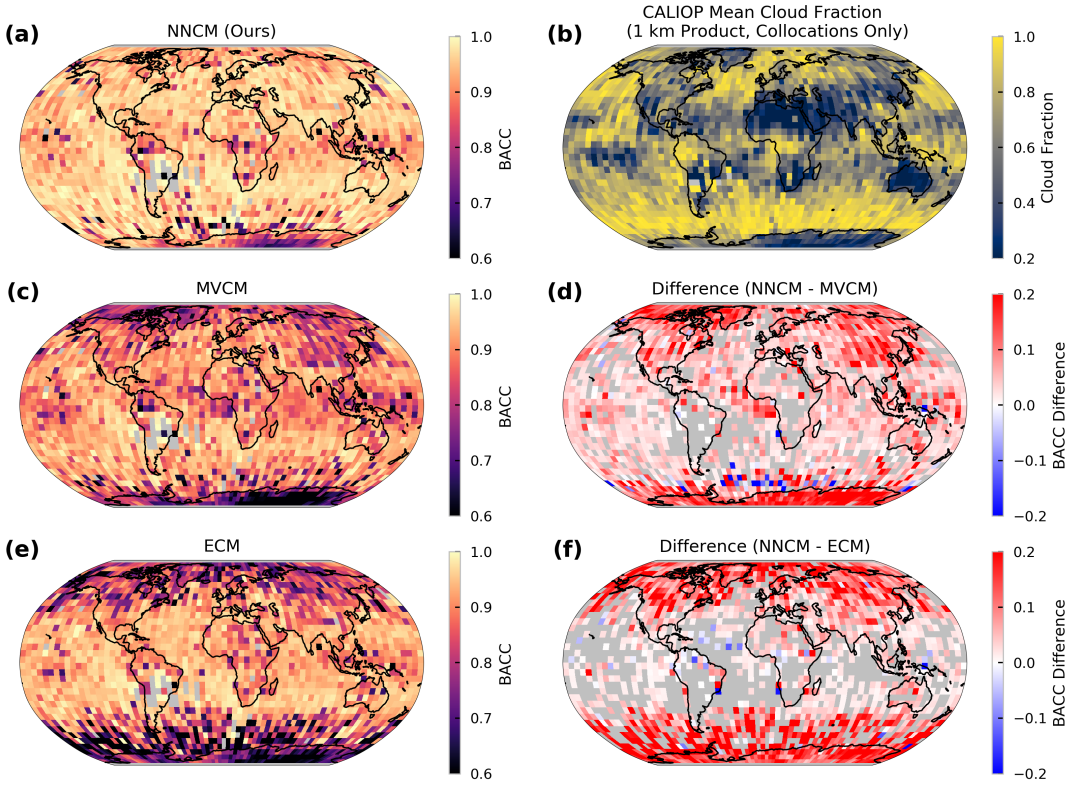

Figure 1.8: Same as Fig. 1.7 but all using BACC instead of ACC. Panel (b) has been replaced with the 1 km CALIOP cloud fraction computed from the VIIRS/CALIOP collocations.

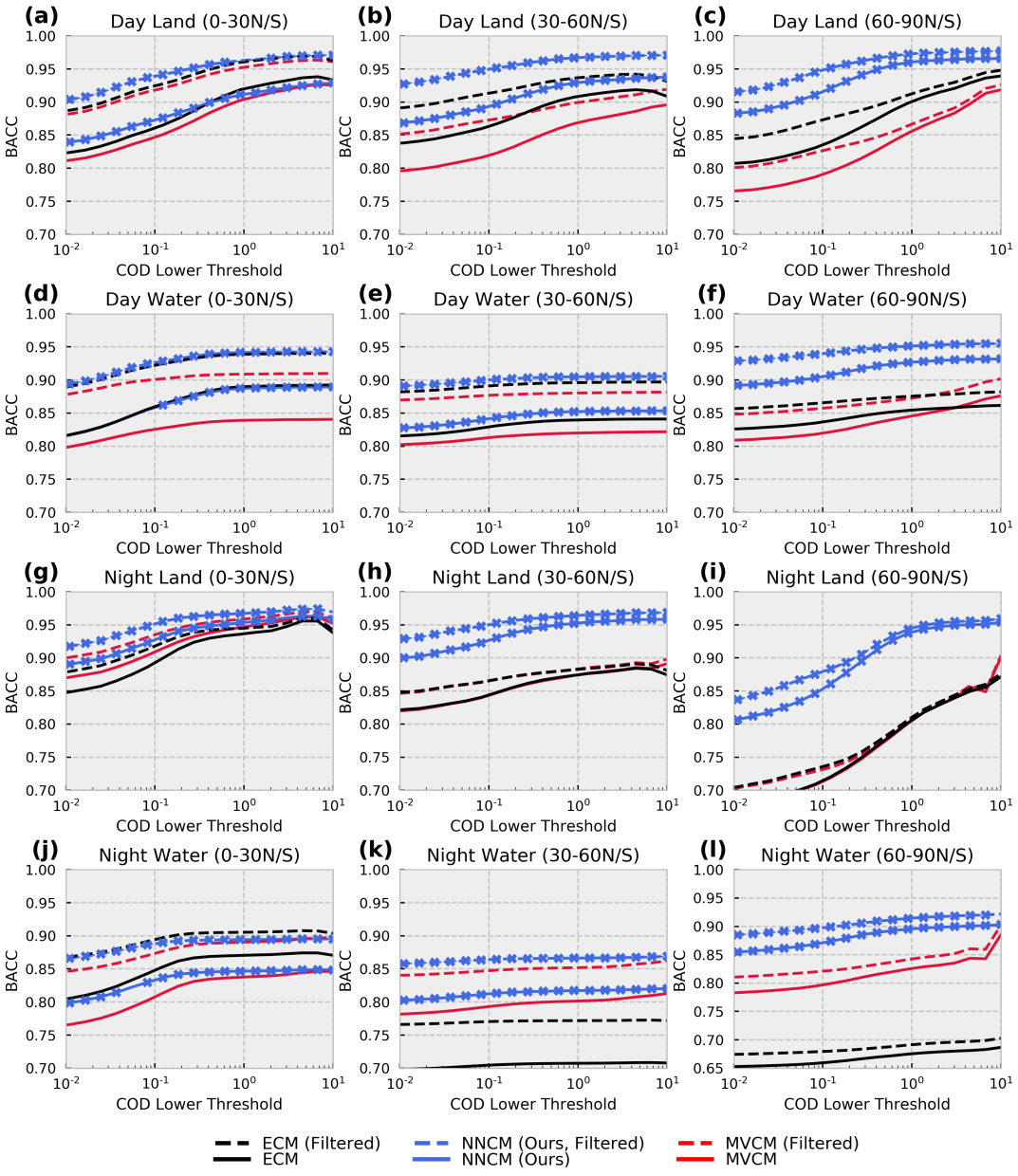

Figure 1.9: Balanced Accuracy (BACC) recalculated after removing clouds below a certain cloud optical depth (COD) threshold. Tick marks on the neural network lines indicate significant differences in performance between the neural network and the best operational model using McNemar's test with p-values less than 0.001. Note that the y-axis limits are different for (l) compared to the other subplots.

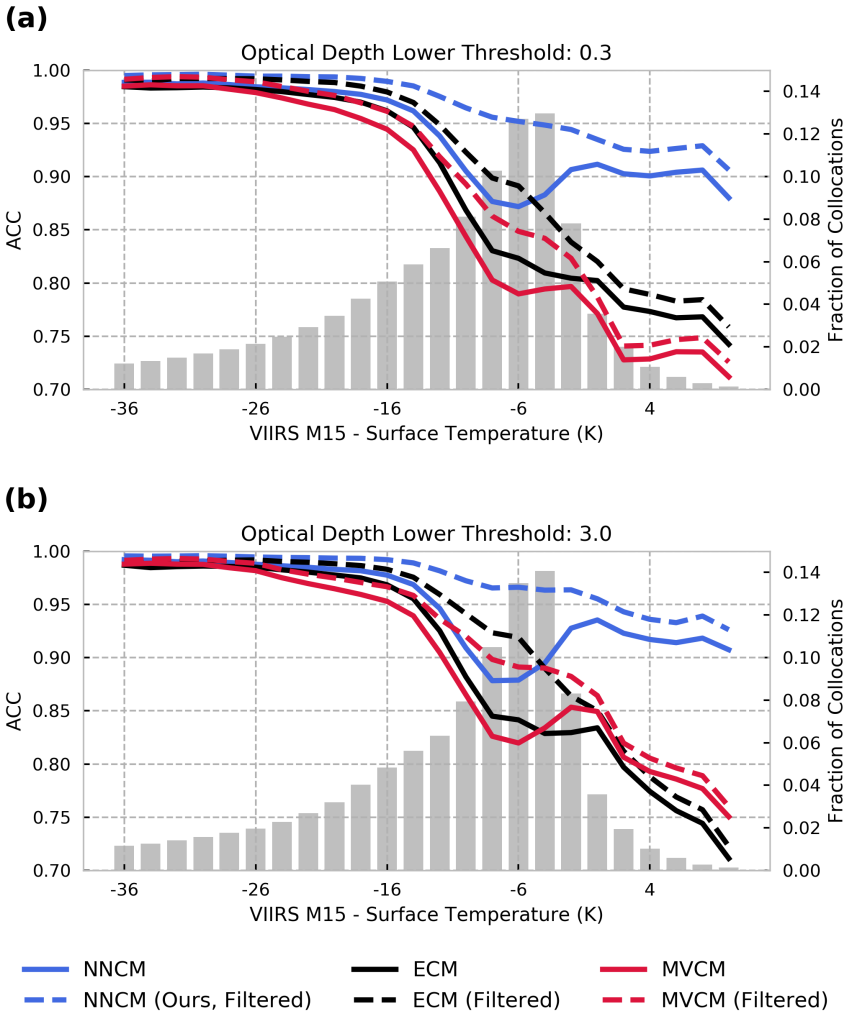

Figure 1.10: ACC calculated as a function of thermal contrast with the surface approximated by the difference between VIIRS M15 (10.8 µm) and surface temperature in Kelvin. Each subplot represents a set of collocations consisting of clear-sky scenes and cloudy scenes with optical depths greater than 0.3 (a) and 3.0 (b).

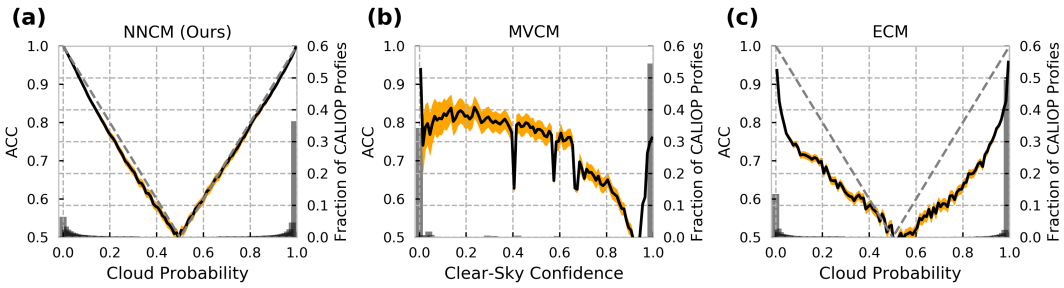

Figure 1.11: Uncertainty assessments for (a) the NNCM (b) the MVCM, and (c) the ECM. ACC values (left y-axis) for cloud probability and clear sky confidence values are calculated for bins of size 0.01. For (a) and (c) a perfectly-calibrated model is plotted with the grey dashed line (see main text). Orange shading indicates the 99.9% confidence interval. Grey bars indicate the fraction of collocations falling within each bin of width 0.01.

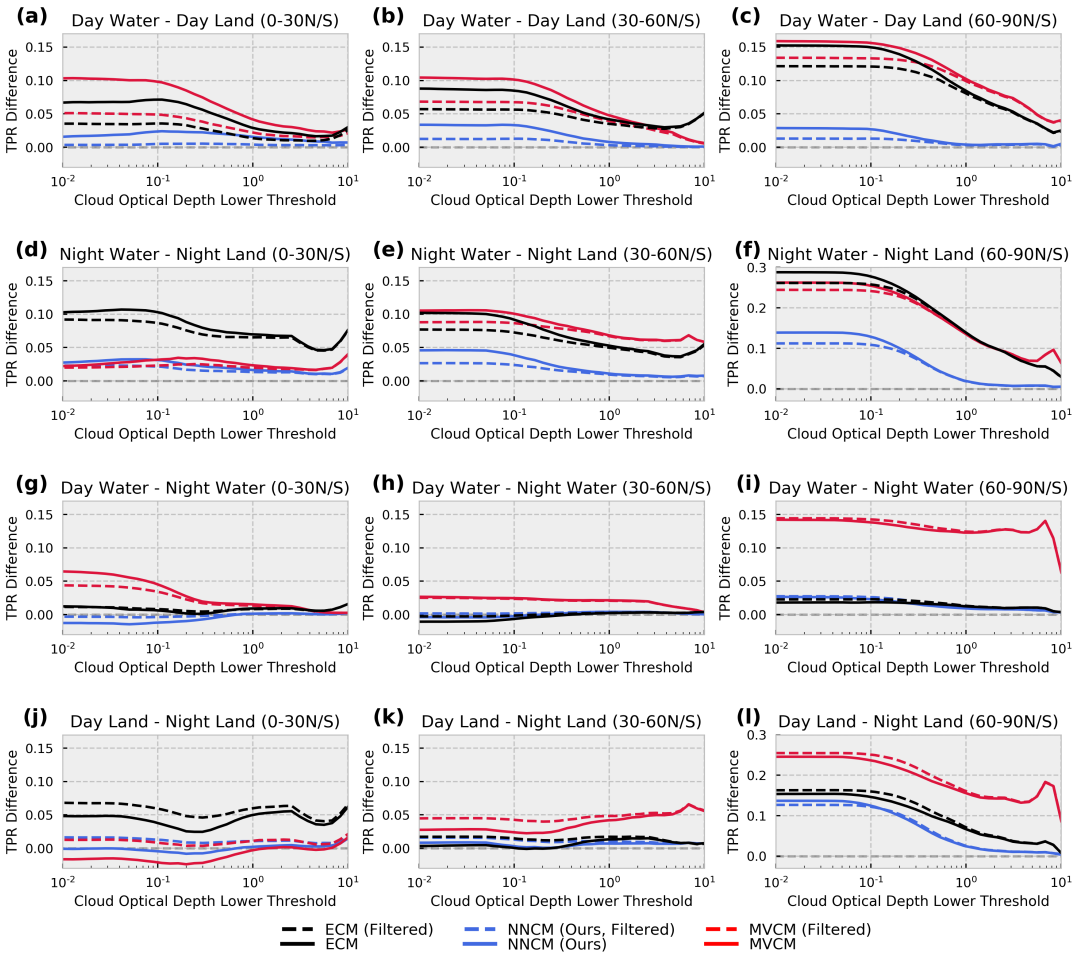

Figure 1.12: TPR differences over combinations of land/water and day/night conditions. The specific TPR difference and latitude is labeled at the top of each subplot. Note that the y-axis limits are different for (f) and (l).

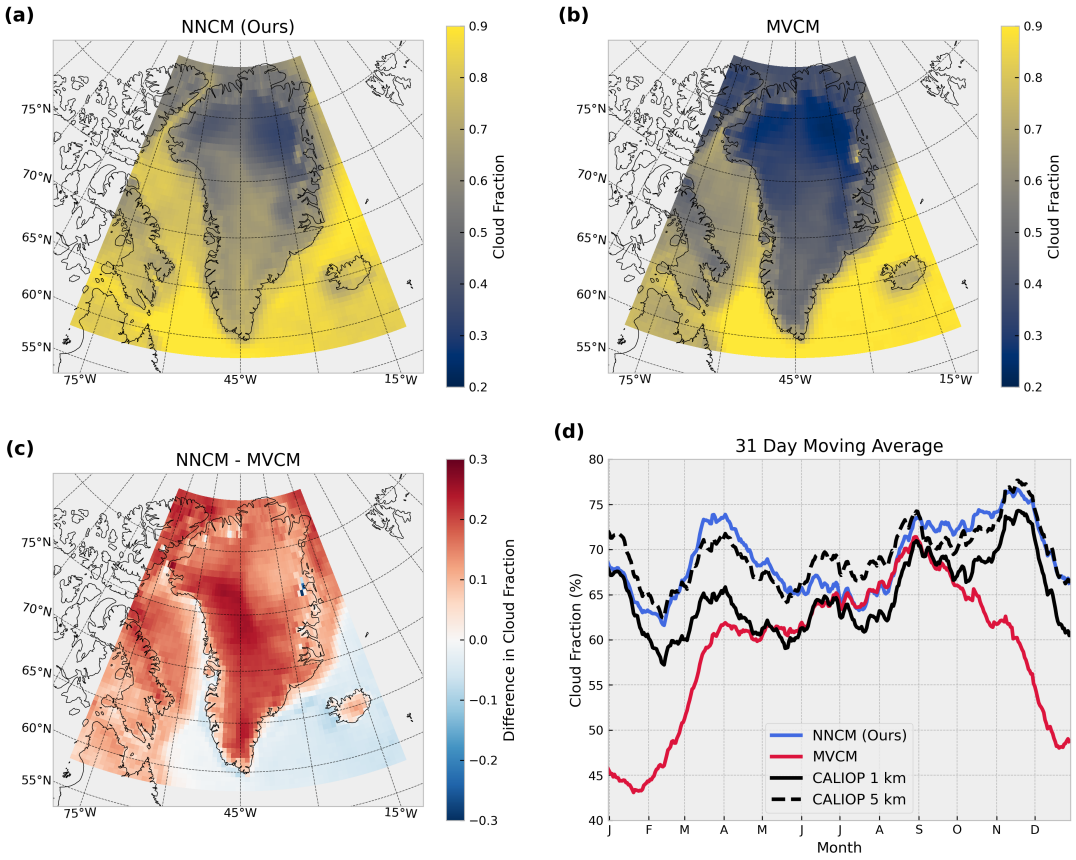

Figure 1.13: Regional analysis of cloud fraction over Greenland. (a) and (b) illustrate the mean cloud fraction for the NNCM and the MVCM for all selected VIIRS scenes in 2019. (c) is the difference between (a) and (b). (d) is the domain-wide 31-day moving average of grid points spatially matched with CALIOP (see main text for details).

2 PROBING THE INTERPRETABILITY OF NEURAL NETWORK CLOUD-TOP PRESSURE MODELS FOR LEO AND GEO IMAGERS

2.1 Introduction

Cloud-top pressure (CTP) is a useful derived product for characterizing clouds and their variability from satellite measurements. CTP can be used in combination with cloud optical depth (COD) to distinguish cloud types such as convective cloud-tops, cirrus, and stratocumulus (Jakob and Tselioudis, 2003). When applied to long-term imager records, such an analysis can be used to identify changes in cloud type (Foster and Heidinger, 2014) or to assess relationships among aerosol loading and cloud type (Oreopoulos et al., 2017). CTP also has applications in downstream cloud products such as the cloud cover layers (CCL) product relevant for aviation nowcasting (Seaman et al., 2017; Noh et al., 2017) and the height assignment of derived motion winds (Daniels et al., 2012).

Several approaches have been developed to estimate CTP from imagers. Many physically-based methods rely on differences between absorbing and non-absorbing infrared channels or require the use of radiative transfer models. Early efforts include Chahine (1974) and Smith and Platt (1978), which explore the use of CO_2-absorbing channels. The Moderate Resolution Imaging Spectroradiometer (MODIS) CTP products (Menzel et al., 2008) employ a similar CO_2-slicing approach. Each MODIS CO_2 channel has differing amounts of CO_2 absorption, so each is sensitive to different levels of the atmosphere. As a result, differences among these channels can be used to infer cloud-top height and pressure.

Inoue (1985) used a split window (11-μm and 12-μm) method to obtain cloud-top tem-

perature for cirrus clouds. Heidinger and Pavolonis (2009) used a similar approach for estimating phase, temperature, and COD for cirrus clouds, relying on multiple channels within the 8-μm to 13-μm region. Specifically, their approach relies on an optimal estimation methodology (Rodgers, 1976) for the Advanced Very High-Resolution Radiometer (AVHRR). This is later formalized as the Algorithm Working Group (AWG) Cloud Height Algorithm (ACHA) for use in operations for the Advanced Baseline Imager (ABI) and Visible Infrared Imaging Radiometer Suite (VIIRS; Heidinger and Li 2017). Optimal cloud analysis (OCA; Poulsen et al. 2012) also uses an optimal estimation method for several cloud properties, including CTP. OCA additionally has the ability to estimate properties of multiple clouds in multi-layer scenes (Watts et al., 2011).

Machine learning (ML) has recently gained popularity in atmospheric science and remote sensing. ML approaches can often be successful in prediction tasks due to their ability to exploit complex relationships between multiple features (predictors) and the corresponding label (predictand). Some methods can rely heavily on feature-engineering, which is the practice of transforming features to make the relationship with the label more suitable for a given model. Neural networks (NNs) are less dependent on feature-engineering due to their use of successive nonlinear operations. NNs have proven useful in a variety of atmospheric science applications, including automated identification of frontal boundaries (Lagerquist et al., 2019), detection of severe convection (Cintineo et al., 2020), and estimation of microphysical properties of snowfall (Chase et al., 2021).

NNs have also been applied to CTP estimation. Kox et al. (2014) used a simple NN trained with the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) to detect and estimate COD and cloud-top altitude of cirrus clouds with the Spinning Enhanced

Visible and InfraRed Imager (SEVIRI). Håkansson et al. (2018) also trained an NN with CALIOP to estimate CTP from MODIS measurements. They compared their NN results with operational algorithms and found large improvement even when using a small subset of channels. Pfreundschuh et al. (2018) extended this work to estimate uncertainties in CTP from NN approaches with a quantile loss function.

Interpretability is a major concern when choosing an NN to solve a given task. Methods like CO_2-slicing, ACHA, and OCA are well-grounded in the physics of radiative transfer. One can often attribute the predictions from these methods to physical aspects of the observations and environment. It can be difficult to explain predictions NN that has successive nonlinear operations. Several efforts have been made to promote the use of various interpretation methods in applications of ML to atmospheric science (McGovern et al., 2019).

We quantify the importance of each channel in NN CTP models for one low-earth orbiting (LEO; VIIRS) and one geostationary (GEO; ABI) imager. These models are trained to match estimates of CTP from CALIOP. Our NN approach largely builds off that of Håkansson et al. (2018) and Pfreundschuh et al. (2018). We first perform a short validation for both VIIRS and ABI and include a comparison with ACHA. We apply several approaches to offer varied perspectives on the importance of the infrared channels used in these models. The overall goal of this analysis is to enhance our understanding of what information is most useful for CTP estimation and is motivated by the substantial increase in performance offered by NNs over more traditional methods. We view model interpretability as an important consideration for applications of operational CTP products. Furthermore, we hope to inform cloud property algorithm development and the channel selection of

instruments focused on remote sensing of cloud properties.

2.2 Data

CALIOP

CALIOP (Winker et al., 2009) is a spaceborne near-nadir-pointing lidar, measuring backscatter intensity at 1064-nm and 532-nm. CALIOP is sensitive to optically thin clouds, making it a suitable source for the validation of several cloud properties from passive imagers. A critical choice in this work is whether to use the 1-km or 5-km CALIOP cloud layer products (Vaughan et al., 2009) to train the NN models. The 1-km product has a spatial resolution most commensurate with both imagers, but the 5-km product has greater sensitivity to cirrus clouds. Thus, there is a trade-off between the representation of fine-scale variability in CTP and the detection of optically thin clouds. Another factor is that COD is only calculated for the 5-km product and will only be representative for clouds detected at 5-km. Cloud-top height estimates are required for the parallax correction of imager measurements, meaning that the choice of CTP product affects the selection of collocated imager pixels. We decide to use the 1-km product for the training and validation of the neural networks, since passive imager measurements are likely not sensitive to optically thin clouds detected by the 5-km product that are missed by the 1-km product.

VIIRS

VIIRS is a LEO imager on the Suomi-National Polar Orbiting Partnership (S-NPP) and NOAA-20 satellites. VIIRS has a nadir spatial resolution of 750-m for 16 moderate-

resolution channels that span visible, near-infrared, and infrared wavelengths. We find coincident observations between S-NPP VIIRS and CALIOP by nearest-neighbor matching of imager pixels and lidar profiles that occur within 2.5 minutes. A parallax correction (Holz et al., 2008) is applied, using the VIIRS viewing geometry and the cloud-top altitude reported from CALIOP.

