A Neural Network-based Cloud Mask for PREFIRE and Evaluation with Simulated Observations

Cameron D. Bertossa

A Thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

(Atmospheric and Oceanic Sciences)

at the

University of Wisconsin-Madison

August 2022

Abstract

A Neural Network-based Cloud Mask for PREFIRE and Evaluation with Simulated Observations

by Cameron D. Bertossa

The Polar Radiant Energy in the Far InfraRed Experiment (PREFIRE) will fill a gap in our understanding of polar processes and the polar climate by offering widespread, spectrally-resolved measurements through the Far InfraRed (FIR) with two identical CubeSat spacecraft. While the polar regions are typically difficult for skillful cloud identification due to cold surface temperatures, the reflection by bright surfaces, and frequent temperature inversions, the inclusion of the FIR may offer increased spectral sensitivity, allowing for the detection of even thin ice clouds. This study assesses the potential skill, as well as limitations, of a neural network-based cloud mask using simulated spectra mimicking what the PREFIRE mission will capture. Analysis focuses on the polar regions. Clouds are found to be detected approximately 90% of time using the derived neural network. The NN's assigned confidence for whether a scene is 'clear' or 'cloudy' proves to be a skillful way in which quality flags can be attached to predictions. Clouds with higher cloud top heights are typically more easily detected. Low-altitude clouds over polar surfaces, which are the most difficult for the NN to detect, are still detected over 80% of the time. The FIR portion of the spectrum is found to increase the detection of clear scenes and increase mid-to-high altitude cloud detection. Cloud detection skill improves through the use of the overlapping fields of view produced by the PREFIRE

instrument's sampling strategy. Overlapping fields of view increase accuracy relative to the baseline NN while simultaneously predicting on a sub-FOV scale.

Acknowledgements

I would like to thank my advisor, Tristan L'Ecuyer, for his support. He has allowed me to fully explore my research interests and has been patient as I have stumbled along the way; much of this work would not have been possible without him. I would also like to thank Aronne Merrelli, Xianglei Huang and Xiuhong Chen for their help preparing the manuscript and offering additional research advisement. Finally, thank you to Elizabeth Maroon and Grant Petty, my two masters committee members, for their feedback on this thesis. Both of these faculty members very clearly care about their students and are willing to help inside and outside of the classroom.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Polar Radiant Energy in the Far InfraRed Experiment (PREFIRE) will fill a major gap in the knowledge of the Arctic energy budget and the role of Far InfraRed (FIR) radiation in Arctic warming, sea ice loss, ice sheet melt, and sea level rise (L'Ecuyer et al., 2021). PREFIRE will make the first spectrally-resolved measurements of FIR radiation, which comprises 60% of Arctic emission. The mission consists of two 6U ("U" for units, where each unit indicates a standard size of 10 cm x 10 cm x 10 cm) CubeSats in different 470-540 km altitude, near-polar (98^o inclination) orbits, each carrying a miniaturized Thermal Infrared Spectrometer (TIRS), covering 5-54 μm with a spectral sampling of 0.86 μm and an effective field of view between 12 and 15 km. To support PREFIRE's objectives, a variety of geophysical and radiative property algorithms are being developed to improve our study of the polar regions. A critical step in the execution of these algorithms is the delineation of clear and cloudy fields of view (FOV) to inform subsequent geophysical retrievals of the scene being viewed. This paper describes a multi-spectral cloud mask for this purpose which utilizes the complete spectral range that will be offered by PREFIRE.

Traditionally, clouds over the polar regions are especially hard to detect due to their similar brightness and emission temperatures compared to the polar surface; cloud masks commonly have a steep degradation in performance in these regions (Liu et al., 2010, Lubin and Morrow, 1998, Raschke, 1987). However, there is evidence that the use of the FIR may offer particular skill in the detection of thin ice clouds which have a lower emitting temperature and ice particles of similar size to the FIR wavelength, allowing for

greater spectral sensitivity (Yang et al., 2003). Furthermore, FIR spectra exhibit distinct signatures for polar surfaces (Huang et al., 2016), meaning clear-sky detection may be aided with its use. These two factors may lead to improvement in cloud detection for PREFIRE in this difficult region. There have been a number of recent studies which have looked at the effect that the FIR may have in more skillful cloud retrievals, whether those observations be ground-based or from a satellite (Cossich et al., 2021, Maestri et al., 2019, Saito et al., 2020).