Models trained on coincident observations between VIIRS and CALIOP can have generalization issues related to viewing angles and solar geometry (White et al., 2021). In this dataset, high latitude collocations are only made at relatively low VIIRS viewing angles. Sun glint poses a significant problem since it is never seen in our collocation dataset (White et al., 2021). To reduce the impact of this issue, we do not include channels that have solar contributions which limits the channels to those with central wavelengths of 8.6-µm, 10.8-µm, and 12.0-µm.

In addition to the native VIIRS channels, we also include information from the VIIRS/Cross-track Infrared Sounder (CrIS) fusion channels (Weisz et al., 2017). These are estimates of absorbing channels created from coarse-spatial-resolution measurements from CrIS that are convolved to match the spectral response functions of MODIS. The fusion channels are mapped to the same resolution as the VIIRS M-bands by exploiting the variability in the native VIIRS infrared channels. Others have found improvement in CTP estimates when including the fusion channels (Li et al., 2020) or CrIS information (Heidinger et al., 2019). They are included here since they represent spectral regions not represented in native VIIRS observations. All bands from VIIRS and the fusion channels used in this work are shown in Table 2.1.

We partition our VIIRS/CALIOP collocations into a training dataset from 2016 and

Source/Band	Central Wavelength
VIIRS - M14	8.6-µm
VIIRS - M15	10.8-µm
VIIRS - M16	12.0-µm
VIIRS/CrIS fusion – MODIS 27	6.7-µm
VIIRS/CrIS fusion - MODIS 28	7.3-µm
VIIRS/CrIS fusion – MODIS 30	9.7-µm
VIIRS/CrIS fusion – MODIS 33	13.3-µm
VIIRS/CrIS fusion – MODIS 34	13.6-µm
VIIRS/CrIS fusion – MODIS 35	13.9-µm
VIIRS/CrIS fusion – MODIS 36	14.2-µm

Table 2.1: Central wavelengths of the infrared channels included in the VIIRS models. The left column indicates whether channels are native VIIRS measurements or derived from the CrIS. Note that fusion channels are named after MODIS bands since they are designed to match spectral response functions of that instrument.

2018, a validation dataset from 2017, and a testing dataset from 2019. The spatial and seasonal distribution of collocations are shown in Figure 2.1. Differences in the spatial distribution between 2019 and the previous years are due to CloudSat and CALIOP's exit from the A-Train (Braun et al., 2019). Gaps in these datasets are primarily due to the unavailability of CALIOP data and a gap in some CrIS channels from April 2019 to June 2019.

ABI

ABI (Schmit et al., 2017) is an imager on the Geostationary Operational Environmental Satellite (GOES) -16 and -17 platforms. The infrared channels considered in this work have a nadir spatial resolution of 2-km. The temporal resolution of ABI full-disk images can vary depending on the scan mode. We use the GOES-16 ABI data from 2019, in which the

ABI Band	Central Wavelength
8	6.2-µm
9	6.9-µm
10	7.3-µm
11	8.4-µm
12	9.6-µm
13	10.3-µm
14	11.2-µm
15	12.3-µm
16	13.3-µm

Table 2.2: Central wavelengths of the infrared channels included in the ABI models.

temporal resolution is mainly ten minutes.

The GOES-16 ABI and CALIOP collocations are found in a similar way to those of VIIRS and CALIOP. One difference is that we relax the time difference requirement to five minutes. We make this change since the nadir resolution of ABI is more than twice as large as the VIIRS M-bands, and it is less likely that a cloud observed by CALIOP is advected out of the matched imager pixel when the area observed by the pixel is larger. In our models we include all ABI channels without solar contributions, which includes bands 8 through 16 (Table 2.2).

The collocations with CALIOP are partitioned into a training dataset from January 2019 through June 2019, a validation dataset from July 2019 through September 2019, and a testing dataset from October 2019 through December 2019 (Figure 2.2).

NWP Data

Numerical weather prediction (NWP) model output fields are included in our NNs in order to characterize the environment of observed clouds. We use the 6-hour forecast from

the 6-hourly Climate Forecast System (CFS) 0.5-degree output (Saha et al., 2014) and match each set of CALIOP collocations by linearly interpolating in space and time. The fields used are the temperature and pressure at the surface and tropopause, total precipitable water, and the temperatures at pressure levels of 20 hPa, 100 hPa, 300 hPa, 500 hPa, 700 hPa, and 900 hPa.

The information contained in many of the infrared channels used is closely related to cloud-top temperature for opaque clouds. If temperature can be determined, then an NWP temperature profile can be used to infer the pressure level. Most clouds occur in the troposphere so the temperature and pressure of the tropopause and the surface might serve as upper and lower bounds. Total precipitable water might serve as an indicator for optically thick cloud cover, and provide information on the expected amount of water vapor absorption. We experimented with including relative humidity, and a greater number of pressure levels (not shown). These did not substantially help model performance so they were not included. Our resulting temperature profile has a similar sparsity to Håkansson et al. (2018).

2.3 Neural Network Training and Validation

Neural Network Details

We use neural network with a quantile loss function that draws from Pfreundschuh et al. (2018) which demonstrated the ability of quantile regression NNs to estimate uncertainties for CTP. The quantile loss is shown in Eq. 2.1 where L is the loss for a prediction $\hat{y}$ for quantile τ and where y is the CALIOP CTP. When multiple quantiles are estimated, L can be

averaged over multiple values of τ. The implications of Eq. 2.1 are that for larger quantiles overestimates are penalized more than underestimates (and the opposite for lower quantiles).

$$L(\tau, y, \hat{y}) = \begin{cases} (1-\tau)\,|y-\hat{y}| & \text{for } y \leq \hat{y} \\ \tau\,|y-\hat{y}| & \text{for } y > \hat{y} \end{cases} \tag{2.1}$$

Each network has four fully connected layers consisting of 64, 32, 16, and 9 units. These values were determined by starting with the architecture used in Håkansson et al. (2018). We found moderate improvement after doubling the number of units and adding an additional layer. Further increases in the number units increased the computational expense, but did not improve performance. All layers except the last are followed by rectified linear unit (ReLU) activations. The last layer represents the nine evenly spaced quantiles we estimate (τ ranging from 0.1 to 0.9 in increments of 0.1) and has no activation function. Predictions for CTP are obtained from the 50th quantile. A frequent problem in quantile regression is the crossing of quantiles since there is no mechanism to ensure the curves do not overlap. In our datasets, we observe crossing quantiles in less than 2% of predictions which we judged to be small enough to ignore for our applications. Other works have suggested solutions for minimizing the crossing of quantiles (Cannon, 2018).

The Adam optimizer (Kingma and Ba, 2015) is used starting with a learning rate of 5×10^{-3} determined using a learning rate range test (Smith, 2017). The batch size was increased from 250 used in Håkansson et al. (2018) to 5,000 where the time taken for each epoch stopped decreasing. The loss is evaluated on the validation dataset after each epoch. The learning rate is reduced by a factor of 10 when the validation loss has not decreased for the last 5 epochs to a minimum of 1×10^{-6}. Training is stopped when the validation loss

has not decreased for the previous 9 epochs. These values are subjectively chosen based on the number of epochs needed for the loss to stop decreasing after each learning rate reduction.

The inputs of each NN include the infrared channels specified in Table 2.1 for VIIRS and Table 2.2 for ABI. In addition to these values and the NWP information, we include several spatial metrics derived from a 5-by-5 pixel array surrounding the central pixel where the prediction is made. These spatial metrics include differences between the central pixel and both the coldest and warmest pixels and the standard deviation of all 25 pixels calculated for all channels. In total, the VIIRS NN includes 51 inputs and the ABI NN includes 47 inputs. All inputs are standardized by subtracting the mean and dividing by the standard deviation calculated from the training dataset. The CALIOP observations of CTP are divided by 1000, meaning predicted CTP values typically lie between 0 and 1. Standardizing the CALIOP observations had no impact on performance, but consistently reduced training time by several epochs.

The NNs are trained using TensorFlow (Abadi et al., 2016) on a Quadro RTX 6000. The following analysis was performed using the NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), and Matplotlib (Hunter, 2007) software libraries.

Neural Network Performance Evaluation

Due to our choice of the 1-km CALIOP CTP product, when we analyze with respect to optical depth, we can only compare instances where the 1-km and 5-km products have identified roughly the same cloud layer. Otherwise, one risks using the optical depth of a cloud to characterize the cloud-top pressure of another cloud lower in the atmosphere.

Where we use optical depth, we limit the collocations to where the products agree on CTP within 150 hPa. For both imagers this removes the overall number of collocations by 16% . The optical depth and location of these collocations that are temporarily removed are shown in Figure 2.3. Clouds with optical depths less than 0.5 are primarily affected with most of these removed profiles occurring in the tropics. Less than 2.5% of clouds with an optical depth near 1 are removed due to this requirement.

The performance of each NN is evaluated on our testing dataset (Figure 2.4). 99% confidence intervals for mean absolute error (MAE) and bias (Figure 2.4.a and Figure 2.4.b) are formed by 1,000 bootstrapped samples of our testing dataset. A two-sided t-test indicates significant differences (p-values less than 0.001) between the NN and ACHA at all levels and COD ranges, except the difference between 1000 hPa and 950 hPa at COD values between 3 and 30. MAE is, as expected, larger for clouds with low COD. ACHA appears to struggle with CTP estimation at the mid-levels between 700 and 500 hPa. The MAE for the entire testing dataset is 58.1 hPa for the NN and 109.3 hPa for ACHA. The NN shows statistically significant improvement in most regions especially at the middle- and high-latitudes (Figure 2.4.d,e).

Both approaches have issues with biases with respect to CALIOP in their predictions of CTP (Figure 2.4.b). The NN systematically fails to predict extreme values of CTP near the surface and places them too high in the atmosphere. The opposite problem occurs at the upper-levels, but is less exaggerated for clouds with high COD. Low COD clouds are most affected, with large positive biases above the 700 hPa level. ACHA has similar, but more extreme behavior, for clouds with low COD. ACHA has different signed biases as a function of CTP for clouds with COD greater than 1. This results in ACHA placing these

clouds between 600 and 900 hPa too low in the atmosphere, and clouds with COD greater than 3 between 600 and 300 hPa too high in the atmosphere. This results in a tendency for ACHA to predict a lower frequency of clouds in the mid-levels and could be a contributor to the larger MAE at these levels. In terms of location, the bias patterns are similar between the NN and ACHA (Figure 2.4.f,g), with a negative bias at the low-latitudes and a positive bias at higher latitudes, but ACHA's mean bias is typically of larger magnitude.

A similar analysis is done for the ABI NN (Figure 2.5). The evaluation of the ABI NN shares many characteristics with that of the VIIRS NN. One difference is that optically thin clouds at the highest levels (<150 hPa) have larger errors compared to the VIIRS NN. Similar issues with the biases occur for ABI with a larger positive bias for optically thin clouds at the upper-levels. The spatial patterns of MAE are similar to those of VIIRS. The spatial pattern of the mean bias differs greatly, as the ABI NN typically has a positive mean bias regardless of location. The MAE for all ABI and CALIOP collocations is 61.6 hPa. While this is similar to VIIRS, the two MAE values are not directly comparable due to the differences in the areas and meteorological conditions viewed.

A comparison between the ABI NN and ACHA is not performed due to the computational expense of running ACHA for the large number of ABI images used. However, we feel that the VIIRS NN/ACHA comparisons above and past evaluations of neural network CTP models (Håkansson et al., 2018; Kox et al., 2014) sufficiently justifies an exploration into their interpretability.

Prediction Uncertainty Assessment

The application of quantile regression for obtaining uncertainty information in NN CTP estimation has been evaluated by previous work (Pfreundschuh et al., 2018). We perform a very brief assessment of the calibration of the predicted distributions from the estimated quantiles for the VIIRS and ABI NNs to ensure we can achieve reasonably similar results. To construct cumulative distribution functions (CDFs) we also use the approach by Pfreundschuh et al. (2018) and extend the first and last quantiles to 0 and 1 using a piecewise linear interpolation.

Figures 2.6.a and d show a probability integral transform (PIT; Dawid, 1984) that indicates the frequency of observations as a function of the predicted value of the CDF. The VIIRS NN appears to estimate CDFs that are too narrow evidenced by the higher frequencies at the tails and the lower frequencies at the middle of the CDFs. The ABI NN appears well calibrated with only small differences in observed frequencies throughout. Figures 2.6.b and 2.6.e show similar information presented differently and again confirm that the VIIRS NN predicts distributions that are slightly narrow, but the ABI distributions accurately capture the range of observed values. Figures 2.6.c and 2.6.f illustrate how the width of the predicted distribution (illustrated by the standard deviation of predicted quantiles) corresponds to a wider range of errors observed when comparing to CALIOP. Altogether, Figure 2.6 shows that the predicted distributions from each of these neural networks are typically well calibrated and correspond to the observed errors in an intuitive manner.

2.4 Interpretability Assessment

Many interpretation methods for ML models have been proposed. Some of these approaches offer the ability to provide local explanations, which attempt to describe how individual features contribute to a single model prediction. This is in contrast to global explanations, which are computed over a set of predictions. We attempt to give different perspectives on feature importance using several methods for explaining NN CTP models.

A key challenge is that many of the features used in these models are correlated with one another. This issue in statistical models is often referred to as multicollinearity and affects several aspects of model development and interpretation (Alin, 2010; Farrar and Glauber, 1967; Dormann et al., 2013). Collinear features contribute to increases in the variance of model parameter estimates (Alin, 2010; Daoud, 2017). They also hamper interpretability (Wheeler and Tiefelsdorf, 2005) since feature importance is often shared between collinear features which can lead to misleading conclusions about their overall ranking relative to other features. Thus, a difficulty we struggle with throughout this analysis is whether a feature is deemed important because it has physical significance related to the task or whether it is correlated with another feature that does. Due to the variance in model parameter estimates as a result of multicollinearity, the following metrics are computed over five models with randomly initialized weights. In our case, these models have negligible differences in overall performance (within 1.5 hPa MAE), but the exact dependencies on particular features can be different.

Sequential Backward Selection

Sequential backward selection (SBS) is commonly used to find reduced feature sets with minimal reduction in model performance. The approach starts with selecting a single feature, retraining the model without the feature, and recording the reduction in model performance. This is done for all features, and the feature that yields the smallest decrease in performance is removed. This process is repeated until the number of desired features is reached. SBS can also be used to understand which feature has the most unique and useful information for the task a model is trained for. A large increase in MAE after a feature is removed implies that the feature has unique information relevant for CTP estimation that the NN was not able to find in other features. A low increase in MAE after a feature is removed could imply that the feature is not useful for CTP estimation in the NN, or that the useful information the feature contained was not unique to the feature and could be obtained from others.

In order to isolate the value of a given channel's information, we perform a full SBS, iterating over conceptually linked groups of features associated with each channel (Figure 2.7 and Figure 2.8). Removing groups of features allows us to determine feature importance as a function of spectral band. Otherwise, a feature's relevant information for the estimation of CTP could also be found in other features from the same channel. This also allows us to quantify the contribution of NWP model output to the NN's performance. Satellite estimates are often useful because they are based on observations (compared to an NWP model forecast). Thus, quantifying the contribution of NWP and comparing it to that of the actual observations could be one way of determining how useful or reliable a given estimate of CTP is, in addition to the uncertainty estimates provided by the neural network.

However, it is worth noting that the NWP group contains a larger number of features than groups associated with each channel.

The feature group SBS analysis shows many intuitive characteristics of CTP estimation. For VIIRS (Figure 2.7), these results imply that the 8.6-μm channel is the most important channel followed by the 10.8-μm, and the 12.0-μm. The 8.6-μm channel, in conjunction with window channels such as 10.8-μm, could be used to identify cloud phase (Strabala et al., 1994) and place a cloud in the upper or lower portion of the troposphere. The most useful fusion channels appear to be the 6.7-μm and 7.3-μm which contain information about water vapor absorption and might be useful for placing a cloud above or below the bulk of the water vapor in a given scene. The low increases in MAE of CO_2 fusion channels (13.3-μm through 14.2-μm) are surprising given that CO_2-slicing has proven a useful approach for CTP estimation.

In most cases, a model's reliance on an individual channel increases when the number of channels decrease. However, there are a few exceptions to this generalization for VIIRS, including the impact of removing information from 8.6-μm and 10.8-μm channels once the 13.9-μm and 14.2-μm channels are removed (Figure 2.7 rounds three to five). NWP information ranks highly in the first few rounds, and as channels are removed, we see an increasing reliance on NWP information. This indicates that the usage of NWP information changes as a function of the channels included in an NN.

The same analysis indicates a few similarities for ABI (Figure 2.8). The 8.4-μm, 10.3-μm, and 12.3-μm all have relatively large increase in MAE when tested in the first several rounds. Unlike VIIRS, the 13.3-μm channel ranks fairly highly. On ABI, the 11.2-μm, ozone channel (9.6-μm) and strongly absorbing water vapor channels (6.2-μm through 7.3-μm) do

not benefit the model strictly in terms of MAE.

In the first round, NWP information appears to be more essential for accurate predictions from ABI compared to VIIRS. This impact becomes more similar after the CO_2 channels (with wavelengths 13.3-μm and above) are removed from the VIIRS models. VIIRS appears to rely more heavily on information from the 8.6-μm channel. The impact of removing the lone CO_2 channel on ABI (13.3-μm) is different than the impact of removing the four fusion CO_2 channels for VIIRS. For instance, in round 7 after the 13.3-μm is removed, the MAE from removing the 12.3-μm or 8.4-μm is nearly tripled on ABI. After removing the fusion CO_2 channels from VIIRS, it is mostly the impact of NWP information which is increased.

Figures 2.7 and 2.8 make it clear that similar performance can be achieved for these CTP models with reduced feature sets. Ignoring differences in reliance on NWP information, similar models could be created using the feature sets after round 5 for both instruments. We continue to use the full feature set to keep the latter experiments consistent.

Neural Network Interpretation Methods

Next, we attempt to characterize these models using approaches specific to NNs that offer local explanations. Both local explanation methods we describe below are relatively complex compared to backward selection. We attempt to provide a concise description of how these approaches work in general terms, but if a detailed explanation is desired we refer the reader to their corresponding references.

The first method used is layer-wise relevance propagation (LRP; Bach et al. 2015). LRP is a popular method for model attribution and has been used to interpret models in applications such as radar reflectivity estimation from satellite imagers (Hilburn et al., 2021)

and for detecting common change patterns among climate models (Barnes et al., 2020). LRP can be generally described as computing a backward pass through an NN, starting with the activations at the last layer. A prediction score is propagated backward through each layer of the model and projected onto the dimensions of the original input at the first layer. There are several different propagation rules that dictate how the prediction score is distributed to the units of each layer. In our application, we use the epsilon rule for all layers, which adds a small positive value to the denominator of the relevance propagation rule in order to improve numerical stability.

The second method we use is Shapley additive explanations (SHAP; Lundberg and Lee 2017), based on the Shapley value from cooperative game theory (Shapley, 1953). Similar to the relevance from LRP, Shapley values attempt to attribute responsibility to features for a given prediction. In the original SHAP paper, a model-agnostic approximation of Shapley values is introduced, called kernel SHAP. However, this approach ignores information available in the structure of the neural network that could be useful for improving computational performance. The same work introduces a NN specific approach, deep SHAP, that leverages principles from DeepLIFT (Shrikumar et al., 2017). Specifically, deep SHAP takes advantage of the per-node attribution rules used in DeepLIFT.

Both the LRP and SHAP can produce signed attributions, according to whether an input feature acted to increase or decrease the prediction. In itself, interpreting the output from local explanation methods can be a difficult task. We attempt to simplify this by standardizing the local explanations. We take the absolute value of the LRP relevance and SHAP values and express them relative to the feature with the greatest value for each input example. For each prediction, each feature has relative importance ranging from 0, which

implies it was not important, to 1, which implies the feature was the most important or tied with the most important.

Figures 2.9 and 2.10 both show the global relative feature importance calculated over conceptually-linked groups of features. These values are calculated by summing the absolute value of the LRP and SHAP attributions for each group of features and dividing by the value from the largest group. The 8.4-μm (ABI) and 8.6-μm (VIIRS) channels are suggested to be the most important channels for CTP estimation. LRP and SHAP assign low relative feature importance to the ozone channels around 9.7-μm on both instruments and the 6.7-μm and 7.3-μm channels on VIIRS. Both methods rank spatial information lower than spectral information for both imagers. Both methods also rank spatial metrics from fusion channels lower than those from native features, despite there being more-than-double the number of features from the fusion channels.

Despite the agreement on a some broad points, there are a few differences between LRP and SHAP. In general, the relative feature importance values from LRP are more distributed across features compared to SHAP, which gives sparser explanations that emphasize the most important features. The high rankings of the 8.4-μm and 8.6-μm channels from SHAP imply that most explanations are dominated by these channels. Both methods agree on the relative ranking of most features, with the significant exception of NWP data for both sensors, where LRP reports values more than twice that of SHAP. SHAP's attribution here also contrasts with the backward selection results which imply that NWP information is very useful for CTP estimation.

There are several other differences between what information these methods suggest that the NNs use compared to backward selection. LRP reports that the fusion channels, ignoring

spatial metrics, have a roughly similar value of relative feature importance as the VIIRS native channels. However, when we remove all fusion channels from the VIIRS models, MAE only increases by less than 5 hPa, which is less than removing information from the 8.6-μm or 10.8-μm channels. Of all fusion channels, the 6.7-μm and 7.3-μm channels appear to the have the largest impact when removed during backward selection but are ranked fairly low when compared to the 13.3-μm, which both LRP and SHAP rank as the most important fusion channel.

Local Explanation Clustering

Next, we explore the local explanations for these models. We attempt to find conceptually similar explanations among the local attributions. We then analyze these explanations as a function of their dominant features, CTP, cloud-top phase, opacity, and location. We find these explanations by using a K-means clustering (from scikit-learn version 0.24; Pedregosa et al. 2011) on the local attributions. Thus, each imager-CALIOP collocation belongs to a specific cluster. For concision, we only perform the following analysis using LRP. We specify four clusters, but do not conclude that it is the optimal number, nor that there are discrete clusters at all. We use the clustering to partition the local explanations into more homogeneous groups and visualize differences among them. The motivation for this analysis is to help understand the kinds of relationships that might be used in the neural networks in predicting CTP for different types of clouds. Figure 2.11 shows the clustering analysis performed for VIIRS and describes each cluster in terms of the the relative feature importance for predictions that belong to each cluster. Figure 2.11 also illustrates the fraction of all collocations belonging to each cluster in terms of CTP, cloud-top phase, opacity, and

location.

VIIRS Cluster 1 has high feature importance in the 8.6-μm channel where it was the leading feature in most predictions. Cluster 1 also had high feature importance in the 10.8-μm channel and low importance in fusion channels. It represents a common explanation across all locations and favors optically thin clouds at all levels regardless of cloud phase. VIIRS Cluster 2 has the highest feature importance in the 12.0-μm, 6.7-μm and 7.3-μm channels, and a high importance in fusion CO_2 channels. It primarily represents upper-level liquid clouds and upper-level opaque ice clouds. Cluster 2 is globally distributed, but is not often found in areas dominated by lower- and middle-level cloudiness. VIIRS Cluster 3 has high feature importance in the 13.3-μm channel, spatial metrics from the 12.0-μm channel, and NWP surface temperature. It is the dominant explanation for lower-level liquid clouds and explains a large fraction of clouds occurring off the western coast of South America, the southwestern coast of Africa, and regions where persistent low-level cloudiness is common. VIIRS Cluster 4 has high feature importance for spatial metrics, NWP information, and moderate values for the fusion CO_2 channels. It is common in many locations, but is frequent over the Southern Ocean. Given the dependence on spatial metrics, and lack of a clear relationship with cloud properties, we expect that Cluster 4 might primarily represent cloud edges where the spatial metrics will take on particularly large values (see Section 4.2.4)

Figure 2.12 illustrates a similar analysis for ABI. Overall, the clusters appear to be less sensitive to opacity and more sensitive to cloud-top pressure. Some similar patterns exist in the spatial distribution of the clusters when comparing ABI to VIIRS.

ABI Cluster 1 shows importance in channels with water vapor absorption (6.2-μm, 6.9-μm and 7.2-μm), many spatial metrics, and NWP information. Cluster 1 primarily

represents clouds at all levels but slightly favors low-level opaque ice clouds. ABI Cluster 2 explains a large fraction of low-level liquid clouds and relies heavily on the 12.3-μm channel where it is the leading feature for over 90% of examples. ABI Cluster 2 is frequent in areas with low-level cloud cover. ABI Cluster 3 has the largest feature importance in the 8.4-μm channel, where it is frequently the leading feature. It is present at various levels, slightly favoring optically thin liquid clouds and mostly occurs at the tropics. ABI Cluster 4 has high feature importance in the 6.2-μm, 6.9-μm, 7.3-μm and 8.4-μm channels and an otherwise low importance in spatial metrics and NWP data aside from 300 hPa temperatures. It describes the vast majority of predictions for ice clouds between 600 hPa and 200 hPa and is primarily located at the high latitudes.