Improved cloud detection in polar regions addresses PREFIRE's objective of improving the characterization of the polar energy budget in several ways. Clouds modulate both incoming shortwave as well as outgoing longwave radiation, leading to strong influences on local weather and climate (Crane and Barry, 1984). Cloud microphysical properties (such as particle size and concentration) and macrophysical properties (such as cloud top height and cloud thickness) each influence its spectral signature throughout the mid and far-infrared. While the broadband effect of clouds is better understood, the signal of clouds at particular IR wavelengths, especially those longer than 15 μm (the FIR), is less clear. This lack of understanding is due to both a lack of observations (L'Ecuyer et al., 2021, Palchetti et al., 2015), as well as water vapor's tendency to be strongly absorbing in the FIR, making the atmosphere relatively opaque when viewing it from the top-down. However, dry environments, such as the polar regions or high altitude locations, allow for greater use of the FIR. PREFIRE aims to exploit this fact to improve the understanding of polar cloud distributions, processes and climate impacts.

The first step to studying and understanding cloud effects on the radiation budget is identifying where they are actually present– this is typically a critical component to any Earth observing mission's algorithm chain. The detection of true clear scenes will determine where surface property and temperature or humidity profile algorithms should be performed, whereas cloudy scenes will be used by those trying to extract cloud microphysical or macrophysical information. The misidentification can lead to increased uncertainty or large errors in the subsequent algorithms. These errors can then propagate further downstream or pose direct problems for users. This is especially true for clear sky retrievals, which, in many cases, can be sensitive to even very thin clouds. In the case of PREFIRE, the mission has an anticipated miss/false detection rate of 10% for clouds, and as such, this benchmark should be kept in mind throughout the study.

Systematically measuring the 'true' surface emission in the Arctic and Antarctic, via the correct identification of clear scenes, has important implications not only for understanding the present state of the climate but also for predicting its evolution. The longwave fluxes of the different characteristic surface types within the polar regions are vital to understanding the surface energy balance and thus the rate at which the surface warms and ice melts, among other processes. However, due to limited observation, estimates of Arctic emission vary by up to 70 W/m^2 between sources with large uncertainty (L'Ecuyer et al., 2021). Being unable to measure the true radiative properties of various surface types does not allow us to then model how they will evolve in a future climate. Correctly identifying 'clear' scenes with little error is vital to PREFIRE's contribution to this important uncertainty.

Many traditional cloud masks do not use a mission's full spectrum of channels, but rather, decision trees based on a subset of channels. The cloud mask for the Moderate Resolution Imaging Spectroradiometer (MODIS), for example, uses a common technique of taking the differences between pairs of brightness temperatures to detect various cloud types (Ackerman et al., 2008, Frey et al., 2008). However, different MODIS channels should be used depending on the type of cloud or environment the cloud mask is being used for. This methodology can result in increased uncertainty in the polar regions, where the surface can be of a similar temperature (or often even warmer with the presence of environmental temperature inversions) to that of a cloud top.

However, the spectral variation across many mid and far-infrared channels may offer additional capability for distinguishing clouds from polar surfaces. As computational power has increased and machine learning (ML) techniques have become more prevalent, the use of a greater number of spectral channels to aid in cloud detection has become computationally practical. ML offers a statistically based method to extract physical patterns from data. For classification problems, these algorithms generally work by taking a set of inputs which are already labeled, going through a training procedure, and then evaluating the skill of the model on a separate test set. These algorithms can be designed in a variety of ways, however, they are typically computationally efficient. This makes their evaluation on large data sets (a common requirement for atmospheric and oceanic studies) viable. A neural network (NN) is a specific type of ML algorithm comprised of many nodes whose information is passed through the network in some nonlinear way. These nodes are optimized in such a way that the neural network 'learns' based on

the minimization of a loss function (Krogh, 2008). This loss function can be defined based on the particular problem at hand. The use of NNs has grown in popularity due to their ability to model nonlinear relationships between inputs and expected outputs. The nonlinear nature of the atmospheric system makes NNs especially useful in their application to weather and climate (Mekanik et al., 2013, Niebler et al., 2021, Wimmers et al., 2019). More specifically, the ability of NNs to detect patterns in multivariate data allow them to offer potential improvements in cloud mask skill by identifying unique spectral signatures of clouds and surfaces. There are many recent examples of cloud masks that use some form of ML algorithm (Jeppesen et al., 2019, Liu et al., 2021, Maestri et al., 2019, Paul and Huntemann, 2020, Wang et al., 2020). The skill to be gained in ML methods versus traditional methods may be found in such references. However, due to the novelty of the FIR measurements that PREFIRE will offer, in this study we choose to not compare traditional methods (like a split-window brightness temperature difference, which does not incorporate the FIR) to the methods proposed.

This manuscript assesses the potential skill, as well as the limitations, of a NN-based cloud mask for the PREFIRE mission. Analysis will be focused on the difficult polar regions (poleward of 60^o). The manuscript is organized as follows: Section 2 discusses the simulation of the training and testing data and the structure of the basic NN cloud mask. Section 3 presents how this cloud mask performs in the evaluation of synthetic data. Section 4 builds upon this baseline NN cloud mask with the addition of model variables and more complex structuring. Section 5 concludes and contains some discussion. While a true validation of skill is not possible until post-launch, the advantages of different techniques and the potential of the FIR may be evaluated throughout this exploratory study.

Chapter 2

Simulation of Dataset and Neural Network Training Process

2.1 Training Dataset

Since there are not a sufficient number of collocated cloud and FIR observations to train a NN, the algorithm presented here is trained and tested using a high spatial resolution simulation from the Geophysical Fluid Dynamics Laboratory (GFDL). The training dataset was developed using GFDL FV3 model output that was part of the first intercomparison project of global storm-resolving models, i.e., the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains (DYAMOND) (Stevens et al., 2019). It simulated weather from August 1 to September 10 in 2016 with a global horizontal resolution of 3km by 3km and 79 vertical levels up to 39km above the surface. The FV3 model has a non-hydrostatic dynamic core (Lin, 2004) that resolves the deep convection at this horizontal resolution. The shallow cumulus convection was still parameterized and a turbulent kinetic energy (TKE)-like scheme was used for the boundary layer parameterization. The FV3 model can skillfully simulate observed cloud system evolution. Further details of the FV3 model configuration and performance can be found in Stevens et al. (2019). While this is typically a tropic and midlatitude targeted model, it still simulates conditions in the polar regions– allowing us to test a variety of scene types. Hereafter, for brevity, the simulation output will be referred to as the GFDL data.

Based on the GFDL data, the upwelling spectral radiance for each gridbox at 3 UTC on August 1, 2016 was computed using a satellite radiance simulator developed by Chen et al. (2013). The simulator is based on the Principal Component-based Radiative Transfer Model (PCRTM; (Liu et al., 2006)) and was designed to interface the PCRTM with reanalyses and numerical model output in a flexible way. It also incorporated surface spectral emissivity as in Huang et al. (2016) in all the calculations. The same radiance simulator has been used before in a variety of IR radiance studies (e.g., Aumann et al., 2018, Bantges et al., 2016, Huang et al., 2014, Peterson et al., 2019). For clear-sky, the PCRTM utilizes temperature, H_2O, and O_3 profiles in 101 levels from the surface to 0.005 hPa. GFDL's profiles extend from the surface to 1 hPa, and are interpolated to the PCRTM levels. Profiles are extended to 0.005 hPa using standard profiles from McClatchey (1972). The profile used depends on the latitude and season. The CH_4, CO, and N_2O concentrations are fixed using U.S. 1976 Standard Atmosphere profile. The CO_2 concentration is similar to CH_4, CO, and N_2O, but is also scaled using ground observation values in August 2016 from Tans and Keeling (2022).