There are a few loose similarities between the clusters identified in the local explanations of both the ABI and VIIRS models. One such similarity is between VIIRS Cluster 3 and ABI Cluster 2. Both of these groups show at least moderate importance in the 12.0-μm (VIIRS), and 12.3-μm (ABI) channels and explain a large proportion of low-level water clouds in similar locations. Both models also have one cluster associated with feature importance in spatial information (ABI Cluster 1 and VIIRS Cluster 4) that occurs somewhat frequently in the high latitudes and more moderately in the tropics. Another loose similarity can be found between VIIRS Cluster 1 and ABI Cluster 3, which have high importance in the 8.4-μm and 8.6-μm channels but have very different spatial distributions.

Local Explanation Example

In an effort to contextualize the LRP attributions and illustrate potential relationships between VIIRS Cluster 4 and cloud edges, we calculate relative feature importance from

LRP for an example VIIRS scene centered over -55°S, 100°E (Figure 2.13). The LRP values are standardized in the same way and are reported as a function of the same conceptual groups in Figure 2.9.a. The neural network is not capable of cloud detection, so predictions are provided in all pixels regardless of whether there is a cloud present.

Shown in Figure 2.13.a and 2.13.b is 10.8-μm channel from VIIRS and the predictions of CTP made by the NN. The width of the predicted distribution can be large near edges of clouds that have high contrast with the surface (Figure 2.13.c). The predicted distributions are also wider where upper-level clouds overlap with mid-level clouds (Figure 2.13.c, lower left). In this scene, NWP information is most important for middle- and lower-level clouds. Native spectral observations (Figure 2.13.e) are most important for upper and lower-level clouds, but there is a strong decrease in the importance of native spectral observations near clouds edges (Figure 2.13.e, right). These low values of the relative feature importance of native spectral observations near cloud edges correspond to large importance of the spatial metrics from native VIIRS observations (Figure 2.13.g). The relative feature importance for spectral fusion feature group is largest for lower- and middle-level clouds, and has more moderate values for upper-level clouds. The feature importance of the spatial metrics from fusion channels (Figure 2.13.h) appear to have the lowest values in this scene overall and only have moderate impact for more spatially uniform low-level clouds and very low importance for upper-level clouds.

We see a few potential explanations for the importance of spatial metrics around cloud edges. At cloud edges, spectral features can difficult to interpret due to the possibility of a pixel being only partially cloudy and the resulting brightness temperature being a mixture of a cloud and an unobscured view of the surface. Spatial metrics such as the

difference between the central pixel and the 5-by-5 pixel maximum and minimum could provide information on the brightness temperature of a nearby fully clear pixel and a nearby fully cloudy pixel. A second explanation could be that this is an artifact of the way our dataset is collected. CALIOP observations are not typically made at the exact same time as VIIRS. This time difference might allow for the cloud observed by CALIOP move outside the view of collocated imager pixel. Spatial metrics might indicate where this is likely and this behavior could alternatively be symptom of training to match an imperfect label.

2.5 Discussion

The efforts to interpret the models in this work give slightly different perspectives on feature importance for NN CTP estimation. In many cases, this is expected when comparing SBS to LRP and SHAP. Ultimately, SBS describes a different set of models, and observing a small increase in error when removing a set of features does not necessarily imply that they are not useful for CTP estimation. It instead might be an indicator that the information is not unique to a feature. This may be the case for the fusion CO_2 channels in the VIIRS model. When removed through backward selection (Figure 2.7), they yield only small increases in model error; however, the LRP and SHAP attribute a moderate amount of feature importance to them. This indicates that while fusion CO_2 channels may be useful for CTP estimation, similar performance, in terms of MAE, can be attained without them. Other differences are less easily explained, such as the differences in importance of NWP information between LRP and SHAP. LRP assigns a large relative FI to NWP information and is in agreement with the first round of the SBS feature group analysis for both instruments, indicating that

this might be a failure of SHAP's attribution.

A few potential sources of the differences between LRP and SHAP could be rooted in our choice of model architecture, the nature of our prediction task, and choice of LRP rules. LRP was initially developed to explain the output of convolutional neural networks trained for image classification (Bach et al., 2015). It is unclear how well these attributions generalize to regression problems. Similarly, during development we noticed slight differences in attributions depending on the exact LRP propagation rules used (Montavon et al., 2019), but qualitatively similar takeaways overall (not shown).

Despite discrepancies in the importance of a few features, there is still some agreement between approaches. All approaches agree that the 8.4-μm and 8.6-μm channels are useful in the estimation of CTP. Similar agreement between methods is found for the importance of the 10.8-μm and 11.2-μm channels for VIIRS and the 12.3-μm channel for ABI. Intuitively, all approaches place a much greater emphasis on the brightness temperatures, which have a more direct physical relationship to cloud-top pressure compared to spatial metrics.

Several methods also agree on the relative unimportance of particular features. These include the ozone channels from both instruments, the 6.2-μm ABI band, which is sensitive to upper tropospheric water vapor, and the spatial metrics calculated from the VIIRS/CrIS fusion channels. In this case it is helpful to remember that the fusion channels are derived from the relatively coarse CrIS observations and interpolated using infrared channels from VIIRS. It is plausible that fine-scale spatial variability on the scale of 3.75 km (the edge length of 5 VIIRS pixels at nadir) is not well-represented.

Regardless of disagreement between LRP and SHAP, we can conclude that the VIIRS/CrIS fusion channels only have small benefit when included in the VIIRS NN since

they increased MAE only by roughly 5 hPa (Figure 2.7) when all are removed. However, several fusion channels indicated no benefit after removal (Figure 2.7 rounds 1 through 5).

One point made earlier in this paper is that removing these channels had the effect of increasing the reliance on NWP information (Figures 2.7 and 2.8). The SBS analysis shows that the fusion CO_2 channels do not substantially reduce model error when included. However, their inclusion is suggested to reduce the reliance on NWP information. This is an important point since it can change how much a given CTP prediction depends on observations, compared to ancillary information from an NWP model forecast. When included in climate records, changing the source of the ancillary NWP data can yield small but meaningful changes in the variability of cloud properties estimated from imagers (Foster et al., 2016). Thus, the physical interpretation CTP estimates can change depending on the reliance on NWP information.

Other difficulties include directly comparing results from the VIIRS and ABI CTP models. Even though spatial metrics are both computed over 5-by-5 pixel arrays, these metrics have different meanings for each sensor. This is due to differing spatial resolutions between sensors and the fact that the spatial resolution of VIIRS varies less at higher viewing angles, due to the aggregation of pixels at lower viewing angles. The spatial resolution of ABI varies much more considerably. Thus, the physical meaning of these metrics is likely quite different between the two instruments.

It is interesting to compare results of this analysis to the physical information exploited in approaches like CO_2-slicing. CO_2 channels do not seem to add much value to a model that already has access to infrared window channels including a channel around 8.6-μm. There is more value in including channels with lower- and middle-level water vapor absorption,

such as the 6.7-μm and 7.3-μm. However, it is not clear if this observation holds for imagers, such as MODIS, where these measurements are made natively. Despite this caveat, most of the feature importance metrics used in this analysis imply that not exploiting variability of the 8.6-μm or infrared window channels between 10-μm and 12-μm will yield a suboptimal result.

The LRP clustering analysis suggests that these models have the capacity to handle CTP predictions for certain types of clouds differently. This is represented by how identified clusters vary with cloud-top phase, opacity, location, and the features used to make a particular prediction. This is an intuitive result, since knowledge of cloud-phase may narrow the range of plausible CTP values. Similarly, knowledge of the opacity of a cloud may inform the NN about the contribution to top-of-atmosphere brightness temperatures from sources below the cloud.

Throughout this work, we note substantial variability in model explanations between sensors and minor differences between random initializations. We stress that even if two CTP NNs for different sensors are trained to match observations from CALIOP, it is unlikely that their local explanations are similar. Our results might not be applicable to other imager-lidar pairings. This has implications for transitioning ML-based approaches to climate records made up of multiple sensors such as VIIRS and MODIS, in which it may be desirable for models for each sensor to have similar explanations in addition to similar predictions for a given example. We suspect that some differences in this analysis come from the fact that VIIRS views a wider range of meteorological conditions. Additionally, our ABI/CALIOP testing dataset is only valid for the last three months of the year, and our VIIRS/CALIOP collocations are collected over an entire year.

Despite the wealth of information provided by the interpretability methods used in this analysis, many questions about how particular features are used in CTP models remain unanswered. For example, why are the 11.2-μm and 12.3-μm channels favored over the 10.3-μm channel on ABI which has less water vapor absorption? Similar questions can be asked about why spatial metrics from one channel might be favored over others or why upper-level water vapor absorption is relatively unimportant for ABI CTP estimation. This analysis gives us an overall idea about which features are useful for an NN, but the task of model interpretation is now shifted to attributing physical significance to these results. Difficulties in attributing physical significance are enhanced by the fact that there is disagreement between interpretability approaches. This motivates future work in verifying local explanations, such as the comparison in Mamalakis et al. (2021), where ground-truth explanations are available.

2.6 Conclusions

We characterize the use of individual channels of LEO and GEO imagers for NN CTP estimation. We first perform a short comparison between our NNs and an operational approach which demonstrates large improvement in CTP estimation with respect to CALIOP. We then use backward selection, LRP, and SHAP to infer the relative importance of features. We find many instances of disagreement between these different perspectives on feature importance, but broad agreement on the importance of a few channels, including the VIIRS 8.4-μm channel (8.6-μm for ABI) and other infrared window channels around 10-μm to 12-μm. We also observe a small benefit in including absorption channels that are sensitive

to middle-level and lower-level water vapor. VIIRS/CrIS fusion CO_2 channels and spatial metrics derived from them appear add little-to-no improvement to CTP models where native infrared channels are already present, but have impact on the reliance of a given model on NWP model output. Clustering local explanations from LRP illustrates how NN models can exploit variability related to CTP, phase, and opacity from infrared measurements. The LRP clustering also suggests, intuitively, that the NNs use different infrared channel combinations for estimating CTP of the different cloud types. While this analysis reveals several interesting aspects of the relative importance of infrared channels for CTP estimation, this work illustrates the immense challenge of attributing physical significance to both global and local explanations for neural networks.

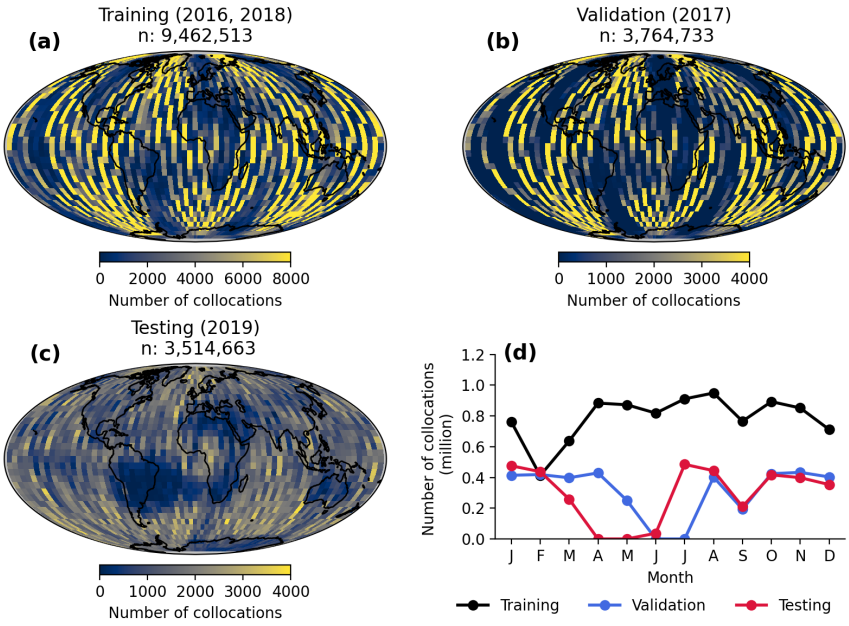

Figure 2.1: Shown are the distributions of VIIRS collocations with CALIOP for the training (a), validation (b), and testing (c) datasets. (d) indicates the seasonal distribution of each.

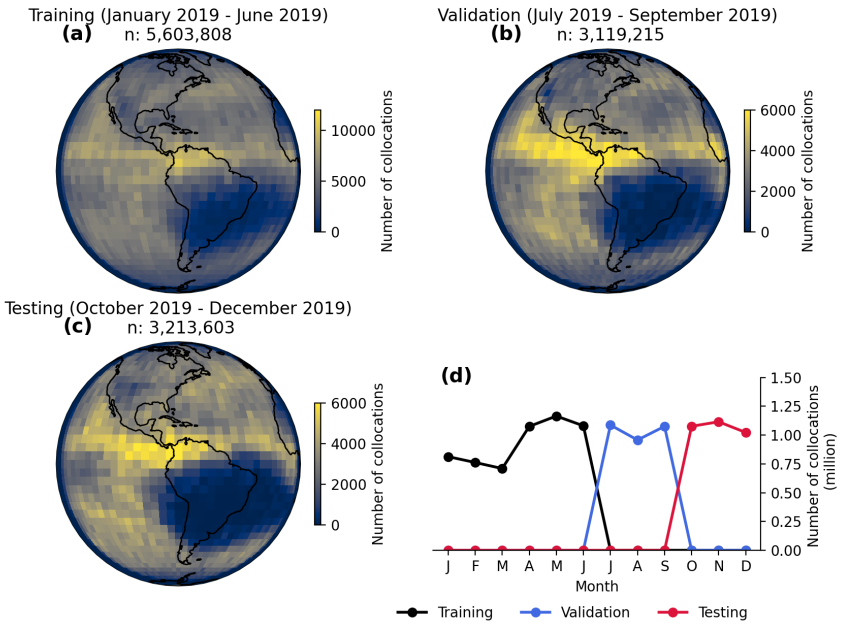

Figure 2.2: Shown are the distributions of ABI collocations with CALIOP for the training (a), validation (b), and testing (c) datasets. Panel (d) indicates the seasonal distribution of each dataset.

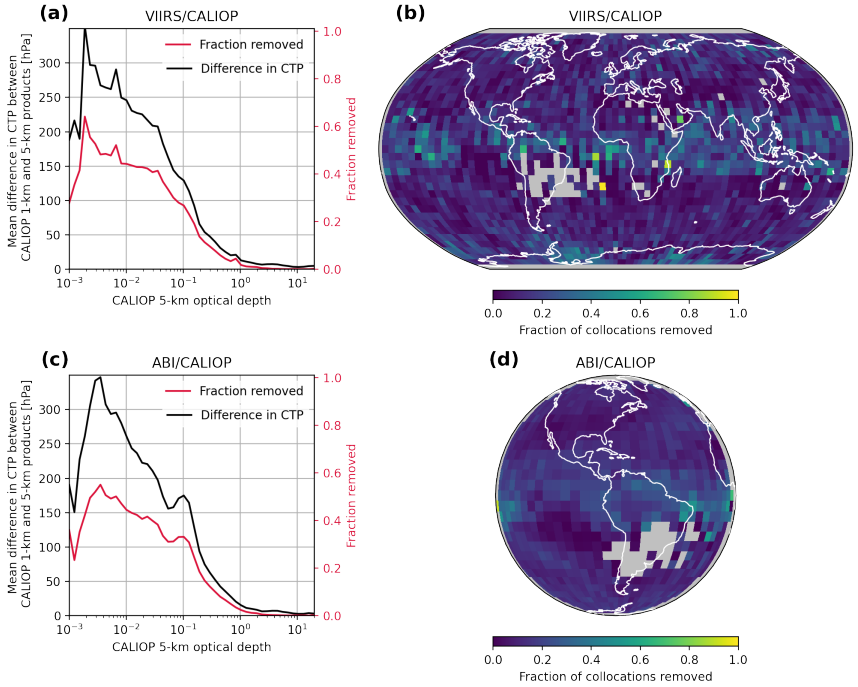

Figure 2.3: (a) and (c) indicate the fraction of collocations that are removed by applying the requirement that the 1-km and 5-km CALIOP products agree within 150 hPa. Also shown is the mean differences between the two products as a function of optical depth. (b) and (d) indicate the fraction of collocations removed on a 5-by-5 degree latitude/longitude grid.

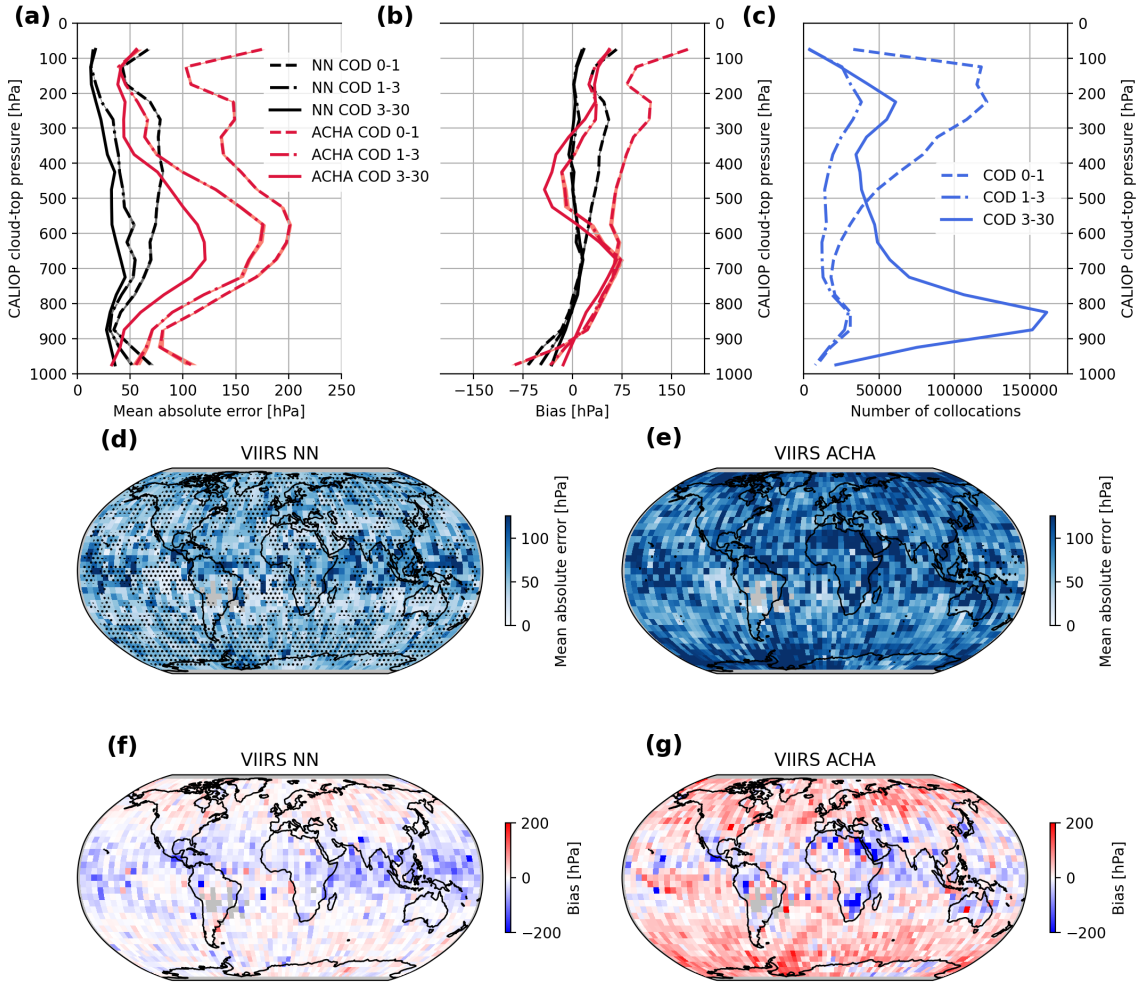

Figure 2.4: (a) shows the mean absolute error of the NN (black) and ACHA (red) for several values of COD. (b) shows the bias of the NN and ACHA compared to CALIOP over the same values of COD. (a) and (b) are shown with 99% confidence intervals in lighter shading, but are often obscured by the mean values due to the narrow intervals. (c) indicates the number of collocations occurring between CALIOP and VIIRS. (d) and (e) indicate the mean absolute error on a 5-by-5 degree latitude/longitude grid. Stippling in (d) and (e) indicate that the respective approach has a statistically significant improvement with a p-value less than 0.001 at the grid point. (f) and (g) indicate the mean bias on the same grid.

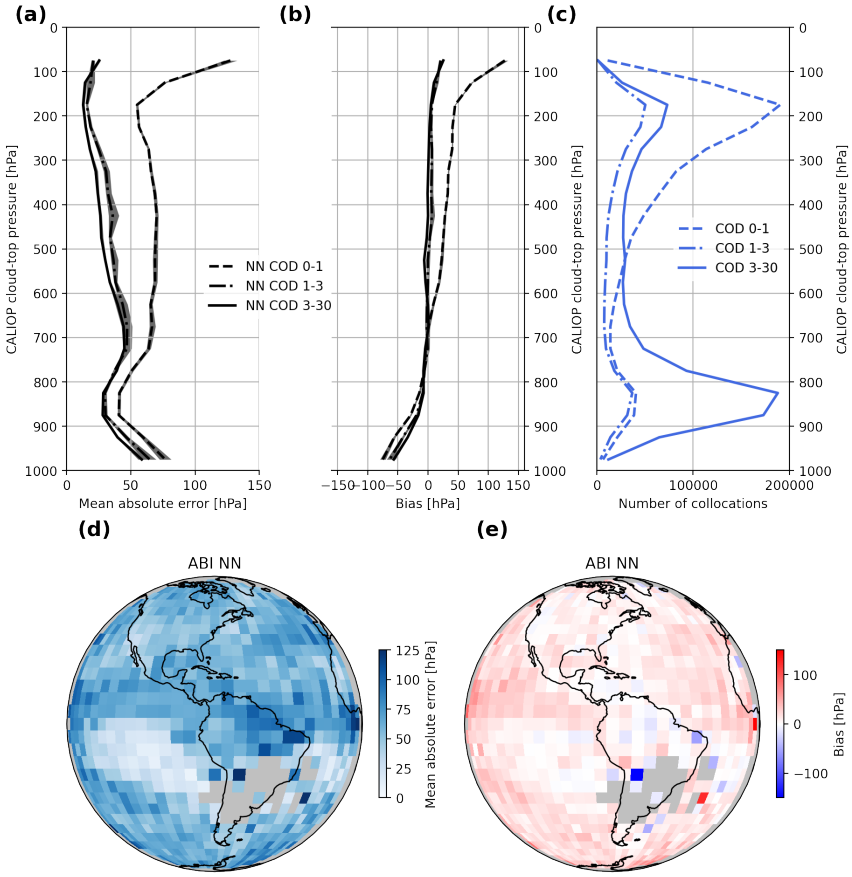

Figure 2.5: (a) shows the MAE for the NN over several ranges of COD. (b) shows the bias over the same ranges of COD. (a) and (b) are shown with 99% confidence intervals in lighter shading, but are often obscured by the mean values due to the narrow intervals. (c) is the number of collocations between ABI and CALIOP. (d) and (e) indicate the MAE and bias of the neural network on a 5-by-5 degree latitude/longitude grid.