For cloudy-sky calculations, the PCRTM also needs cloud phase, cloud optical depth and cloud effective size for each level with clouds. Given the high spatial resolution of the GFDL fields, any profile is cloudy if the cloud water content (liquid + ice) is not zero at any level. Note that the cloud fraction is assumed to be 100% in these calculations, which is not unreasonable given the small gridboxes in comparison to the TIRS FOV. Clouds above 440 hPa are deemed as ice clouds in PCRTM, clouds below 440 hPa are deemed as water clouds. Liquid and ice cloud optical depths are computed separately. The optical depth of each is computed based on the liquid and ice water content (Ebert and Curry, 1992, Fouquart, 1988), respectively, where an effective cloud size of 20 μm is used for liquid clouds and a temperature-dependent value is used for ice clouds (Ou and Liou, 1995). Then, water and ice cloud optical depth are summed together and input into PCRTM. Surface spectral emissivity variations over the entire longwave, including the FIR region, are also taken into account in the PCRTM calculation. Typical surface emissivities (e.g., mostly ice and water over polar region) are adapted from Huang et al. (2016). For sea ice regions, surface emissivity is a weighted average of ice and water emissivity based on sea ice fraction from NSIDC (NSIDC, 2022). Figure 2.1 depicts four of the fields produced by the GFDL simulation that are used in PCRTM to compute the gridbox upwelling spectral radiance.

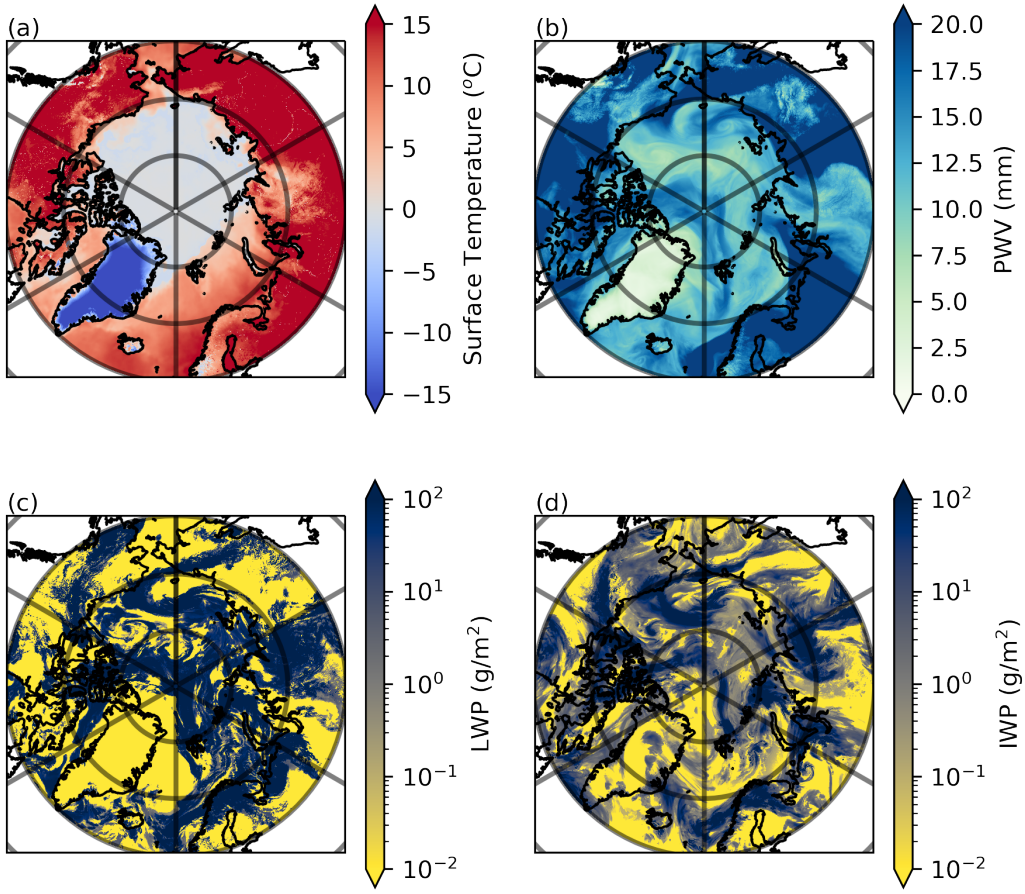

FIGURE 2.1: Four example GFDL fields that go into calculating upwelling radiance using PCRTM: (a) Surface temperature, (b) Precipitable water vapor (PWV), (c) Liquid water path (LWP), (d) Ice water path (IWP). Everything except the northern hemisphere polar region (defined as poleward of 60^oN) is masked out.

From the PCRTM radiances, which contain 5421 spectral channels whose centers range from 50 cm^{-1} to 2760 cm^{-1} with an even spacing of 0.5 cm^{-1}, 54 valid TIRS channels are derived for each gridbox using the appropriate spectral response functions. However, the two remaining shortest wavelength channels (central wavelengths at 3.37 μm and 4.42 μm) will have contribution from reflected solar radiation. Since the PCRTM radiance calculations do not include any reflected solar radiation, those two channels will not be realistically simulated. In addition, our intention is that the cloud mask method should not have any diurnal dependence, which implies that the utilized channels should be free from reflected solar radiation; thus, these two channels are removed as inputs for this study, resulting in a 52-channel spectrum for each 3km-by-3km GFDL gridbox.

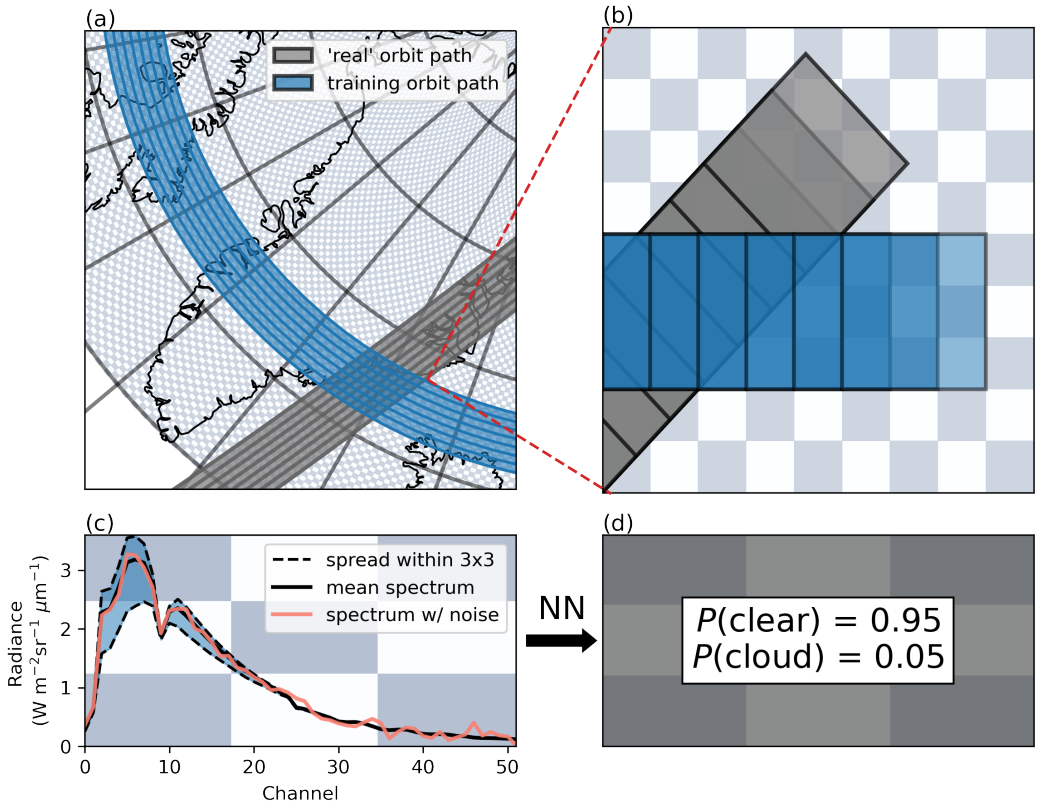

FIGURE 2.2: (a) An example orbit path of a PREFIRE CubeSat (gray) versus what is chosen to be representative of an orbit for the NN training and evaluation procedure (dark blue). The blue-and-white checkerboard represents the GFDL grid size. Note that training orbit paths are simulated to cover the entirety of both polar regions. (b) Zooming in on one of the along-track paths. Large grey boxes indicates the orientation of a simulated orbit path's footprints, while the large blue box indicates the size and orientation of the simulated training footprints. FOV overlap 2/3 with one another, mimicking the TIRS instrument sampling time. (c) An example mean spectrum representative of an aggregated 3-by-3 GFDL 'patch' (large blue box shown in (b)). Orange represents that spectrum after estimates of channel noise has been added to it. (d) The final output of a trained NN, returning the probability that the scene is clear versus cloudy.

While there are assumptions and approximations inherent the surface emissivity models, and in the PCRTM itself, we believe they are the most accurate available given the state of knowledge. The surface emissivity spectra are derived from physical models that are accurate in the mid-IR (Chen et al., 2014). Since the spectral shape of surface emissivity is fundamentally determined by the spectral shape of the refractive index and scattering model used to compute reflection of such dense medium, accuracy of modeled

mid-IR surface emissivity can give us more confidence for the accuracy of modeled far-IR surface emissivity. Of course, the far-IR surface emissivities eventually need to be validated against observations, which is one of the main objectives for the entire PREFIRE mission. The PCRTM has also been shown to accurately reproduce spectral radiance from a reference line-by-line radiative transfer model (Chen et al., 2014).