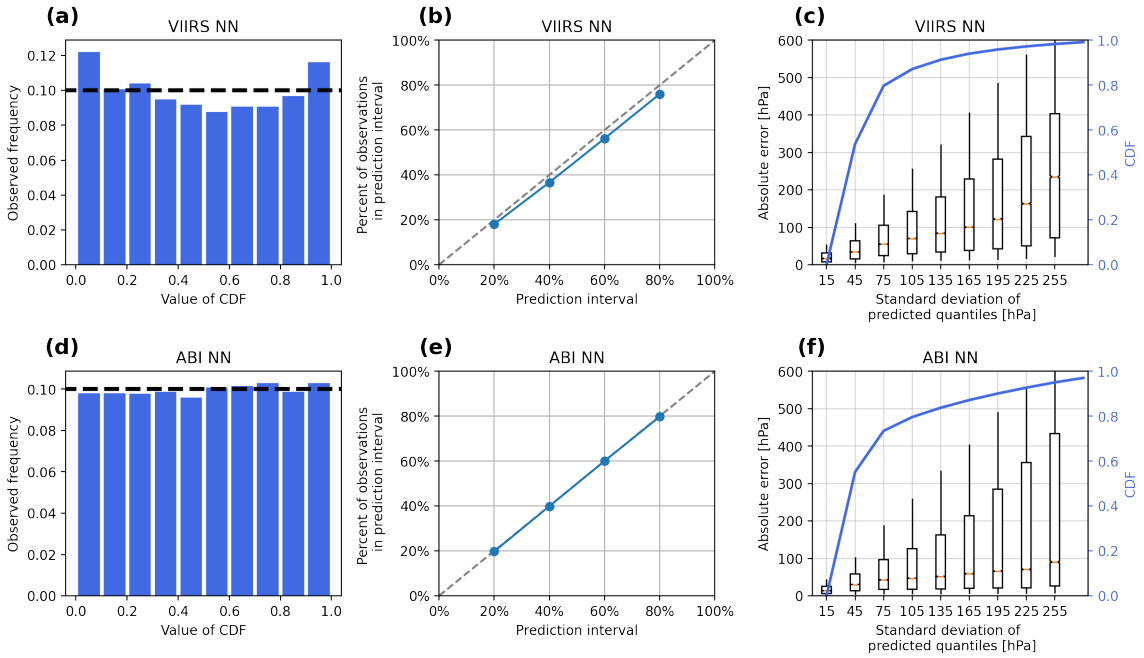

Figure 2.6: (a) and (b) indicate the observed frequency of CALIOP observations as a function of the value of the predicted cumulative distribution function (CDF) from the predicted quantiles. (b) and (e) show the fraction of CALIOP observations that fall within the prediction intervals derived from the predicted quantiles. Dashed lines in (a), (b), (d), and (e) indicate a well calibrated model. (c) and (f) show the distribution of absolute errors with CALIOP as a function of the standard deviation of predicted quantiles. The middle orange line represents the 50th percentile, box edges represent the 30th and 70th percentile, and the whiskers represent the 10th and 90th percentile of absolute error (left x-axis) with respect to CALIOP. The cumulative distribution function is shown in blue and represented on the right x-axis.

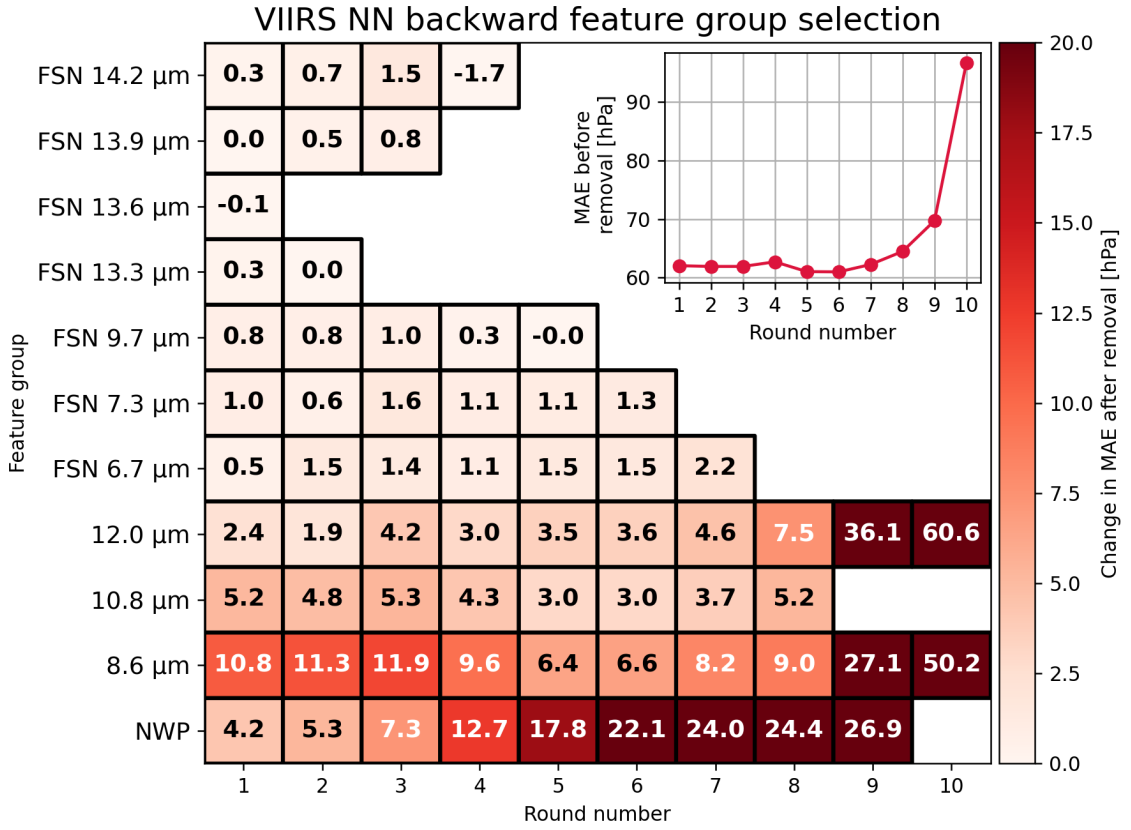

Figure 2.7: Shown are the results of a backward selection performed on features linked to each channel used in the VIIRS CTP models. This figure is most easily interpreted by considering each column from left to right. Each column represents a single round of backward selection. The inset plot shows the MAE of a model that includes all remaining features present in a given column. In each round, a feature's impact is tested by training five identical but randomly initialized models without that feature and recording the MAE. The value in each box represents the mean increase in MAE (of the three best-performing models) relative to a model that includes all features present in a column. Note that the feature group that increases MAE the least in a given round is permanently removed from the model and is no longer tested in the following rounds.

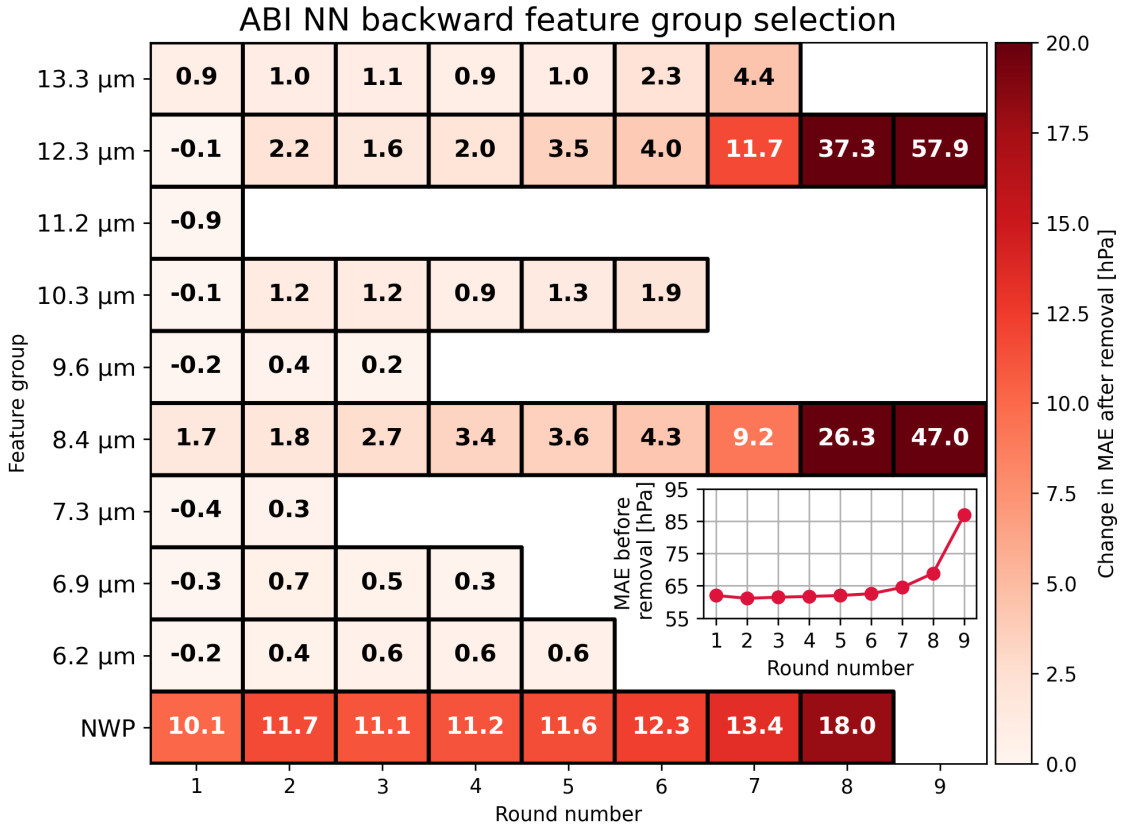

Figure 2.8: Shown are the results of a backward selection performed on features linked to each channel used in the ABI CTP models. This figure is most easily interpreted by considering each column from left to right. Each column represents a single round of backward selection. The inset plot shows the MAE of a model that include all remaining features after each round. In each round, a feature's impact is tested by training five identical but randomly initialized models without that feature and recording the MAE. The value in each box represents the mean increase in MAE (of the three best-performing models) relative to a model that includes all features present in a column. Note that the feature group that increases MAE the least in a given round is permanently removed from the model and is no longer tested in the following rounds.

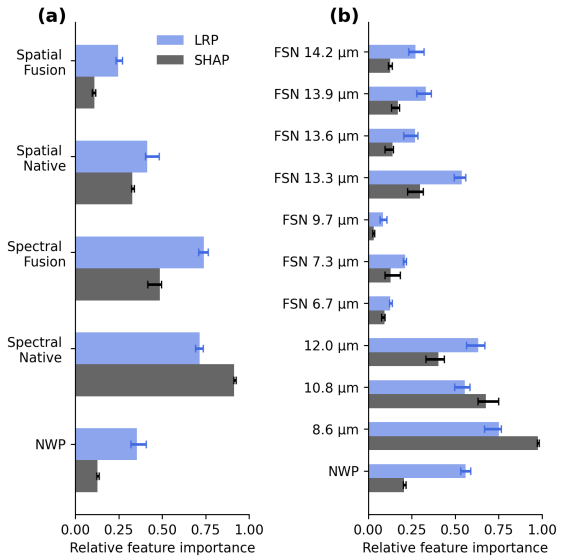

Figure 2.9: Relative feature importance for different groups of features calculated over five VIIRS NN models. (a) separates channel brightness temperatures from their associated spatial metrics and fusion channels from native VIIRS channels. (b) separates features based on their associated channel.

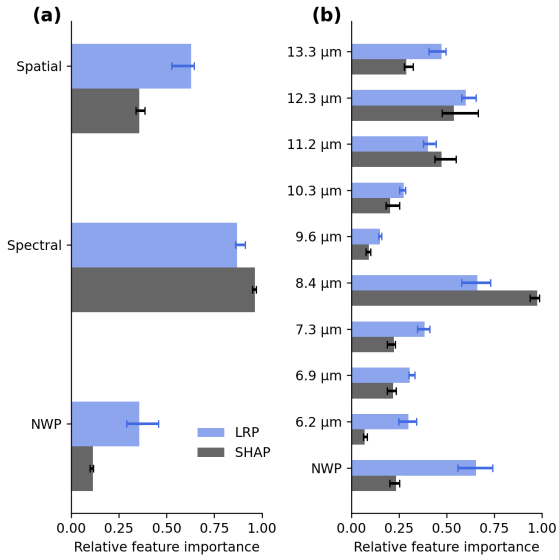

Figure 2.10: Cluster analysis of relative feature importance values calculated from LRP for the ABI CTP models. (a) represents the distribution of feature importance values for each cluster, where the black middle line, box edges, and whiskers represent the 50th, 30th/70th, and 10th/90th quantiles of each feature. (b) and (c) show the distribution of each cluster with respect to CTP and optical depth of the uppermost cloud for ice clouds and liquid clouds respectively. (d), (e), (f), (g) show the spatial distribution of each cluster on a regular 5-degree grid and the proportion of collocations falling within each cluster listed above. Note that the color bars represent slightly different ranges and are chosen to emphasize spatial variability within each cluster. The analysis in this figure is subject to the requirement described in Figure 2.3 due to the use of optical depth.

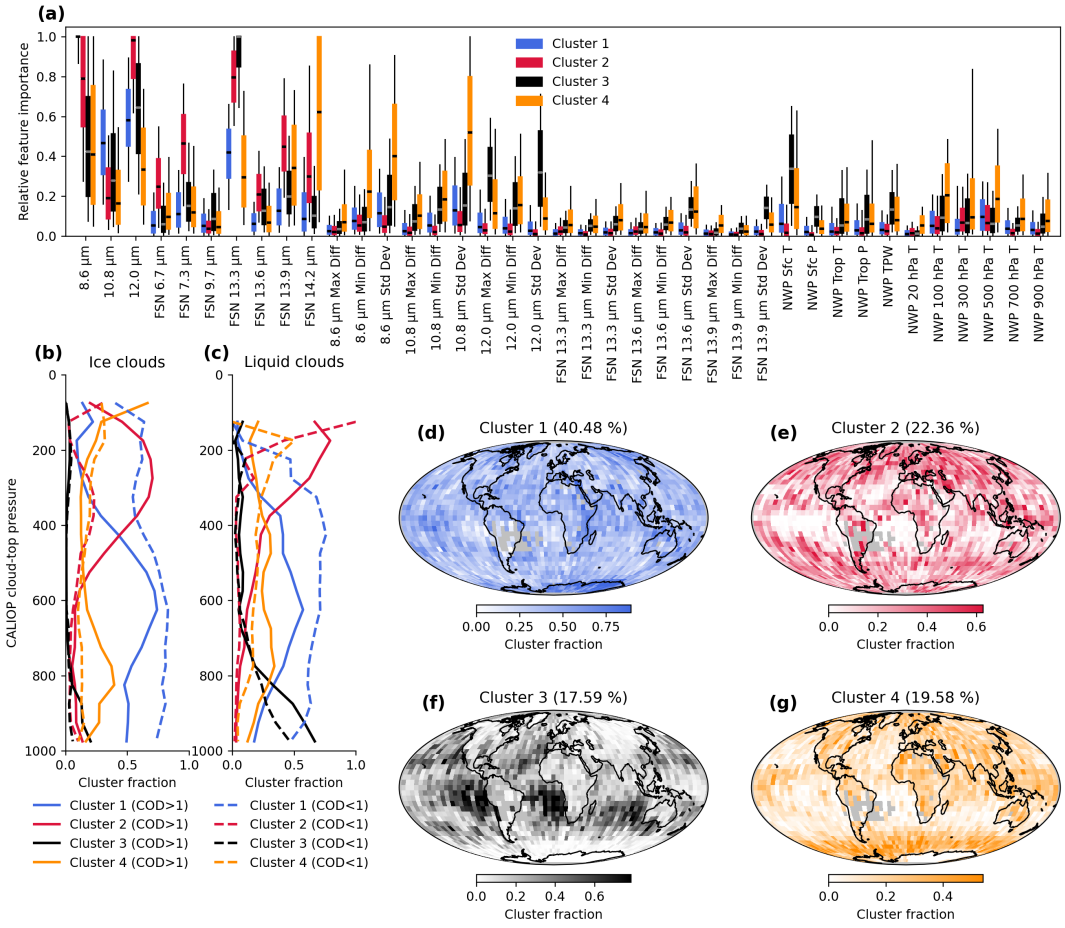

Figure 2.11: Cluster analysis of relative feature importance values calculated from LRP for the VIIRS CTP models. (a) represents the distribution of feature importance values for each cluster where the black middle line, box edges, and whiskers represent the 50th, 30th/70th, and 10th/90th quantiles of each feature. (b) and (c) show the distribution of each cluster with respect to CTP and optical depth of the uppermost cloud for ice clouds and liquid clouds respectively. (d), (e), (f), (g) show the spatial distribution of each cluster on a regular 5-degree grid and the proportion of collocations falling within each cluster listed above. Note that the color bars represent slightly different ranges and are chosen to emphasize spatial variability within each cluster. The analysis in this figure is subject to the requirement described in Figure 2.3 due to the use of optical depth. Some fusion channel spatial metrics are not shown in (a) due to very low values and to ease visualization.

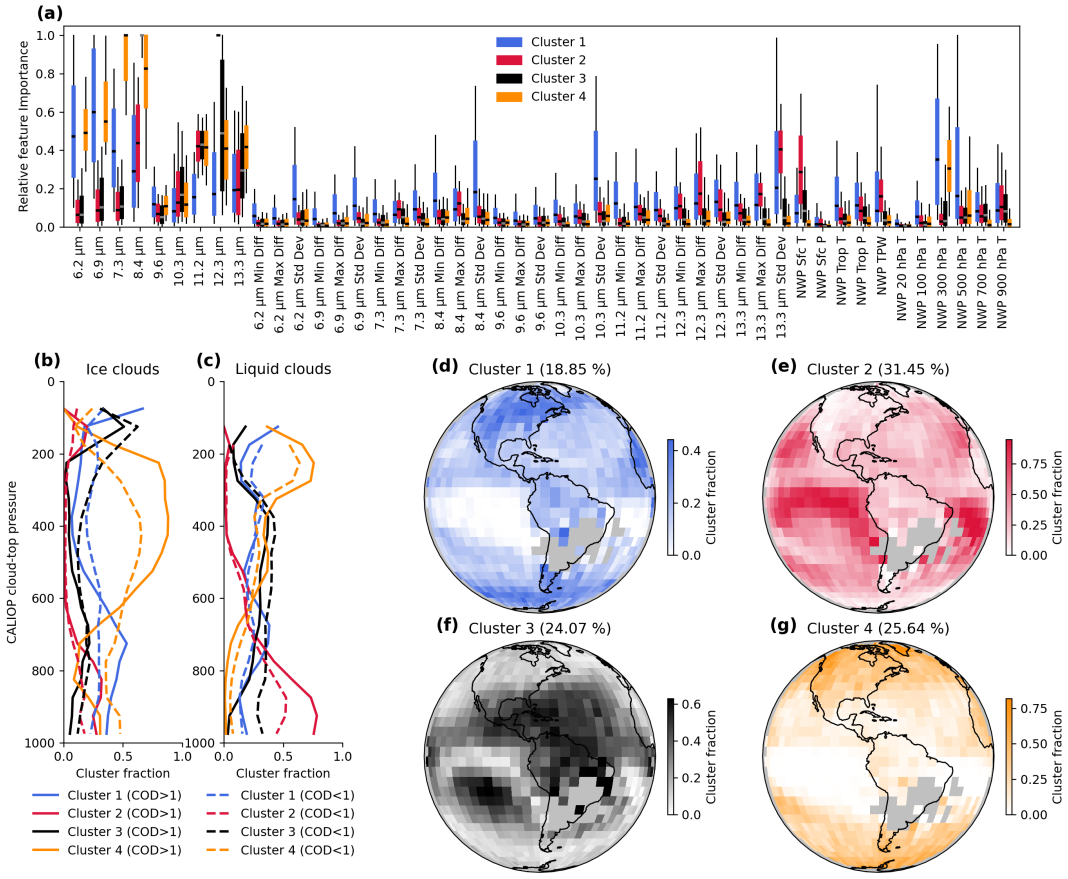

Figure 2.12: Cluster analysis of relative feature importance values calculated from LRP for the ABI CTP models. (a) represents the distribution of feature importance values for each cluster, where the black middle line, box edges, and whiskers represent the 50th, 30th/70th, and 10th/90th quantiles of each feature. (b) and (c) show the distribution of each cluster with respect to CTP and optical depth of the uppermost cloud for ice clouds and liquid clouds respectively. (d), (e), (f), (g) show the spatial distribution of each cluster on a regular 5-degree grid and the proportion of collocations falling within each cluster listed above. Note that the color bars represent slightly different ranges and are chosen to emphasize spatial variability within each cluster. The analysis in this figure is subject to the requirement described in Figure 2.3 due to the use of optical depth.

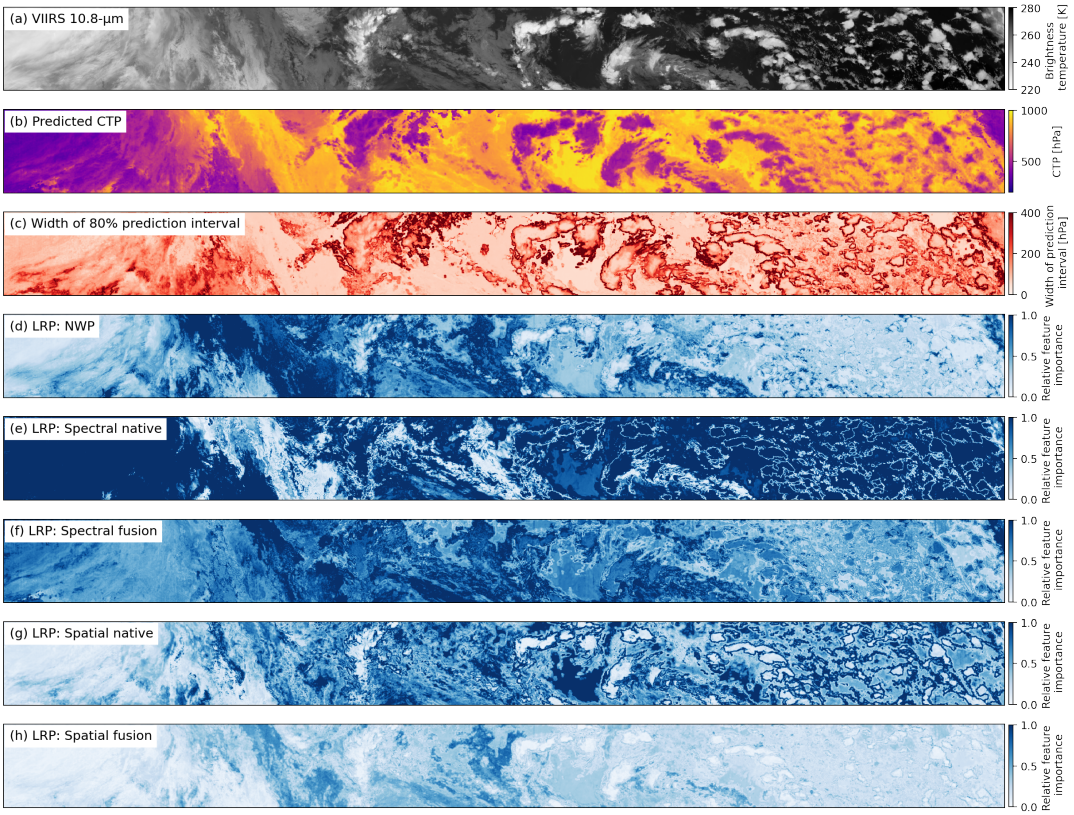

Figure 2.13: Example of CTP predictions for VIIRS scene centered over -55°S, 100°E. (a) is the 10.8-μm infrared channel. (b) is the estimates of CTP from the 50th quantile. (c) is the width of the 80% prediction interval constructed from the 10th and 90th quantiles. (d), (e), (f), (g), and (h) are the LRP relative feature importance for the NWP, spectral native, spectral fusion, spatial native, and spatial fusion groups discussed in the text.

3 OPTIMIZING FOR CONSISTENCY IN NEURAL NETWORK CLOUD PROPERTY RETRIEVALS FOR MULTI-SENSOR SATELLITE RECORDS

3.1 Introduction

Satellite imagers have long been used to estimate properties of clouds using both physical and statistical approaches. Examples include cloud-top height (Inoue, 1985; Chahine, 1974), cloud-base height (Seaman et al., 2017; Noh et al., 2017), cloud optical depth and effective radius (Nakajima and King, 1990), cloud droplet number concentration (Brenguier et al., 2000) and many others. These derived products can be used to study long-term variation in clouds given a sufficiently long observational record. There have been many efforts to create stable observational records from sensors in both low-earth and geostationary orbit to facilitate such studies (Popp et al., 2020). However, intercomparisons have shown varying levels of disagreement in seasonal and long-term variability of cloud properties from imager observations (Stubenrauch et al., 2013; Karlsson and Devasthale, 2018).

The Advanced Very-High Resolution Radiometer (AVHRR) is the longest imager record with a similar sensor design beginning in 1979. Given the need for stable long-term records of cloud properties, there have been several projects aimed at reprocessing AVHRR data. These include the Pathfinder Atmospheres - Extended dataset (PATMOS-x; Heidinger et al., 2014), the Climate Monitoring Satellite Application Facility (CM SAF) Cloud, Albedo And Surface Radiation dataset (CLARA-A2; Karlsson et al. and the Climate Change Initiative Cloud project (Cloud CCI; Stengel et al., 2020). Regarding more recent imagers, there are efforts to create continuous datasets of or assess the differences between the

relatively modern Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) instruments (Uprety et al., 2013; Li et al., 2018; Skakun et al., 2018; Hall et al., 2019; Xiong et al., 2020). Especially relevant to this work, are the NASA MODIS and VIIRS climate data record continuity cloud properties (CLDPROP) (Platnick et al., 2021; Frey et al., 2020). Geostationary (GEO) records include the international satellite cloud climatology project (ISCCP; Schiffer and Rossow 1983) which is a popular dataset made up of relatively coarse resolution observations from multiple GEO imagers. With the recent launch of more capable GEO imagers such as the Advanced Baseline Imager (ABI; Schmit et al. 2017), Advanced Himawari Imager (AHI; Bessho et al. 2016) and the forthcoming Flexible Combined Imager (FCI; Durand et al. 2015) a next-generation version (ISCCP-NG) is in early stages of development (Heidinger et al., 2021).

Constructing a cloud climatology from a single or multiple different instruments comes with many challenges. These can include differences in channel availability, or even slight differences in spectral response functions among channels with similar central wavelengths (Meyer et al., 2020). Sensor calibration can drift over time for a single sensor, or differ from another sensor of the same design requiring adjustments (Heidinger et al., 2010; Bhatt et al., 2016). Spatial resolution can play a role due to the detection of features at fine scales that may go unresolved by others (Frey et al., 2020). Sensors among a single record could have different orbits making intercomparisons difficult (e.g. Aqua and Terra MODIS). Many satellites carrying AVHRR drift in their orbits causing the local observation time to change (e.g. NOAA-14, and NOAA-15). The specific orbit and time period covered by a satellite mission also affects what sources can be used to validate derived products, and the amount

of spatiotemporal overlap between other sensors in which consistency is needed. All of these considerations contribute to the difficult task of creating reliable long-term records of derived products from measurements made by multiple sensors.

Machine learning (ML) has become a popular tool for estimating characteristics of clouds and climate (Beucler et al., 2021). Several ML approaches have used the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP; Winker et al. 2009) as a labeled dataset to train these algorithms (Kox et al., 2014; Håkansson et al., 2018; Wang et al., 2020; White et al., 2021). For supervised ML applications, having a labeled dataset that spans a wide range of plausible conditions is a critical component to ensuring the resulting model is capable of generalizing to unseen observations. Several satellite imagers make observations collocated with CALIOP, but there are specific limitations that need to be accounted for. MODIS/CALIOP collocations only occur at low MODIS viewing angles. Collocations between VIIRS and CALIOP do not contain sun glint and models trained with these datasets can require adjustments for algorithms that make use of visible channels (White et al., 2021). Thus, while CALIOP is a useful resource, there are caveats regarding the representativeness of these collocations that can require special treatment.

In this work, we explore strategies to transition ML algorithms to imager cloud climate records made up of multiple sensors. Specifically we concern ourselves with creating continuous record of cloud-top pressure (CTP) from the VIIRS and MODIS instruments that makes use of neural networks trained to match CALIOP. This is a challenging task since, as previously mentioned, VIIRS and MODIS both have different distributions of collocations with CALIOP. Therefore it is likely that models trained separately for each sensor may exhibit different behavior. Both VIIRS and MODIS have common channels

around 8.6 μm, 11 μm and 12 μm, but their central wavelengths and spectral response functions differ. Furthermore, their spatial resolutions differ making it difficult to obtain comparable spatial metrics between both instruments. This means there is no guarantee that a model trained on VIIRS and CALIOP collocations will perform accurately when supplied with MODIS data. While there have been many efforts to demonstrate the utility of ML-based methods for cloud property estimation, there are not many evaluations of how well these methods generalize to sensors they are not trained with.

We aim to explore whether special treatment is needed for transitioning ML-based cloud-top pressure (CTP) models to climate records made up of two satellite imagers. We perform experiments where neural network models for each sensor have access to different sets of channels and where one or both imagers have access to labeled data. We show that a simple modification to the loss function of a neural network can generally improve consistency between VIIRS and MODIS cloud-top pressure if we take advantage of collocations between these two instruments. In all instances we compare our trained neural networks to the CLDPROP CTP product and neural networks without our proposed modification. This analysis is motivated by the increase in performance of ML-based CTP algorithms relative to current operational approaches, the increasing prevalence of ML-based remote sensing applications, and the growing need for reliable cloud climate records with high intersensor consistency.

3.2 Data

Instruments

VIIRS is a polar-orbiting satellite imager with 16 moderate resolution channels spanning visible, near-infrared and infrared wavelengths with a spatial resolution of 750 m at nadir. Its swath width of 3,600 km allows for at least twice daily views of earth with more frequent coverage at higher latitudes. VIIRS is currently on board the Suomi National Polar-orbiting Partnership (Suomi-NPP) and NOAA-20 satellites, and is additionally planned to launch on the future JPSS-2, -3, and -4 satellites. For this study we only use data from the Suomi-NPP VIIRS. In order to avoid issues with the representativeness of sun glint in the collocation dataset with CALIOP, we restrict ourselves to using only the three infrared channels without any solar contribution (Table 2.1). Their spectral response functions are shown in Fig. 3.1.

MODIS is a polar-orbiting satellite with 36 spectral channels at resolutions up to 250 m. The channels used from MODIS in this analysis all have a spatial resolution of 1 km. Relative to VIIRS, MODIS has a more narrow swath width resulting in some gaps in daily coverage near the equator. MODIS is currently on board the Terra and Aqua satellites launched in 2001 and 2006 respectively. MODIS has three channels that approximately match the spectral response function of the M14, M15 and M16 channels on VIIRS (Table 2.1). In addition to these channels, we use MODIS channels 27, 28 and 30 with central wavelengths of 6.8 μm, 7.3 μm and 9.7 μm in some of the following analysis. Their spectral response functions are also shown in Fig. 3.1.

MODIS and VIIRS differ on how the spatial resolution of the sensor varies with viewing angle. The VIIRS processing aggregates measurements at low viewing angles making for

a nadir resolution of 750 m, but as the viewing angle increases, these measurements are disaggregated allowing for a more consistent spatial resolution at higher viewing angles. A thorough explanation of this feature of VIIRS can be found in Cao et al. 2014. MODIS does not have this functionality and its nadir spatial resolution of 1 km increases monotonically at higher viewing angles.

CALIOP serves as our reference instrument for training our neural networks to estimate cloud-top pressure. CALIOP is particularly sensitive to thin cirrus clouds making it a suitable validation source for many cloud properties including cloud detection and cloud-top pressure estimation from passive imagers (Holz et al., 2008). However, CALIOP only makes observations near-nadir to the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite and only views the same ground location at roughly 16-day intervals. Furthermore, CALIOP data have only been available since its launch in 2006 making validation efforts in early portions of many imager records difficult. Even so, CALIOP has been used to validate cloud property estimates for a wide array of imagers and its lifetime overlaps significantly with MODIS Aqua and entirely with S-NPP VIIRS. In this work, we use the cloud-top pressure estimates from version 4.2 of the 1 km CALIOP cloud layer product (Vaughan et al., 2009) to train and evaluate our neural network approaches.

Collocations between Imagers and CALIOP

Due to the slightly differing orbits of S-NPP and the CALIPSO satellites, VIIRS makes collocations with CALIOP roughly every 2 to 3 days. Collocations are found by matching each 1 km profile from the CALIOP cloud layers product to the nearest imager pixel. The differences in viewing angles between CALIOP and imagers can cause slight issues matching

observations. Thus, a parallax correction is applied using the altitude of the uppermost cloud layer observed in the CALIOP observations and the VIIRS viewing geometry. CALIOP collocations with MODIS are found in a similar manner. MODIS and CALIOP have been in the same A-train constellation (L'Ecuyer and Jiang, 2011) for the vast majority of their operational lifetimes until the formation of the C-train (Braun et al., 2019). Due to this, there is a much larger availability of collocations between MODIS and CALIOP than with VIIRS and CALIOP. The C-train formation also changes the spatial distribution of collocations found between VIIRS and CALIOP.

These datasets are split into a training, validation, and testing sets. Training datasets are used to train the neural networks. Validation datasets are used during development to tune hyperparameters and for early-stopping. Testing datasets are used to assess the overall performance of the models on unseen data. The VIIRS/CALIOP training set comes from 2019 and 2017, the validation set comes from 2018, and the testing data comes from 2016. MODIS/CALIOP collocations are split into the training set which consists of all except the first 7 days in each month in 2017, a validation set of the first 7 days of each month in 2017, and a testing dataset of the first 15 days of each month in 2016. Note that there is a small amount overlap between the time periods used in training set for VIIRS and the validation set for MODIS. However there is no overlap in time with the testing datasets meaning that the data used during training and development are independent from the performance results reported later in this work. The spatial distribution of collocations are shown in Fig. 3.2.

VIIRS and MODIS Collocations

Our proposed approach includes the use of matching observations between two imagers. Since viewing angle and cloud-top altitude pose a significant challenge in selecting matching observations between two imagers, we find a set of approximately ray-matched collocations between VIIRS and MODIS. This is done by collecting observations from these two instruments that are made within a ground distance of 250 m, at sensor zenith and azimuth angle differences of less than 3 degrees, and at time differences less than 2 minutes. The majority of these collocations that meet these requirements occur at the high latitudes. The somewhat loose time difference requirement is specified to allow for a small amount of VIIRS/MODIS collocations at the equator. We remove all collocations where the MODIS-VIIRS Continuity Cloud Mask (MVCM; Frey et al. 2020) did not observe a cloud from both sensors. Furthermore we remove all collocations where the M15 band on VIIRS and the band 31 on MODIS differ by more than 20 K. While differences are to be expected due to differing spectral response functions, a difference as large as 20 K is more likely due to the advection of cloud cover.

CLDPROP

As mentioned previously, CLDPROP is a project with the intent of creating a continuous cloud climate record from MODIS and VIIRS. We use the CLDPROP CTP product as a benchmark to evaluate the effectiveness of our approach in creating an intersensor consistent product. The CLDPROP CTP is based on an optimal estimation (Rodgers, 1976) retrieval described in Heidinger and Li 2017 and is similarly used in NOAA operations for VIIRS and ABI as well as the PATMOS-x AVHRR climate record. CLDPROP CTP will be evaluated

with respect to CALIOP and differences between MODIS and VIIRS for ray-matched collocations will be assessed. Given the significant effort that has gone into creating the CLDPROP products and the underlying algorithm it is based on, we believe it serves as a reasonable benchmark to evaluate the consistency of our proposed approach.

3.3 Methodology

Neural Network Architecture and Training

For each imager, we train a neural network with a total of five fully connected (FC) layers with 64, 32, 32, 16, and 1 units (Fig.3.3) . All except the last FC layer are followed by rectified linear unit (ReLU) activation functions. The weights of the last three layers are shared between both imagers and the first two layers are trained specifically for each imager. The goal of this design is to incentivize the sensor-specific layers to learn similar representations from each sensor. These representations are then passed to the shared layers with the goal of reaching a similar prediction. We believe that sensor-specific layers are necessary due to the differing nature of the features derived from VIIRS and MODIS.

The inputs to the neural network are made of up brightness temperatures (BTs) from the channels in Table 2.1, spatial metrics derived from these channels, and ten features from Numerical Weather Prediction (NWP) model output. Since the VIIRS and MODIS resolutions differ, we calculate the spatial metrics from each instrument differently. For VIIRS, we take a 5 pixel by 5 pixel array and calculate the difference between the central pixel and both the minimum and maximum BT. We also include the standard deviation of BT in the same 25 pixels. A 5 by 5 pixel array has an edge length of 3.75 km at nadir for

VIIRS data and an edge length of 5 km at nadir for MODIS data. To bring the spatial metrics into closer agreement in terms of area observed, we calculate the same spatial metrics for MODIS over a 3 by 3 pixel array. The NWP information comes from 6-hour forecasts from Climate Forecast System (CFS; Saha et al. 2014) model output at 6 hour intervals and are interpolated linearly in both space and time. The exact fields used are the temperatures at 900, 700, 500, 300, 100, and 20 hPa as well as pressure and temperature at the surface and tropopause.

Before training, we standardize the features by dividing by the mean and standard deviation calculated from the training dataset. We also scale the predictand by dividing by 1000. This is done to reduce issues related to the relative scales of the each feature and generally improves performance and reduces training time of the neural network. In some instances (described later in section 3.4) we fit a linear regression and adjust analogous MODIS channels to more closely align with their VIIRS counterparts.

Each batch is comprised of 10,000 examples randomly sampled separately from each of the VIIRS/CALIOP, MODIS/CALIOP, and VIIRS/MODIS datasets for a total of 30,000 examples per batch. The Adam optimizer (Kingma and Ba, 2015) is used with learning rate of 0.005 and is reduced by a factor of 10 when there is no reduction in loss calculated on the validation dataset after 3 epochs. Training is stopped after 5 epochs with no improvement on the validation dataset. All models in this study finished training in less than 50 epochs.

Loss Function

$$L_{Total} = L_{VIIRS-CALIOP} + L_{MODIS-CALIOP} + \alpha_c w_\phi L_c \tag{3.1}$$

$$L_c = |\hat{y}_{VIIRS} - \hat{y}_{MODIS}| + \frac{1}{N}\sum_{i=1}^{N}\left|A_{i,VIIRS} - A_{i,MODIS}\right| \tag{3.2}$$

$$w_\phi = \frac{1}{\frac{|\phi|}{60} + 0.1} \tag{3.3}$$

We use a loss function that incentivizes both low differences between predictions and the CALIOP reference labels and low differences between predictions made by VIIRS and MODIS for the ray-matched collocations (Eq 3.1). L_{Total} is the total loss comprised of the absolute error between the VIIRS prediction and the CALIOP label ($L_{VIIRS-CALIOP}$), the absolute error between the MODIS prediction and the CALIOP label ($L_{MODIS-CALIOP}$), and a consistency term (L_c; Eq 3.2) multiplied by a constant (α_c) and a latitude-dependent weight (w_ϕ; Eq 3.3). Since the ray-matched collocation dataset under-represents the lower-latitudes, w_ϕ acts to increase the contribution of L_c to L_{Total} from the relatively small amount of low-latitude samples in our datasets. The consistency loss is a two component loss where the first component is the absolute value of the difference between the VIIRS and MODIS predictions ($\hat{y}_{VIIRS}$ and $\hat{y}_{MODIS}$). The second component is the difference between intermediate activations. Specifically, $A_{i,VIIRS}$ and $A_{i,MODIS}$ are the activations of the last sensor-specific layer before the ReLU operation is applied. In this case, $N = 32$ since this layer has 32 units. Eq. 3.1 and Eq. 3.2 are both expressed for a single example, and are averaged over the entire batch during training. During training, $L_{VIIRS-CALIOP}$ and $L_{MODIS-CALIOP}$ are both calculated from datasets of collocations between CALIOP and their respective imager. L_c is calculated from samples of the MODIS/VIIRS collocation dataset.

It may not initially be obvious why two components might be needed in Eq. 3.2. We expect that the second component is beneficial since it provides more direct feedback to the weights of the sensor-specific layers. We expect the first component is beneficial since there is no upper or lower bound on the activations, and relatively small differences in activations may still yield large differences in predicted CTP. Thus minimizing the second component ensures that the preprocessing layers find a common representation, and the first component ensures that any small differences in representations that remain do not yield large differences in predicted CTP when passed through the shared layers. When the differences in activations are removed, both intersensor differences and training time increased. When the differences in predictions are removed, intersensor differences increased relative to including both terms. We also note that removing the differences in activations increased the sensitivity of our results to the chosen value of α_c and required more fine-tuning to achieve reasonable performance. It is likely that weighting each component of L_c differently could yield better results. We found that equal weights worked reasonably well in our application, but we caution that this may not be a generally applicable result and could depend on the relative scales of each term.

3.4 Results

We divide our experiments into four separate categories defined by two characteristics. The first characteristic is whether CALIOP labels from only a single imager or both imagers are available during training. We refer to the single imager case where we only use VIIRS/-CALIOP labels as single-source training (SST) and the multiple imager case where both

VIIRS/CALIOP and MODIS/CALIOP labels are used as multiple-source training (MST). The second characteristic is whether or not we restrict the neural networks to use only shared channels between each sensor. We refer to the case where only information from the VIIRS M14, M15 and M16 channels and their MODIS counter parts (MODIS-29, -31, -32) are used as similar channel subsets (SCS). We then compare this to different channel subsets (DCS) where we include additional channels in the MODIS models that do not exist on VIIRS (MODIS-27, -28, and -30).

In each experiment, we assess performance over a range of values of α_c to explore the impact of L_c on the agreement with CALIOP and the consistency between the two imagers. For each experiment we include the CLDPROP CTP product and a neural network baseline (where possible) for comparison. In the MST-SCS and the MST-DCS experiments, the baseline neural networks are trained independently for each sensor without L_c and without shared layers using otherwise identical hyperparameters. In the SST-SCS experiment, a MODIS neural network baseline cannot be trained due to the lack of MODIS/CALIOP labels. In this case, we assume VIIRS and MODIS data are interchangeable and the VIIRS baseline neural network is used to make predictions with MODIS data. This is not possible for the SST-DCS experiment, since the size of the inputs and nature of the considered channels differ between VIIRS and MODIS. Thus, our only baseline for the SST-DCS experiment is the CLDPROP CTP product. These experiments are meant to represent a variety of plausible scenarios when attempting to develop neural network models for one or more imagers with the objective of creating a consistent record of derived cloud products.

In the SCS experiments there is an additional preprocessing step we apply before the mean and standard deviation standardization. We fit a linear regression between the MODIS

features and the VIIRS features (Fig. 3.4). This is done to account for biases between channels and spatial metrics and to generally bring information from the two sensors into closer alignment. A similar approach is applied in the Cloud CCI retrieval to apply a neural network cloud mask trained with AVHRR data to other sensors such as MODIS and AATSR (Sus et al., 2018). Our linear regression parameters are determined from the VIIRS/MODIS ray-matched collocation dataset.

Multiple-Source Training

We first present the results from the MST-SCS and MST-DCS experiments. We choose values for α_c of 0, 0.01, 0.1, 0.3, 0.5, 1.0, and 2.0, and 5.0. $\alpha_c = 0$ is a special case where the L_c term has no effect on L_{Total} and represents a scenario where there are no comparable observations between two imagers, but both imagers make collocations with CALIOP. Thus, there is no mechanism for explicitly enforcing consistency between the two neural networks aside from the last two layers with shared weights. For each neural network we calculate the MAE and bias between CALIOP and each imager individually, as well the MAE and bias between the two imagers. All results are reported for the testing datasets (Fig. 3.2).

Figure 3.5 shows a summary of our MST experiments and a comparison of our results to the CLDPROP CTP product. We observe that the value of α_c has only small impact on each model's performance with respect to CALIOP (Fig. 3.5.a, b). It is only at large values ($\alpha_c = 2$) that we begin to see MAE with CALIOP increase as consistency between imagers is prioritized in the loss function. Across all values of α_c we observe only small biases with respect to CALIOP. The neural networks for MODIS show similar behavior (Fig. 3.5.c and d). However, the DCS case on MODIS has a slightly lower MAE with CALIOP likely

due to the use of channels not available to VIIRS. In all of these cases the baseline neural networks without L_c and shared layers perform similarly.

When we consider the intersensor differences (Fig 3.5.e,f) it is clear that the neural network baseline models fail to make reasonably similar predictions between both imagers. The MST-DSC neural network baseline is the worst model in terms of intersensor differences with a VIIRS/MODIS MAE more than twice that of CLDPROP. The baselines perform similarly to the neural networks with shared layers and $\alpha_c = 0$. This may indicate that shared layers are not enough to improve consistency between these models. However, once α_c is increased, the consistency improves considerably. The neural networks trained with L_c reach parity with CLDPROP roughly where $\alpha_c = 0.1$. Further decreases in VIIRS/MODIS MAE are observed with increasing α_c. We expect that the DCS case requires a slightly larger value of α_c to reach parity with CLDPROP due to the need to deter the MODIS model from exploiting variability only observed in channels not available to the VIIRS model. We also note that the Neural Networks trained with L_c and non-zero α_c can achieve lower VIIRS/MODIS biases compared to CLDPROP.

Single-Source Training

Next, we summarize the results from the SST experiments to simulate a scenario where only one of the two imagers (VIIRS) has labeled data from CALIOP (Fig.3.6). This is a particularly difficult task since the only way the MODIS-specific layers can be trained is through the L_c term where the goal is to match the intermediate activations and predictions of the VIIRS model. We do not include the α_c=0 since L_{Total} reduces to $L_{VIIRS-CALIOP}$ which means that there is no mechanism for training the MODIS-specific layers in Fig.3.3.

Starting with Fig. 3.6.a and b, we observe that the MAE and bias between VIIRS and CALIOP are very similar both between the SCS and DCS scenarios, and also with the MST experiments. Results differ greatly with the MST experiments when considering the MODIS/CALIOP results for the SST experiments in 3.6.c and d. As α_c is increased, the MODIS/CALIOP MAE increases sharply. We expect this difference occurs due to $\alpha_c w_\phi L_c$ making up an overall larger portion of L_{Total} compared to MST experiments. The MODIS preprocessing layers are also only made to minimize differences with the intermediate representations and predictions of the VIIRS layers, and there is otherwise no mechanism to incentivize low MAE between MODIS and CALIOP as there is in the MST experiments. This is representative of an undesirable solution to minimizing L_{Total} when L_c becomes the dominant term. In such a case both models would be incentivized to predict the same value for CTP regardless of whether it successfully minimizes $L_{VIIRS-CALIOP}$ or $L_{MODIS-CALIOP}$. In all VIIRS/CALIOP and MODIS/CALIOP metrics, the SST-SCS neural network baseline slightly outperforms all models trained with L_c.

When considering the consistency between VIIRS and MODIS, the SST-SCS baseline does not outperform CLDPROP or any of the neural networks trained with L_c. At $\alpha_c = 0.01$ the neural network matches the VIIRS/MODIS MAE of CLDPROP with near-zero bias. Further improvements in VIIRS/MODIS MAE occur at larger values of α_c although they come with the expense of significantly reduced performance in the MODIS/CALIOP metrics. These results imply that α_c=0.01 or 0.1 may be an optimal choice since they roughly match the MODIS/VIIRS consistency of CLDPROP with low bias and have reasonable performance with respect to CALIOP. However, we stress that in a practical SST scenario, one might not have access to a set of labeled data for the secondary imager, and one may

not be able to determine an appropriate value of α_c this way.

Time Series Analysis

The dataset used to evaluate the intersensor consistency has a few important caveats. Due to the orbits of each imager and the requirements on obtaining an approximately ray-matched collocations between them, there are several cases where this dataset is not representative. This is particularly true in the low latitudes (Fig 3.2.g,h,j). A second issue is that cloud motion, even at small time differences between MODIS and VIIRS, can impact how mutually representative the ray-matched collocations are.

To provide an alternative perspective to the problem of VIIRS/MODIS intersensor consistency, we collect a time series of CTP estimates from several of these neural network models and CLDPROP starting in 2013 through the end of 2015. From each experiment we select one neural network trained with L_c based on the trade off between each imager's performance with respect to CALIOP and the consistency between MODIS and VIIRS. We then choose four regions defined by the latitude-longitude bounding boxes shown in Table 2.2. Data from VIIRS and MODIS are sampled to regularly spaced 0.05 degree grid daily composites within each region. For each day, only the most nadir observations are used. Predictions from the neural networks are masked by the MVCM cloud mask and only grid points where both sensors observed the location within 1 hour of each other are retained.

Earth Mover's Distance

In this instance, one cannot calculate grid-point-level differences between the two imagers and expect representative results. Instead, we compare the overall distributions of

CTP predictions for 20-day composites. We then use the Earth Mover's distance (EMD; often referred to as the Wasserstein metric) to calculate the differences in CTP probability distribution functions (PDFs) between MODIS and VIIRS. EMD is a general approach for quantifying differences between distributions and can be described in general terms as the minimum cost of transforming one distribution into another. A thorough discussion of EMD and its useful characteristics in several contexts can be found in Panaretos and Zemel 2019. We use an implementation provided by version 1.7.1 of the SciPy python library Virtanen et al. (2020). Our expectation is that this approach may be more robust to advection of cloud cover within the specified domains that would otherwise affect grid-point-level differences.

Differences in the CTP distributions among the VIIRS and MODIS predictions are shown in Fig.3.7 for the MST experiments. A low EMD value implies small differences between CTP distributions. The neural network baselines perform reasonably well compared to CLDPROP in all regions except region 3 and the MST-DCS baseline in region 4. The MST-SCS neural network trained with the consistency loss has the lowest EMD in the majority of 20-day composites particularly in the lower-latitudes (regions 1 and 2). The MST-SCS neural network trained with L_c also appears to be the most consistent in time despite not always being the approach with the lowest EMD. The MST-DCS neural network trained with L_c shows comparable EMD to CLDPROP, and much lower EMD relative to its respective baseline in the higher latitudes (regions 3 and 4). Some of these approaches show some seasonality in their EMD values such as CLDPROP in region 4, the MST-SCS baseline in region 3, and several models in region 2. This indicates that certain environmental conditions could contribute to the differences observed between CTP algorithms for each sensor.

Similar characteristics of the EMD analysis are shown for the SST approaches in Fig. 3.8. All SST neural network approaches have lower comparable EMD with CLDPROP with the exception of the SST-SCS baseline in regions 2 and 3. It also shows strong seasonality in region 3 and more moderate seasonality similar to CLDPROP in region 4. The neural networks trained with L_c have the lowest EMD values throughout with both the SCS and DCS scenarios showing very similar results.

Cloud Fraction Differences

EMD is a useful metric for comparisons of PDFs. However, they do not illustrate in a physically intuitive way the impact these differences may have on quantities frequently used in trend analysis. To better put these results in perspective, we frame the previous time series analysis in terms of differences in cloud fraction at various levels. For each 20-day composite we calculate the frequency of high (0-400 hPa), middle (400-700 hPa) and low (700-1050 hPa) level clouds. We then calculate the difference in these quantities for both imagers.

Fig. 3.9. shows the differences in high middle and low cloud fraction between VIIRS and MODIS for the MST experiments in the regions used previously in this analysis. As suggested by EMD, the MST-DCS baseline model shows the largest differences at the upper- and middle-levels occasionally exceeding 8% but has relatively moderate differences typically less than 4% at the lower levels. The other neural networks typically have differences less than 4% at all levels, although the seasonality issues with the MST-SCS baseline in region 3 are apparent. Disagreement in CLDPROP is relatively low at the middle-levels with most differences less than 2%. CLDPROP has larger issues with upper-level cloud

fraction in regions 1 and 4 and lower-level cloud fraction in region 1. Across all regions CLDPROP tends to predict more lower-level clouds from VIIRS, and more upper-level clouds in MODIS. In most cases, implementing the consistency loss when training neural networks appears to improve consistency between sensors. An exception is region 2, where the MST-DCS neural network trained with L_c has larger differences compared to its respective baseline, but smaller differences in other regions.

Fig. 3.10 shows the same differences with the SST experiments plotted with CLDPROP as a reference. Comparisons of the SST-SCS neural network trained with L_c and its corresponding baseline are favorable. Region 1 is an example where these two approaches are roughly comparable, but seem to be a slight improvement upon CLDPROP. Outside region 1, the SST-SCS Baseline has numerous issues including large differences at all levels in region 2, and seasonality in the upper and middle levels of region 3 and 4. On the whole the SST-SCS neural network trained with L_c appears to be an improvement upon the baseline and a slight improvement upon CLDPROP. The SST-DCS neural network mostly performs similarly to the SST-SCS neural network with L_c with larger differences appearing occasionally.

Changes in Relative Feature Importance

Much of the previous analysis demonstrates that we can bring two neural networks closer in agreement by modifying the loss function to optimize for intersensor consistency. One might expect that when L_c is added to the loss function, that the two neural networks are incented to only rely on features that are more similar between the two instruments. With this line of reasoning we might expect a decrease in the usage of spatial features that depend

heavily on the resolution of the sensor, and how that resolution changes with increasing viewing angle. Similarly we might observe an increase in the use of the 8.4 μm where the spectral response functions between the instruments are most similar or a heavier reliance on NWP model output which are the same for each imager.

To investigate how the usage of particular features changes with L_c, we use Layerwise Relevance Propagation (LRP; Bach et al. 2015) to estimate feature importance. LRP is a popular interpretability tool for neural networks and has been used in a variety of applications in atmospheric science and remote sensing (Hilburn et al., 2021; Barnes et al., 2020; Toms et al., 2020). There are several variants of the LRP rules that typically differ on how relevance values are calculated throughout the model. We use the LRP $\epsilon = 1$ rule, which improves the numerical stability of the relevance values. Since interpreting LRP relevance values can be a difficult task in itself, we simplify this analysis by expressing the LRP relevance relative to the most important predictor. We do this by taking the absolute value of the LRP output, and dividing by the largest relevance in each example. Thus, a relative feature importance of 1.0 indicates that the feature was the most important and a relevance of 0 implies that the feature was not at all important for CTP prediction. The feature importance is calculated on the ray-matched collocation dataset so that comparisons may be made between VIIRS and MODIS. We perform this analysis for the MST-SCS and -DCS scenarios and compare the baseline neural networks with the neural networks trained with L_c.

Figure 3.11 shows how the relevance values change after L_c is added to the loss function in the MST-SCS experiments. There are a few unexpected changes evident in this analysis. We observe an increase in reliance on low-level temperatures from NWP from both VIIRS

and MODIS, but there are mixed results for upper-level temperatures as well as surface and tropopause information. Both sensors show an intuitive decrease in the importance of spatial features, which can be attributed to the different spatial resolutions. By far, the largest change comes from the spectral features where there is a relatively large decrease in the importance of M14/MODIS29 and M15/MODIS31 and a large increase in the usage of M16/MODIS32. This is somewhat unintuitive due to the difference in spectral response functions of these channels shown in Fig. 3.1. A potential explanation might be that neural networks are relying more on a single spectral channel, and less on differences between channels, where BT differences could be more heavily impacted by the specific SRFs of each instrument. However it is unclear why the 12.0 μm channel is favored over others.

Figure 3.12 shows a similar analysis for the MST-DCS experiments. Again, we observe a mix of changes in the usage of NWP features, and a general decrease in the usage of spatial metrics. We also observe a similar pattern in the importance values of the M14/MODIS29, M15/MODIS31 and M16/MODIS32. The channels specific to the MODIS models are the 9.7 μm, 7.3 μm, and the 6.8 μm. As expected, when L_c is added to the loss function the importance of the these channels dramatically decreases. Their impact, however, is not completely eliminated and is comparable to several spatial metrics and NWP features. Potential explanations could be that these features are still useful for matching CALIOP, or can be used to correct for differences in the shared channels. We note that there are a few rare cases where the MST-DCS neural network with L_c has lower differences in EMD and cloud fraction compared to its SCS counterpart so it is plausible that these MODIS-specific channels may be useful for matching VIIRS CTP despite not being shared with VIIRS. Overall, while adding L_c to the loss function can sometimes improve the intersensor

consistency from a number of different perspectives, it has changed the importance of particular features in some unexpected ways for this application.

3.5 Discussion

We have explored the effectiveness of this methodology to improve intersensor consistency among a CTP neural network developed for VIIRS and MODIS under a variety of scenarios and compared it to standard neural network baselines and the CLDPROP operational CTP product. This fairly simple approach can improve our ability to transition ML algorithms to long-term records made up of multiple imagers in select circumstances by reducing inconsistency in derived products. Compared to more standard neural network applications such as in Kox et al. (2014), Håkansson et al. (2018) and White et al. (2021), the proposed methodology requires adding sensor-specific input layers, an additional term to the loss function, collecting coincident observations between two imagers, and choosing an appropriate value of α_c.

The MST cases represent the most favorable scenarios where both imagers have labeled data. Since the distributions of labeled data for each instrument are different, it is reasonable to expect that a neural network developed for each sensor may not be consistent in terms of their predictions. We test this possibility and find that the baseline neural networks perform relatively poorly in terms of intersensor consistency compared to CLDPROP despite outperforming CLDPROP in their comparisons with CALIOP. Our proposed methodology improves consistency in the ray-matched collocation dataset when the contribution of L_c is increased by increasing the value of α_c. This increase in consistency occurs for both the

SCS and DCS experiments but at the expense of decreased performance with respect to CALIOP.

When we examine the time series of EMD and cloud fraction differences for several regions, we still see improvement for the MST models that are trained with L_c, but we occasionally observed mixed results for the MST-DCS model. This may indicate that our approach is sensitive to the distribution of collocations in the ray-matched dataset especially considering that the MST-SCS and MST-DCS models achieve similar VIIRS/MODIS MAE for the ray-matched dataset. The slight disparities in consistency between the lower- and higher-latitudes of the cloud fraction analysis offer some evidence for this. Overall, the models trained with the consistency loss (L_c) appear to have lower differences and less seasonality high latitudes.

The SST cases illustrates a more challenging scenario where only a single imager (VIIRS) has labeled data with CALIOP but we have matching observations with the secondary imager (MODIS). We compared our methodology to a baseline that assumes that after a linear adjustment to MODIS data, that MODIS and VIIRS data are interchangeable if we only consider shared channels. This solution performs well with respect to CALIOP, but has high intersensor differences between MODIS and VIIRS. The SST-SCS and -DCS models outperform CLDPROP at moderate values of α_c in all metrics in Fig3.6.

When considering the time series analysis of the SST experiments, we see that the SST-SCS baseline has similar EMD values to CLDPROP but has issues with seasonality in regions 2,3 and 4. Our proposed method consistently has lower EMD in all regions, but these differences are not as easily seen in the cloud fraction analysis with the exception of region 2. This indicates that the differences observed in the EMD analysis may not occur

over the 700 and 400 hPa thresholds that define our high-, middle-, and low-level categories. The SST-DCS neural network performs similarly to its SCS counterpart albeit with slightly higher EMD in region 1 and occasionally larger cloud fraction differences.

If α_c becomes too large, the MODIS predictions can be severely affected in all experiments. For the MST experiments, we have a labeled dataset with both sensors so one could monitor the performance of each neural network to ensure that increasing α_c does not degrade performance with respect to CALIOP. However, this is not the case with SST experiments and it is not obvious when α_c should stop increasing. Thus, one must look for other indicators that the CTP performance might be degraded.

This points to a key challenge in transitioning this methodology to practice in SST scenarios: how does one choose an optimal value of α_c that improves intersensor consistency without unrealistically narrowing the range of CTP predictions? To illustrate this point, we find the predicted CTP distribution for the SST-SCS and MST-SCS experiments for a small (0.01) and a large (2.0) value of α_c. These predictions are made for the imager-CALIOP collocations. Figure 3.10 shows that the MST CTP distributions are only slightly affected under this range of values. However, the impacts are more severe for MODIS where α_c is large in the SST experiments. We observe that the overall CTP distribution for large α_c skews heavily towards the mean CTP despite the fact that this same model outperforms CLDPROP with respect to CALIOP (Fig 3.6.c). Since we can expect the distribution of CTP values between VIIRS and MODIS to be roughly similar, α_c could be selected by inspecting the changes in predicted CTP distributions. One could also collect a limited sample of labeled data perhaps from ground-based or in-situ instruments and track the differences with these data with increasing α_c.

There are few other important caveats that should be considered when interpreting the analysis in this work. First, is that our ray-matched collocations are not perfectly representative of either the differences between imagers or the entire range of conditions viewed by them. Despite our filters applied to the ray-matched dataset, it is still susceptible to advection of cloud cover. Thus when minimizing L_c, we are to some degree lowering differences in predicted CTP between different clouds in addition to mitigating real differences between the two imagers. This may contribute to the narrowing of the CTP distributions at large α_c. When we apply a cloud mask to remove clear-sky observations, our ray-matched collocations are subject to the errors of the MVCM, which can be quite large in the arctic regions (White et al., 2021; Frey et al., 2020). Our time series analysis provides some indication that is an issue due to the discrepancy between the intersensor MAE estimated in ray-matched collocations compared to the EMD values and cloud fraction difference in the time series analysis. Thus, the success of this methodology to other applications could be sensitive to the quality and representativeness of the collocated dataset.

In the DCS analysis, we use a set of similar channels and add 3 channels to the MODIS model that don't exist on VIIRS. However, this approach does not necessarily require sharing similar channels at all. Moreover, there is no requirement that the primary imager (VIIRS) only use channels that are available on the secondary imager (MODIS). This is made possible by the separate preprocessing layers included in our proposed model. We expect the potential benefit of using non-shared channels depends on whether similar intermediate representations can be learned by the neural network. One may need to increase the number of units in each layer or the number of layers in the preprocessing sections of the neural network for more disparate inputs. Although we have not specifically tested this here,

we expect that using separate preprocessing layers for each sensor increases the capacity of the neural networks to accommodate differences among the imagers. As more ML-based algorithms are transitioned to high-stakes operational tasks and analyses of climate records, future work is certainly needed to inform what solutions are optimal for mitigating differences associated with changing observation platforms.

Given the large number of imagers that don't have ray-matched collocations with other imagers it may be useful to explore alternative ways of minimizing differences among statistical cloud property models. In a portion of this work we use EMD to quantify differences in CTP distribution. One option could be to replace the L_c term of Eq 3.1 with the EMD of two distributions of comparable scenes. Using EMD might allow for the use of observations at differing viewing angles, or larger time differences. Minimizing EMD has been used in many ML applications, but has particularly seen wide use in Generative Adversarial Models (GANs; Arjovsky et al. 2017). This would also be a path for extending this approach to S-NPP and NOAA-20 VIIRS which make coincident observations at a time difference of roughly 50 minutes. Such an approach could also be tested with ISCCP and ISCCP-NG data where geostationary imager observations overlap significantly in space and time but have very different viewing geometries.

3.6 Conclusions

In this analysis, we demonstrated that intersensor consistency in neural network cloud-top pressure retrievals between two imagers can be improved in select circumstances by a simple change to the loss function and by exploiting coincident observations between two imagers.

The experiments performed illustrate a variety of scenarios regarding the availability of labeled data for each imager and the spectral channels available to each imager. A key challenge in implementing this methodology for the scenario where only a single imager has labeled data is choosing an appropriate weight for the consistency term in the loss function and evaluating the accuracy the model created for the secondary imager with no labeled data. If the weight is set too high, we observe an unrealistic narrowing of the predicted cloud-top pressure distribution. We suggest some strategies for choosing a reasonable value for this parameter. Nonetheless, we show that the neural network approach can outperform the CLDPROP cloud-top pressure product in terms of accuracy with respect to CALIOP and intersensor consistency which is a major concern for satellite cloud climate records. We expect that this methodology could be one pathway for improving the generalization of ML-based remote sensing algorithms and facilitating their transition to climate records.

VIIRS		MODIS	
Band Name	Central Wavelength	Band Name	Central Wavelength
		MODIS 27	6.76 µm
		MODIS 28	7.33 µm
M14	8.55 µm	MODIS 29	8.55 µm
		MODIS 30	9.72 µm
M15	10.76 µm	MODIS 31	11.03 µm
M16	12.01 µm	MODIS 32	12.02 µm

Table 3.1: Shown are the infrared channels without solar contributions from VIIRS and MODIS that are used in this analysis.

Region	Latitude Range	Longitude Range
1	10S - 10N	140E - 160E
2	30S - 10S	0E - 20E
3	30N - 50N	110W - 130W
4	40N - 60N	10W - 10E

Table 3.2: Coordinates of the regions in which comparisons are performed between VIIRS and MODIS CTP distributions.

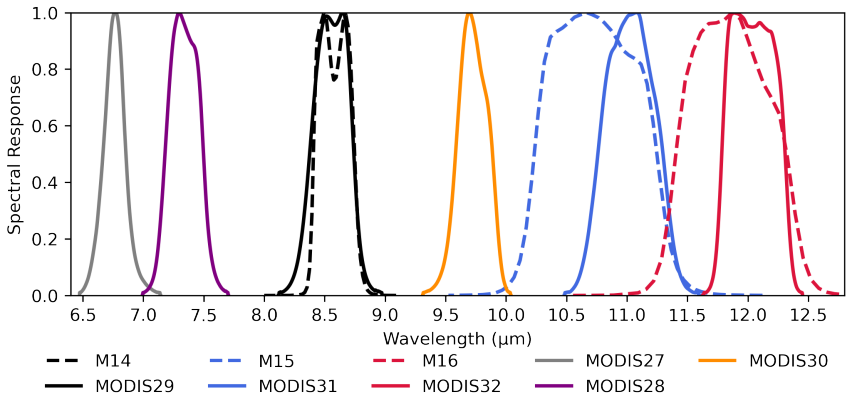

Figure 3.1: The normalized (to 1) spectral response functions of the VIIRS and MODIS channels used.

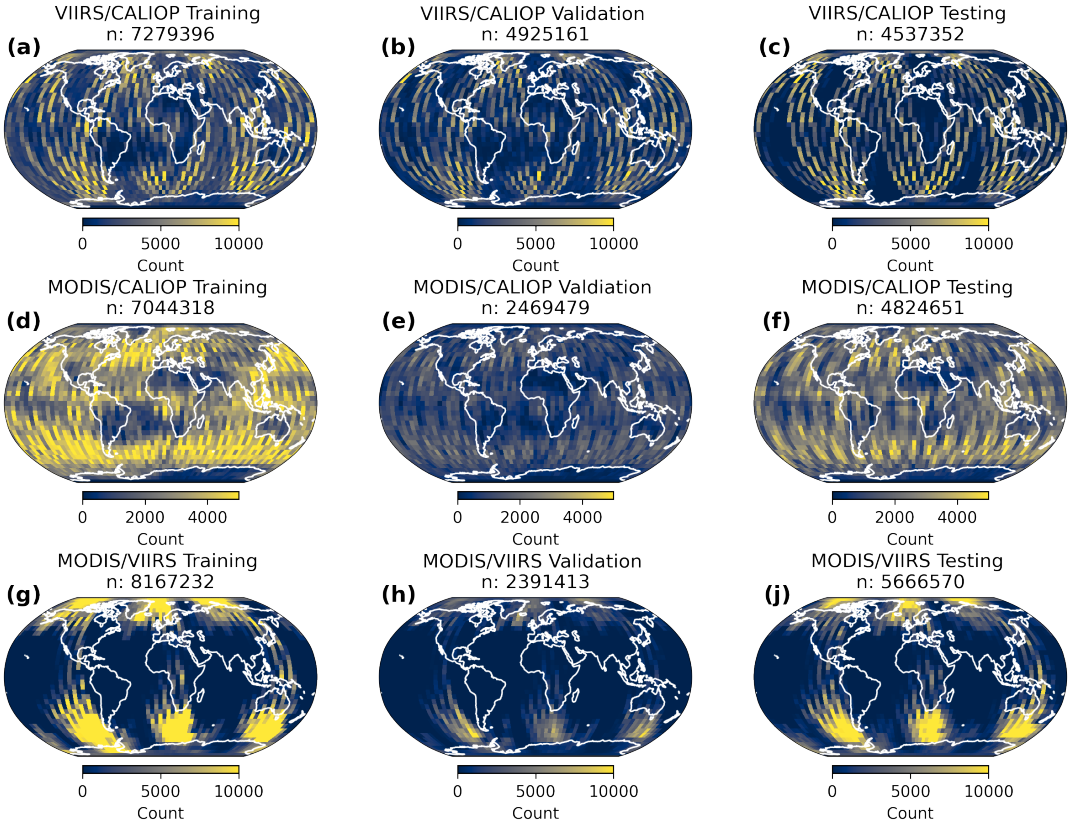

Figure 3.2: The distribution of collocations between VIIRS and CALIOP (a,b,c), MODIS and CALIOP(d,e,f) and MODIS and VIIRS (h,i,j). Shown in each subplot title is the number of total collocations in each dataset. Note the differences in color bars between each sub plot.

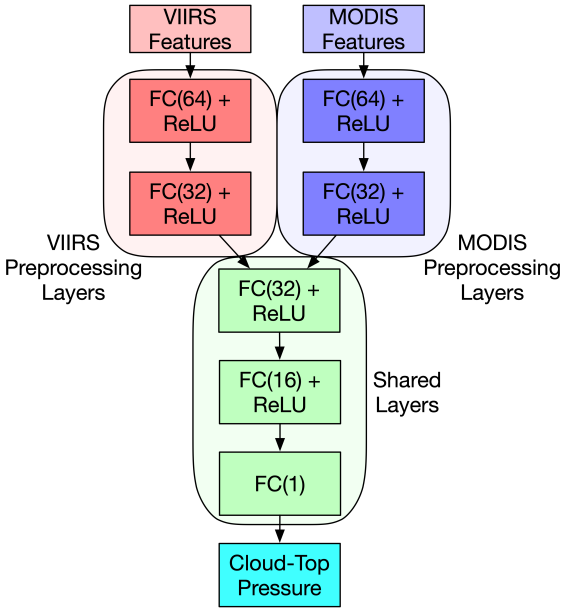

Figure 3.3: Schematic of the neural network used in this work. Each block (except the last) represents a fully-connected layer followed by a rectified linear unit activation. Layers specific to VIIRS are shown in red, MODIS in blue, and shared layers are shown in green.

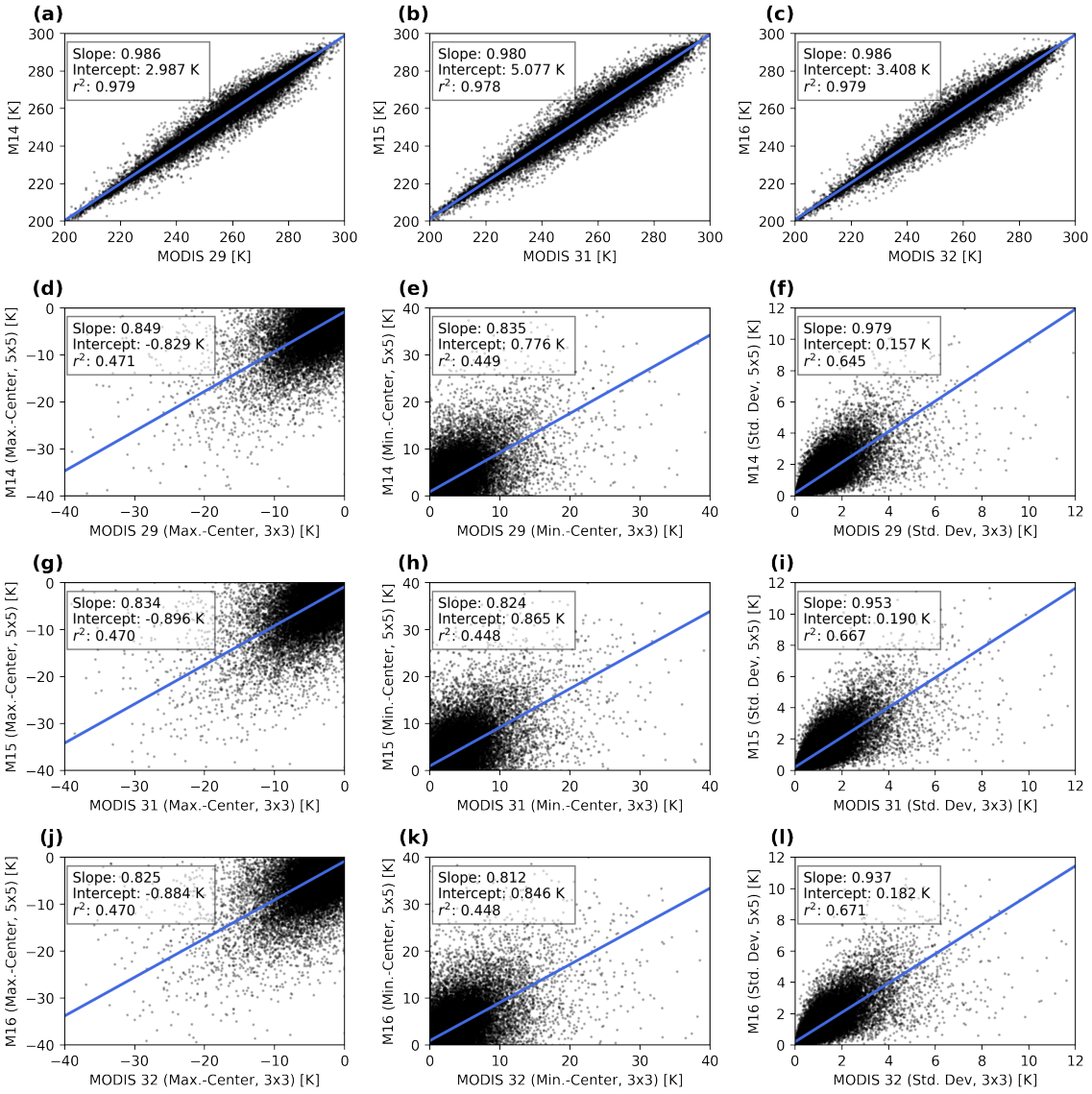

Figure 3.4: Comparison of spectral and spatial features derived from the VIIRS (y-axes) and MODIS (x-axes). The scatter plots represent the values from each sensor before the linear fit (blue) is applied to MODIS data. Only one out of every hundred points are shown to ease visualization.

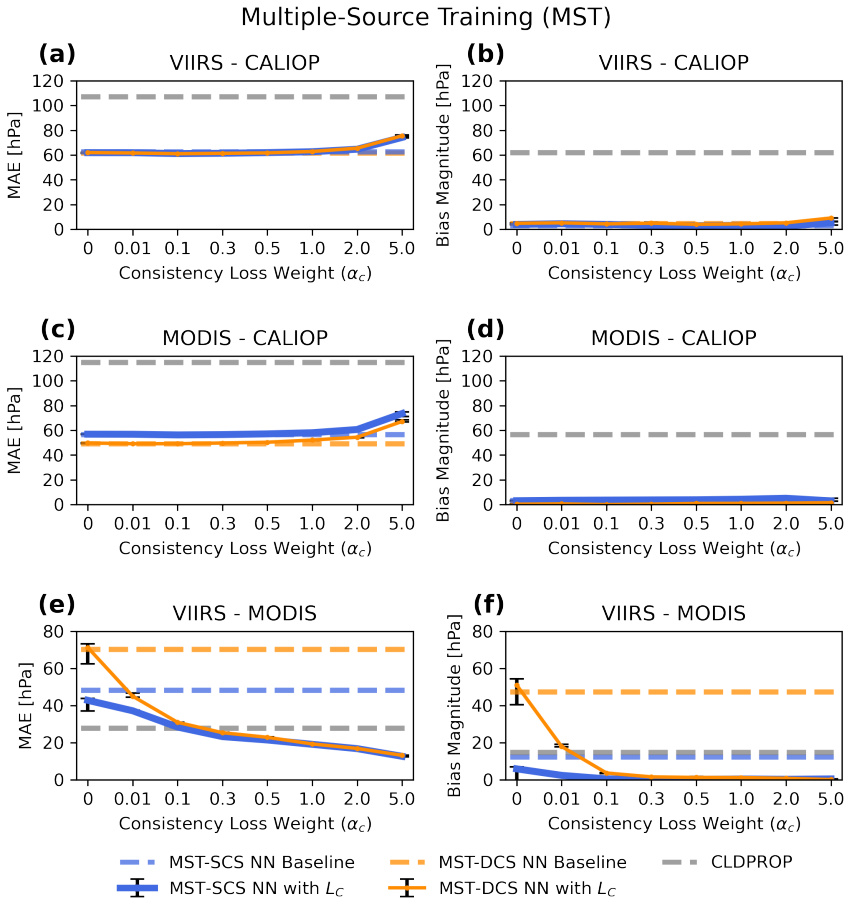

Figure 3.5: Evaluation of the MST experiments for both the SCS and DCS scenarios. Shown are the MAE (a,c,e) and bias magnitude (b,d,f) of each of the three pairings of VIIRS, MODIS and CALIOP. Note that the x-axis intervals are not evenly spaced. Each value of α_c is run for three neural networks with randomly initialized weights that are otherwise identical. Error bars indicate the highest and lowest value of the three models.

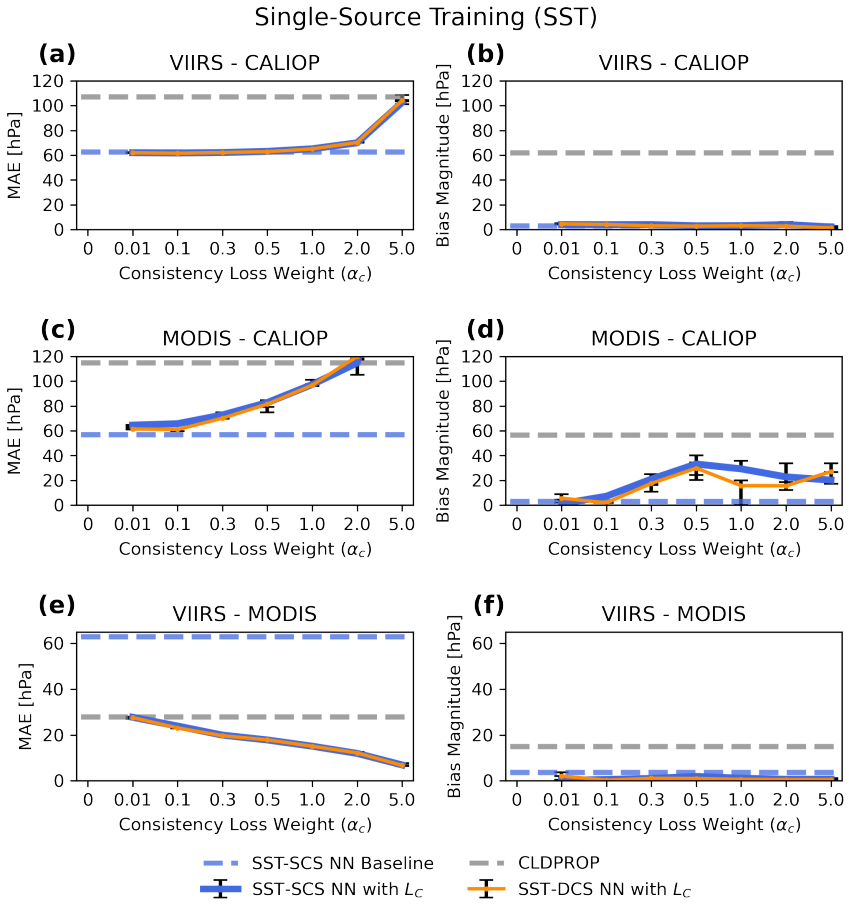

Figure 3.6: Evaluation of the SST experiments. Shown are the MAE (a,c,e) and bias magnitude (b,d,f) of each of the three possible pairings of VIIRS, MODIS and CALIOP. Note that the x-axis intervals are not evenly spaced.

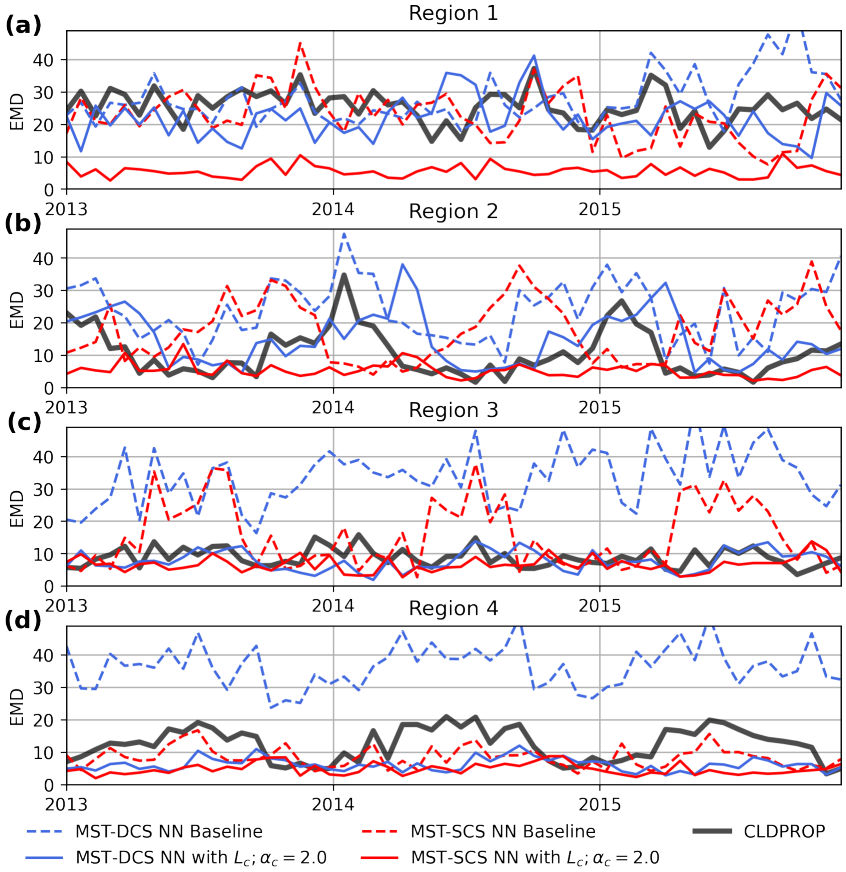

Figure 3.7: Comparison of CTP distributions for geographic regions 1 (a), 2 (b), 3(c) expressed in earth mover's distance for the MST experiments. See table 2 for coordinates of the geographic regions.

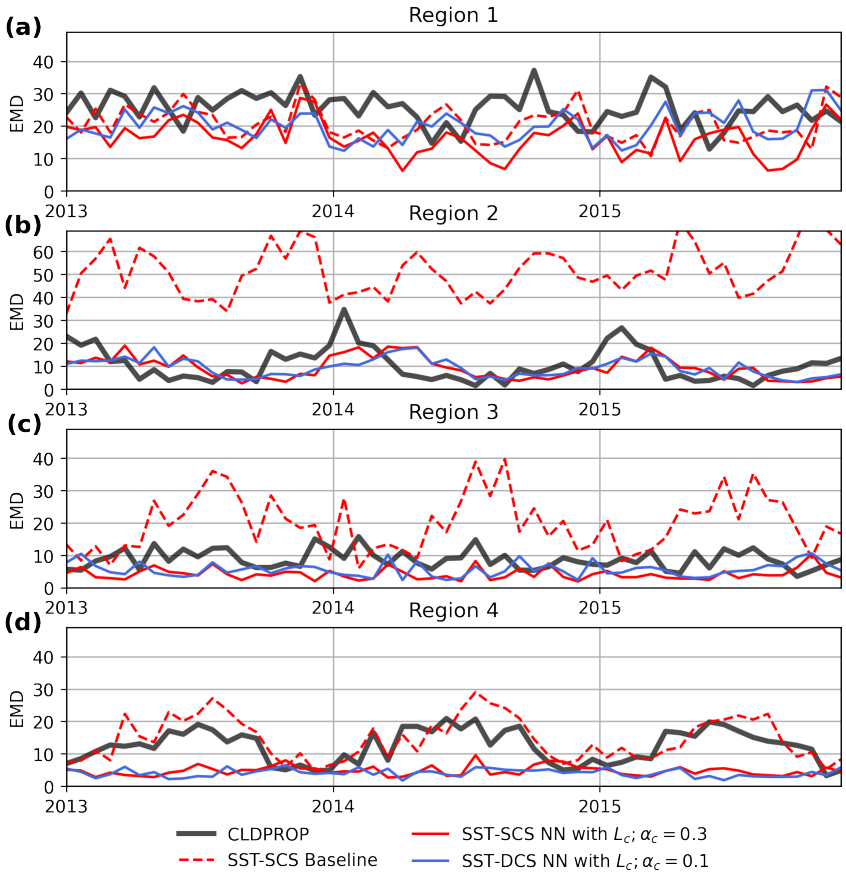

Figure 3.8: Comparison of CTP distributions for geographic regions 1 (a), 2 (b), 3(c) expressed in earth mover's distance for the SST experiments. See table 2 for coordinates of the geographic regions.

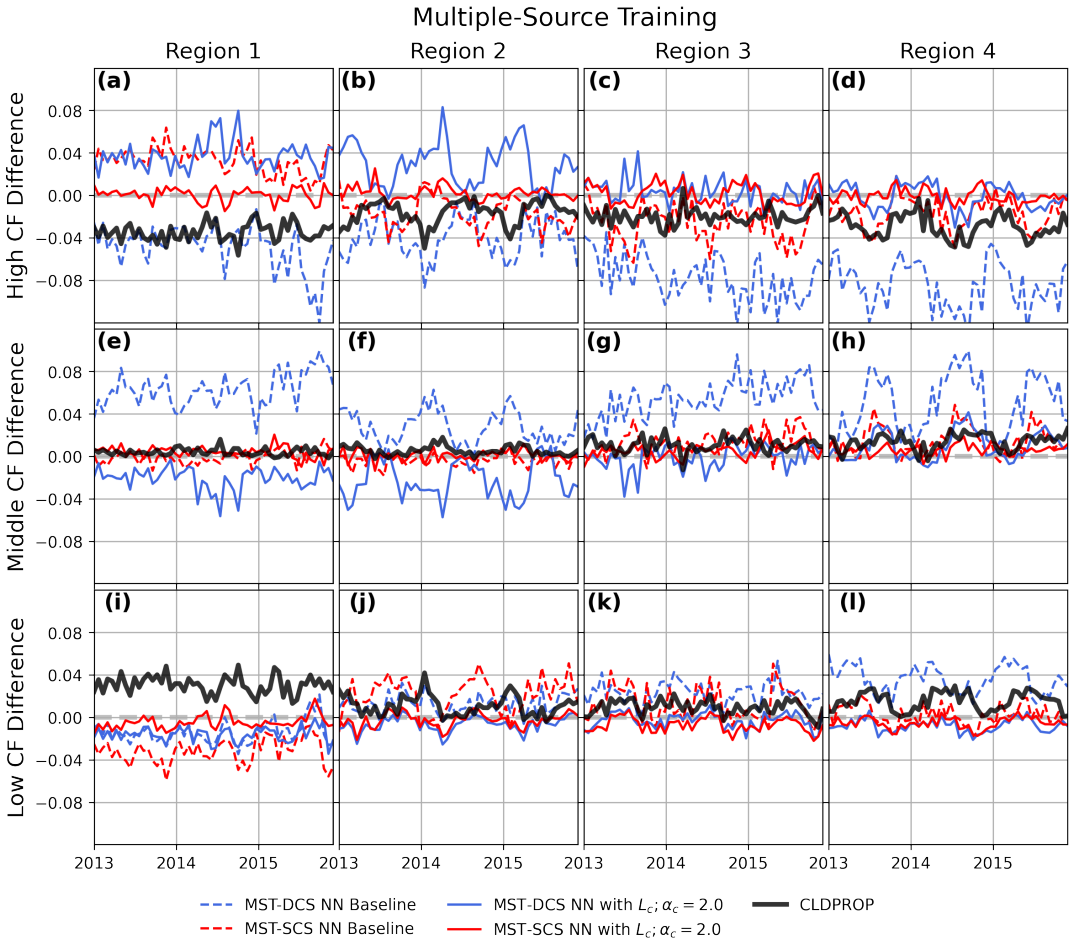

Figure 3.9: Differences in the frequency of (CF) of high (a), middle (b) and low (c) level clouds from 2013 to 2015 for the MST experiments. See Table 2 for the geographic coordinates of regions 1,2, and 3.

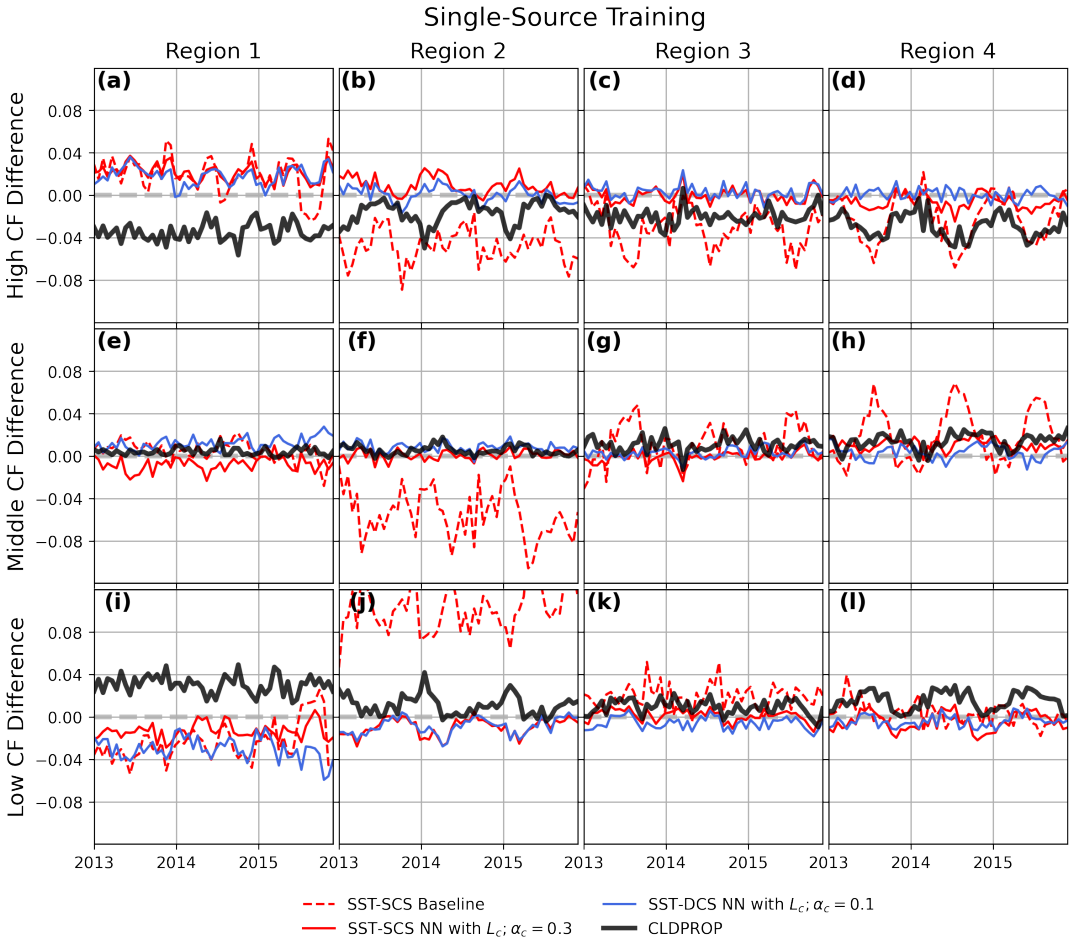

Figure 3.10: Differences in the frequency of (CF) of high (a), middle (b) and low (c) level clouds from 2013 to 2015 for the SST experiments. See Table 2 for the geographic coordinates of regions 1,2, and 3.

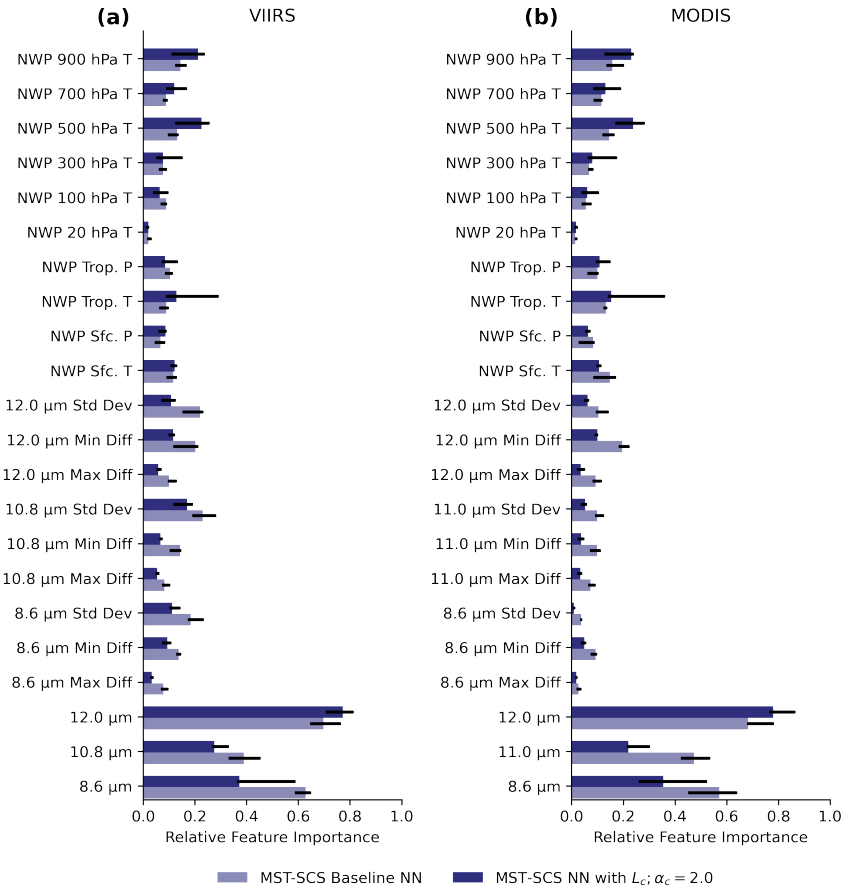

Figure 3.11: Relative feature importance of each of the feature in the neural networks for the MST-SCS experiment. Results are shown for MODIS (a) and VIIRS(b) over two values of α_c. Each value of α_c is tested over three different random initializations of the neural network represented by the error bars.

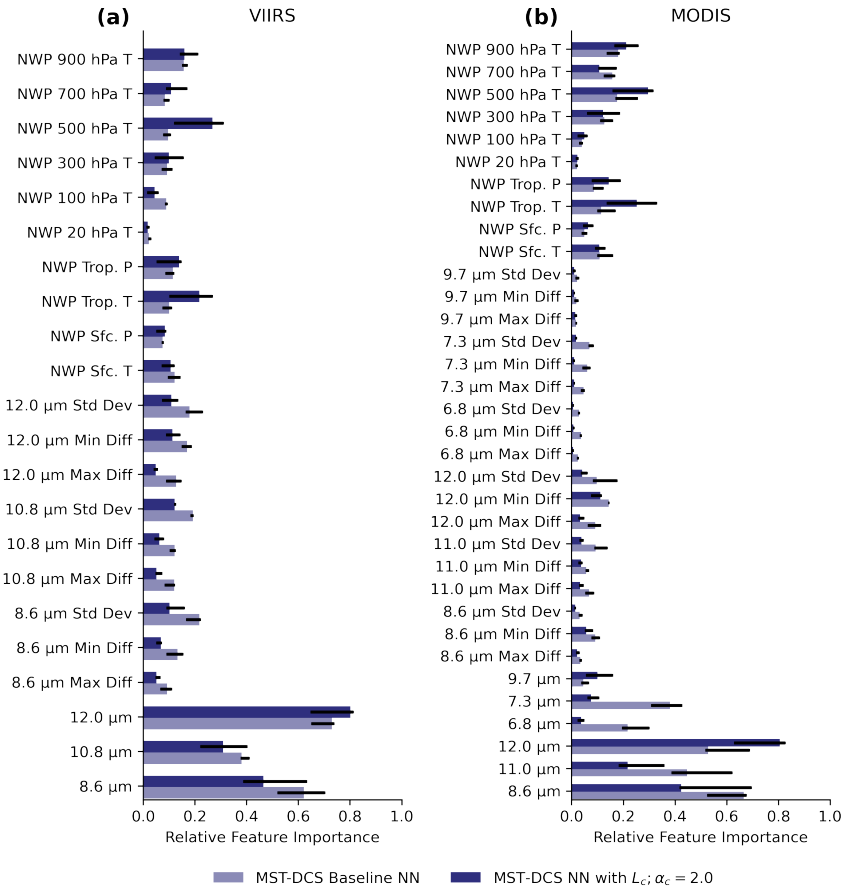

Figure 3.12: Relative feature importance of each of the feature in the neural networks for the MST-SCS experiment. Results are shown for MODIS (a) and VIIRS (b) comparing the baseline neural network to one trained with the consistency loss (L_c). Each model is tested over three neural networks with different randomly initialized weights.

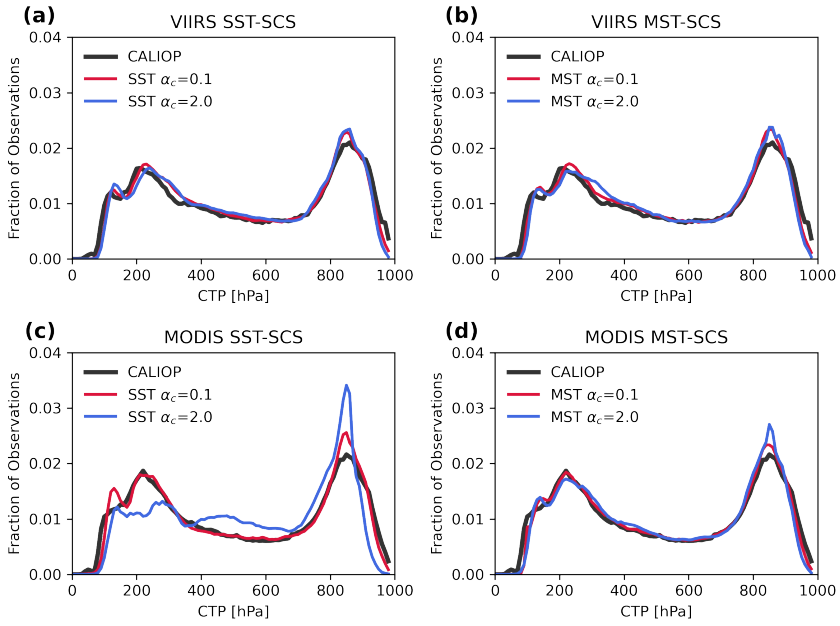

Figure 3.13: Distributions of CTP from the SST (a,c) and MST (b,d) experiments for VIIRS and MODIS. Histograms are calculated over bins with 10 hPa width.

REFERENCES

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016* 265–283. 1605.08695.

Ackerman, Steve, Frey Richard, Strabala Kathleen, Liu Yinghui, Gumley Liam, Baum Bryan, and Menzel Paul. 2010. Discriminating Clear-Sky from Cloud with MODIS, Algorithm Theoretical Basis Document (MOD35) - Version 6.1. Tech. Rep., NASA.

Ackerman, Steven A., R. E. Holz, R. Frey, E. W. Eloranta, B. C. Maddux, and M. McGill. 2008. Cloud Detection with MODIS. Part II: Validation. *J. Atmos. Ocean. Technol.* 25(7): 1073–1086.

Alin, Aylin. 2010. Multicollinearity. *WIREs Computational Statistics* 2(3):370–374.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. 1701.07875.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

Barnes, Elizabeth A, Benjamin Toms, James W Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. 2020. Indicator patterns of forced change learned by

an artificial neural network. *Journal of Advances in Modeling Earth Systems* 12(9): e2020MS002195.

Bessho, Kotaro, Kenji Date, Masahiro Hayashi, Akio Ikeda, Takahito Imai, Hidekazu Inoue, Yukihiro Kumagai, Takuya Miyakawa, Hidehiko Murata, Tomoo Ohno, Arata Okuyama, Ryo Oyama, Yukio Sasaki, Yoshio Shimazu, Kazuki Shimoji, Yasuhiko Sumida, Masuo Suzuki, Hidetaka Taniguchi, Hiroaki Tsuchiyama, Daisaku Uesawa, Hironobu Yokota, and Ryo Yoshida. 2016. An introduction to Himawari-8/9- Japan's new-generation geostationary meteorological satellites. *Journal of the Meteorological Society of Japan. Ser. II* 94(2):151–183.

Beucler, Tom, Imme Ebert-Uphoff, Stephan Rasp, Michael Pritchard, and Pierre Gentine. 2021. Machine Learning for Clouds and Climate (Invited Chapter for the AGU Geophysical Monograph Series "Clouds and Climate"). *Earth and Space Science Open Archive* 27.

Bhatt, Rajendra, David R. Doelling, Benjamin R. Scarino, Arun Gopalan, Conor O. Haney, Patrick Minnis, and Kristopher M. Bedka. 2016. A Consistent AVHRR Visible Calibration Record Based on Multiple Methods Applicable for the NOAA Degrading Orbits. Part I: Methodology. *Journal of Atmospheric and Oceanic Technology* 33(11):2499 – 2515.

Braun, Barbara Manganis, Theodore H. Sweetser, Clifford Graham, and Joseph Bartsch. 2019. CloudSat's A-Train Exit and the Formation of the C-Train: An Orbital Dynamics Perspective. In *2019 ieee aerospace conference*, 1–10.

Brenguier, Jean-Louis, Hanna Pawlowska, Lothar Schüller, Rene Preusker, Jürgen Fischer, and Yves Fouquart. 2000. Radiative Properties of Boundary Layer Clouds: Droplet

Effective Radius versus Number Concentration. *Journal of the Atmospheric Sciences* 57(6):803 – 821.

Bulgin, Claire E., Jonathan P.D. Mittaz, Owen Embury, Steinar Eastwood, and Christopher J. Merchant. 2018. Bayesian Cloud Detection for 37 Years of Advanced Very High Resolution Radiometer (AVHRR) Global Area Coverage (GAC) Data. *Remote Sens.* 10(1): 97.

Cannon, Alex J. 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment* 32(11):3207–3225.

Cao, Changyong, Frank J. De Luccia, Xiaoxiong Xiong, Robert Wolfe, and Fuzhong Weng. 2013. 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(2):1142–1156.

Cao, Changyong, Frank J. De Luccia, Xiaoxiong Xiong, Robert Wolfe, and Fuzhong Weng. 2014. Early On-Orbit Performance of the Visible Infrared Imaging Radiometer Suite Onboard the Suomi National Polar-Orbiting Partnership (S-NPP) Satellite. *IEEE Transactions on Geoscience and Remote Sensing* 52(2):1142–1156.

Chahine, Moustafa T. 1974. Remote sounding of cloudy atmospheres. I. The single cloud layer. *Journal of Atmospheric Sciences* 31(1):233–243.

Chase, Randy J, Stephen W Nesbitt, and Greg M McFarquhar. 2021. A Dual-Frequency Radar Retrieval of Two Parameters of the Snowfall Particle Size Distribution Using a Neural Network. *Journal of Applied Meteorology and Climatology* 60(3):341–359.

Cintineo, John L., Michael J. Pavolonis, Justin M. Sieglaff, and Daniel T. Lindsey. 2014. An Empirical Model for Assessing the Severe Weather Potential of Developing Convection. *Weather Forecast.* 29(3):639–653.

Cintineo, John L, Michael J Pavolonis, Justin M Sieglaff, Anthony Wimmers, Jason Brunner, and Willard Bellon. 2020. A Deep-Learning Model for Automated Detection of Intense Midlatitude Convection Using Geostationary Satellite Images. *Weather and Forecasting* 35(6):2567–2588.

Daniels, Jaime, Wayne Bresky, Steve Wanzong, Chris Velden, and Howard Berger. 2012. GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document For Derived Motion Winds - Version 2.5. Tech. Rep., NOAA NESDIS Center for Satellite Applications and Research.

Daoud, Jamal I. 2017. Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series* 949:012009.

Dawid, A. P. 1984. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)* 147(2):278–292.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1):27–46.

Düben, Peter, Umberto Modigliani, Alan Geer, Stephan Siemen, Florian Pappenberger, Peter Bauer, Andy Brown, Martin Palkovic, Baudouin Raoult, Nils Wedi, and Vasileios Baousis. 2021. Machine learning at ECMWF: A roadmap for the next 10 years (878).

Durand, Yannig, Pascal Hallibert, Mark Wilson, Mounir Lekouara, Semen Grabarnik, Donny Aminou, Paul Blythe, Bruno Napierala, Jean-Louis Canaud, Olivier Pigouche, Julien Ouaknine, and Bernard Verez. 2015. The flexible combined imager onboard MTG: from design to calibration 9639:1 – 14.

Farrar, Donald E., and Robert R. Glauber. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics* 49(1):92–107.

Foster, Michael J, and Andrew Heidinger. 2014. Entering the Era of +30-Year Satellite Cloud Climatologies: A North American Case Study. *Journal of Climate* 27(17):6687–6697.

Foster, Michael J., Andrew Heidinger, Michael Hiley, Steve Wanzong, Andi Walther, and Denis Botambekov. 2016. PATMOS-x Cloud Climate Record Trend Sensitivity to Reanalysis Products. *Remote Sensing* 8(5).

Frey, Richard A., Steven A. Ackerman, Robert E. Holz, Steven Dutcher, and Zach Griffith. 2020. The continuity MODIS-VIIRS cloud mask. *Remote Sens.* 12(20):1–18.

Håkansson, Nina, Claudia Adok, Anke Thoss, Ronald Scheirer, and Sara Hörnquist. 2018. Neural network cloud top pressure and height for MODIS. *Atmos. Meas. Tech.* 11(5): 3177–3196.

Hall, D. K., G. A. Riggs, N. E. DiGirolamo, and M. O. Román. 2019. Evaluation of MODIS and VIIRS cloud-gap-filled snow-cover products for production of an Earth science data record. *Hydrology and Earth System Sciences* 23(12):5227–5241.

Harris, Charles R., K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585(7825): 357–362.

Heidinger, A., Knapp K., Phillips C. Walther A., and Setngel M. 2021. Initial Plans for the Calibration and Data Content of the Next Generation of the International Satellite Cloud Climatology Project (ISCCP-NG). In *American meteorological society 2021 annual meeting*.

Heidinger, Andrew, and Yue Li. 2017. AWG Cloud Height Algorithm (ACHA) - Version 3.1. Tech. Rep., NOAA NESDIS Center for Satellite Applications and Research.

Heidinger, Andrew K, Nicholas Bearson, Michael J Foster, Yue Li, Steve Wanzong, Steven Ackerman, Robert E Holz, Steven Platnick, and Kerry Meyer. 2019. Using sounder data to improve cirrus cloud height estimation from satellite imagers. *Journal of Atmospheric and Oceanic Technology* 36(7):1331–1342.

Heidinger, Andrew K., Amato T. Evan, Michael J. Foster, and Andi Walther. 2012. A Naive Bayesian Cloud-Detection Scheme Derived from CALIPSO and Applied within PATMOS-x. *Journal of Applied Meteorology and Climatology* 51(6):1129 – 1144.

Heidinger, Andrew K., Michael J. Foster, Andi Walther, and Xuepeng (Tom) Zhao. 2014. The Pathfinder Atmospheres–Extended AVHRR Climate Dataset. *Bulletin of the American Meteorological Society* 95(6):909 – 922.

Heidinger, Andrew K., William C.Straka III, Christine C. Molling, Jerry T.Sullivan, and Xiangqian Wu. 2010. Deriving an inter-sensor consistent calibration for the AVHRR solar reflectance data record. *International Journal of Remote Sensing* 31(24):6493–6517. `https://doi.org/10.1080/01431161.2010.496472`.

Heidinger, Andrew K, and Michael J Pavolonis. 2009. Gazing at cirrus clouds for 25 years through a split window. Part I: Methodology. *Journal of Applied Meteorology and Climatology* 48(6):1100–1116.

Hilburn, Kyle A, Imme Ebert-Uphoff, and Steven D Miller. 2021. Development and Interpretation of a Neural-Network-Based Synthetic Radar Reflectivity Estimator Using

GOES-R Satellite Observations. *Journal of Applied Meteorology and Climatology* 60(1): 3–21.

Holz, R. E., S. A. Ackerman, F. W. Nagle, R. Frey, S. Dutcher, R. E. Kuehn, M. A. Vaughan, and B. Baum. 2008. Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP. *Journal of Geophysical Research: Atmospheres* 113(D8).

Hunter, J. D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 9(3):90–95.

Inoue, Toshiro. 1985. On the Temperature and Effective Emissivity Determination of Semi-Transparent Cirrus Clouds by Bi-Spectral Measurements in the 10μm Window Region. *Journal of the Meteorological Society of Japan. Ser. II* 63(1):88–99.

Jakob, Christian, and George Tselioudis. 2003. Objective identification of cloud regimes in the Tropical Western Pacific. *Geophysical Research Letters* 30(21).

Karlsson, K.-G., K. Anttila, J. Trentmann, M. Stengel, J. Fokke Meirink, A. Devasthale, T. Hanschmann, S. Kothe, E. Jääskeläinen, J. Sedlar, N. Benas, G.-J. van Zadelhoff, C. Schlundt, D. Stein, S. Finkensieper, N. Håkansson, and R. Hollmann. Clara-a2: the second edition of the cm saf cloud and radiation data record from 34 years of global avhrr data.

Karlsson, Karl Göran, Erik Johansson, Nina Håkansson, Joseph Sedlar, and Salomon Eliasson. 2020. Probabilistic Cloud Masking for the Generation of CM SAF Cloud Climate Data Records from AVHRR and SEVIRI Sensors. *Remote Sens.* 12(4):713.

Karlsson, Karl-Göran, and Abhay Devasthale. 2018. Inter-Comparison and Evaluation of the Four Longest Satellite-Derived Cloud Climate Data Records: CLARA-A2, ESA Cloud CCI V3, ISCCP-HGM, and PATMOS-x. *Remote Sensing* 10(10).

Kingma, Diederik P., and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd int. conf. learn. represent. iclr 2015 - conf. track proc.* International Conference on Learning Representations, ICLR. 1412.6980.

Kox, Stephan, Lucca Bugliaro, and A Ostler. 2014. Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing. *Atmospheric Measurement Techniques* 7(10):3233–3246.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25:1097–1105.

Kühnlein, Meike, Tim Appelhans, Boris Thies, and Thomas Nauß. 2014. Precipitation Estimates from MSG SEVIRI Daytime, Nighttime, and Twilight Data with Random Forests. *J. Appl. Meteorol. Climatol.* 53(11):2457–2480.

Lagerquist, Ryan, Amy McGovern, and David John Gagne II. 2019. Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting* 34(4): 1137–1160.

LeCun, Yann, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361(10):1995.

L'Ecuyer, Tristan S., and Jonathan H. Jiang. 2011. Touring the Atmosphere Aboard the A-Train. *AIP Conference Proceedings* 1401(1):245–256. https://aip.scitation.org/doi/pdf/10.1063/1.3653856.

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

Li, Fangjun, Xiaoyang Zhang, Shobha Kondragunta, and Ivan Csiszar. 2018. Comparison of Fire Radiative Power Estimates From VIIRS and MODIS Observations. *Journal of Geophysical Research: Atmospheres* 123(9):4545–4563.

Li, Y., B. A. Baum, A. K. Heidinger, W. P. Menzel, and E. Weisz. 2020. Improvement in cloud retrievals from VIIRS through the use of infrared absorption channels constructed from VIIRS+CrIS data fusion. *Atmospheric Measurement Techniques* 13(7):4035–4049.

Liu, Yinghui, Steven A. Ackerman, Brent C. Maddux, Jeffrey R. Key, and Richard A. Frey. 2010. Errors in Cloud Detection over the Arctic Using a Satellite Imager and Implications for Observing Feedback Mechanisms. *J. Clim.* 23(7):1894–1907.

Lundberg, Scott M, and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in neural information processing systems*, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30. Curran Associates, Inc.

Mamalakis, Antonios, Imme Ebert-Uphoff, and Elizabeth A. Barnes. 2021. Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. 2103.10005.

Marais, Willem, Robert Holz, Jeffrey Reid, and Rebecca Willett. 2020. Leveraging spatial textures, through machine learning, to identify aerosol and distinct cloud types from multispectral observations. *Atmos. Meas. Tech.* 13(10):1–35.

McGovern, Amy, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith. 2019. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society* 100(11):2175–2199.

McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Menzel, W Paul, Richard A Frey, Hong Zhang, Donald P Wylie, Chris C Moeller, Robert E Holz, Brent Maddux, Bryan A Baum, Kathy I Strabala, and Liam E Gumley. 2008. MODIS global cloud-top pressure and amount estimation: Algorithm description and results. *Journal of Applied Meteorology and Climatology* 47(4):1175–1198.

Meyer, Kerry, Steven Platnick, Robert Holz, Steve Dutcher, Greg Quinn, and Fred Nagle. 2020. Derivation of Shortwave Radiometric Adjustments for SNPP and NOAA-20 VIIRS for the NASA MODIS-VIIRS Continuity Cloud Products. *Remote Sensing* 12(24).

Minnis, Patrick, Gang Hong, Szedung Sun-Mack, William L. Smith, Yan Chen, and Steven D. Miller. 2016. Estimating nocturnal opaque ice cloud optical depth from MODIS

multispectral infrared radiances using a neural network method. *J. Geophys. Res.* 121(9): 4907–4932.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-wise relevance propagation: An overview*, 193–209. Cham: Springer International Publishing.

Nakajima, Teruyuki, and Michael D. King. 1990. Determination of the Optical Thickness and Effective Particle Radius of Clouds from Reflected Solar Radiation Measurements. Part I: Theory. *Journal of Atmospheric Sciences* 47(15):1878 – 1893.

Noh, Yoo-Jeong, John M. Forsythe, Steven D. Miller, Curtis J. Seaman, Yue Li, Andrew K. Heidinger, Daniel T. Lindsey, Matthew A. Rogers, and Philip T. Partain. 2017. Cloud-Base Height Estimation from VIIRS. Part II: A Statistical Algorithm Based on A-Train Satellite Data. *Journal of Atmospheric and Oceanic Technology* 34(3):585 – 598.

Oreopoulos, Lazaros, Nayeong Cho, and Dongmin Lee. 2017. Using MODIS cloud regimes to sort diagnostic signals of aerosol-cloud-precipitation interactions. *Journal of Geophysical Research: Atmospheres* 122(10):5416–5440.

Panaretos, Victor M., and Yoav Zemel. 2019. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application* 6(1):405–431.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot,

and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85):2825–2830.

Pfreundschuh, S., P. Eriksson, D. Duncan, B. Rydberg, N. Håkansson, and A. Thoss. 2018. A neural network approach to estimating a posteriori distributions of Bayesian retrieval problems. *Atmospheric Measurement Techniques* 11(8):4627–4643.

Platnick, Steven, Kerry Meyer, Galina Wind, Robert E. Holz, Nandana Amarasinghe, Paul A. Hubanks, Benjamin Marchant, Steven Dutcher, and Paolo Veglio. 2021. The NASA MODIS-VIIRS Continuity Cloud Optical Properties Products. *Remote Sensing* 13(1).

Platnick, Steven, Kerry G. Meyer, Michael D. King, Galina Wind, Nandana Amarasinghe, Benjamin Marchant, G. Thomas Arnold, Zhibo Zhang, Paul A. Hubanks, Robert E. Holz, Ping Yang, William L. Ridgway, and Jerome Riedi. 2017. The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua. *IEEE Trans. Geosci. Remote Sens.* 55(1):502–525.

Popp, Thomas, Michaela I. Hegglin, Rainer Hollmann, Fabrice Ardhuin, Annett Bartsch, Ana Bastos, Victoria Bennett, Jacqueline Boutin, Carsten Brockmann, Michael Buchwitz, Emilio Chuvieco, Philippe Ciais, Wouter Dorigo, Darren Ghent, Richard Jones, Thomas Lavergne, Christopher J. Merchant, Benoit Meyssignac, Frank Paul, Shaun Quegan, Shubha Sathyendranath, Tracy Scanlon, Marc Schröder, Stefan G. H. Simis, and Ulrika Willén. 2020. Consistency of Satellite Climate Data Records for Earth System Monitoring. *Bulletin of the American Meteorological Society* 101(11):E1948 – E1971.

Poulsen, C. A., R. Siddans, G. E. Thomas, A. M. Sayer, R. G. Grainger, E. Campmany, S. M. Dean, C. Arnold, and P. D. Watts. 2012. Cloud retrievals from satellite data using optimal estimation: evaluation and application to ATSR. *Atmospheric Measurement Techniques* 5(8):1889–1910.

Rodgers, Clive D. 1976. Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Reviews of Geophysics* 14(4):609–624.

Rumelhart, D., G Hinton, and R Williams. 1986. Learning representations by back-propagating errors. *Nature* (323):533–536.

Saha, Suranjana, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, et al. 2014. The NCEP climate forecast system version 2. *Journal of climate* 27(6):2185–2208.

Schiffer, R. A., and W. B. Rossow. 1983. The International Satellite Cloud Climatology Project (ISCCP): The first project of the World Climate Research Programme. *Bull. Amer. Meteorol. Soc.* 64:779–784.

Schmit, Timothy J, Paul Griffith, Mathew M Gunshor, Jaime M Daniels, Steven J Goodman, and William J Lebair. 2017. A closer look at the ABI on the GOES-R series. *Bulletin of the American Meteorological Society* 98(4):681–698.

Seaman, Curtis J, Yoo-Jeong Noh, Steven D Miller, Andrew K Heidinger, and Daniel T Lindsey. 2017. Cloud-base height estimation from VIIRS. Part I: Operational algorithm validation against CloudSat. *Journal of Atmospheric and Oceanic Technology* 34(3): 567–583.

Shapley, Lloyd S. 1953. *17. A value for n-person games*. Princeton University Press.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* 6(1):1–48.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. 1605.01713.

Skakun, Sergii, Christopher O. Justice, Eric Vermote, and Jean-Claude Roger. 2018. Transitioning from MODIS to VIIRS: an analysis of inter-consistency of NDVI data sets for agricultural monitoring. *International Journal of Remote Sensing* 39(4):971–992. PMID: 29892137.

Smith, Leslie N. 2017. Cyclical Learning Rates for Training Neural Networks. In *Proc. - 2017 ieee winter conf. appl. comput. vision, wacv 2017*, 464–472. Institute of Electrical and Electronics Engineers Inc. 1506.01186.

Smith, WL, and CMR Platt. 1978. Comparison of satellite-deduced cloud heights with indications from radiosonde and ground-based laser measurements. *Journal of Applied meteorology* 17(12):1796–1802.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15(56):1929–1958.

Stengel, M., S. Stapelberg, O. Sus, S. Finkensieper, B. Würzler, D. Philipp, R. Hollmann, C. Poulsen, M. Christensen, and G. McGarragh. 2020. Cloud_cci Advanced Very High Res-

olution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties. *Earth System Science Data* 12(1):41–60.

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

Strabala, Kathleen I., Steven A. Ackerman, and W. Paul Menzel. 1994. Cloud Properties inferred from 8-12-μm Data. *Journal of Applied Meteorology and Climatology* 33(2):212 – 229.

Stubenrauch, C. J., W. B. Rossow, S. Kinne, S. Ackerman, G. Cesana, H. Chepfer, L. Di Girolamo, B. Getzewich, A. Guignard, A. Heidinger, B. C. Maddux, W. P. Menzel, P. Minnis, C. Pearl, S. Platnick, C. Poulsen, J. Riedi, S. Sun-Mack, A. Walther, D. Winker, S. Zeng, and G. Zhao. 2013. Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel. *Bull. Am. Meteorol. Soc.* 94(7): 1031–1049.

Sus, O., M. Stengel, S. Stapelberg, G. McGarragh, C. Poulsen, A. C. Povey, C. Schlundt, G. Thomas, M. Christensen, S. Proud, M. Jerg, R. Grainger, and R. Hollmann. 2018. The Community Cloud retrieval for CLimate (CC4CL) – Part 1: A framework applied to multiple satellite imaging sensors. *Atmospheric Measurement Techniques* 11(6):3373–3396.

Thampi, Bijoy Vengasseril, Takmeng Wong, Constantin Lukashin, and Norman G. Loeb. 2017. Determination of CERES TOA Fluxes Using Machine Learning Algorithms. Part

I: Classification and Retrieval of CERES Cloudy and Clear Scenes. *J. Atmos. Ocean. Technol.* 34(10):2329–2345.

Toms, Benjamin A., Elizabeth A. Barnes, and Imme Ebert-Uphoff. 2020. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems* 12(9):e2019MS002002.

Uddstrom, Michael J, Warren R Gray, Richard Murphy, Niles A Oien, and Talbot Murray. 1999. A Bayesian Cloud Mask for Sea Surface Temperature Retrieval. *J. Atmos. Ocean. Technol.* 16(1):117–132.

Uprety, Sirish, Changyong Cao, Xiaoxiong Xiong, Slawomir Blonski, Aisheng Wu, and Xi Shao. 2013. Radiometric Intercomparison between Suomi-NPP VIIRS and Aqua MODIS Reflective Solar Bands Using Simultaneous Nadir Overpass in the Low Latitudes. *Journal of Atmospheric and Oceanic Technology* 30(12):2720 – 2736.

Vaughan, Mark A., Kathleen A. Powell, David M. Winker, Chris A. Hostetler, Ralph E. Kuehn, William H. Hunt, Brian J. Getzewich, Stuart A. Young, Zhaoyan Liu, and Matthew J. McGill. 2009. Fully Automated Detection of Cloud and Aerosol Layers in the CALIPSO Lidar Measurements. *Journal of Atmospheric and Oceanic Technology* 26(10):2034 – 2050.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng,

Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17:261–272.

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. *Atmospheric Measurement Techniques* 13(5):2257–2277.

Watts, P. D., R. Bennartz, and F. Fell. 2011. Retrieval of two-layer cloud properties from multispectral observations using optimal estimation. *Journal of Geophysical Research: Atmospheres* 116(D16).

Weisz, Elisabeth, Bryan A. Baum, and W. Paul Menzel. 2017. Fusion of satellite-based imager and sounder data to construct supplementary high spatial resolution narrowband IR radiances. *Journal of Applied Remote Sensing* 11(3):1 – 14.

Welch, R. M., S. K. Sengupta, A. K. Goroch, P. Rabindra, N. Rangaraj, and M. S. Navar. 1992. Polar Cloud and Surface Classification Using AVHRR Imagery: An Intercomparison of Methods. *J. Appl. Meteorol.* 31(5):405–420.

Wheeler, David, and Michael Tiefelsdorf. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7(2):161–187.

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. *Atmospheric Measurement Techniques* 14(5):3371–3394.

Wimmers, Anthony, Christopher Velden, and Joshua H. Cossuth. 2019. Using Deep Learning to Estimate Tropical Cyclone Intensity from Satellite Passive Microwave Imagery. *Mon. Weather Rev.* 147(6):2261–2282.

Winker, David M., Mark A. Vaughan, Ali Omar, Yongxiang Hu, Kathleen A. Powell, Zhaoyan Liu, William H. Hunt, and Stuart A. Young. 2009. Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms. *Journal of Atmospheric and Oceanic Technology* 26(11):2310 – 2323.

Xiong, Xiaoxiong, Amit Angal, Tiejun Chang, Kwofu Chiang, Ning Lei, Yonghong Li, Junqiang Sun, Kevin Twedt, and Aisheng Wu. 2020. MODIS and VIIRS Calibration and Characterization in Support of Producing Long-Term High-Quality Data Products. *Remote Sensing* 12(19).

Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. MixUp: Beyond empirical risk minimization. In *6th int. conf. learn. represent. iclr 2018 - conf. track proc.*