In order to simulate the inhomogeneity that may be present within a TIRS scene, 3-by-3 'patches' of neighboring GFDL gridboxes are aggregated and the channel-by-channel radiances are averaged. This procedure results in a mean 52-channel spectrum representative of 9 different GFDL gridboxes (Fig. 2.2(c)). As will be discussed more extensively later on, the accuracy of the model is not significantly affected by the number of gridboxes aggregated within these patches (for example, 9 versus 16). Noise is then added to each channel drawing from a Gaussian distribution whose mean is 0 and standard deviation is equal to that of the expected TIRS channel noise. Estimates of instrument noise for each channel as well as the TIRS spectral response functions can be found in Fig. B.1.

Since the PREFIRE mission is polar targeted, this cloud mask will focus on cloud detection in the polar regions specifically (defined as poleward of 60^o). Thus, only those GFDL gridboxes within this region are used for model training and evaluation (refer to Fig. 2.1 and Fig. 2.2(a) for a reference of the size and spatial inhomogeneity within the northern hemisphere portion of this domain). A sliding window is adopted such that two-thirds of a coarsened FOV will contain the same GFDL gridboxes as its neighbor to the west or to the east (blue in Fig. 2.2(b)), see Fig. B.2 for another representation of this sliding window. This sliding window represents the TIRS detector read timing, where the read out time is one-third the time length it takes the satellite to move one whole scene footprint. While PREFIRE's orbit paths will cross lines of latitude and longitude (gray in Fig. 2.2(a)), the training FOV are chosen to be oriented such that the aggregated patches share vertices with the native GFDL grid (blue in Fig. 2.2(b)). This allows one to fully exploit the polar domain of the GFDL grid; ultimately resulting in approximately $8.4\text{x}10^6$ different (but not independent, due to overlap) spectra for the model to use during training and evaluation. The difference in orientation of the training swath relative to that of the planned PREFIRE orbit has no impact on the algorithm or results. Since the GFDL simulation uses an evenly spaced latitude and longitude grid, the GFDL gridboxes will be smaller near the poles. This results in physically smaller aggregated patches and a general over-representation of the simulated conditions that occur near the poles. However, the

coarsening procedure still simulates FOV inhomogeneity and the oversampling is found to not significantly affect the results (for example, evaluation accuracy is not correlated with latitude). Furthermore, while only one timestep from the FV3 simulation is used for training (due to computational limitations), there is still a large variety of scenes sampled since the simulation covers both poles, with relatively fine resolution gridboxes. However, more variability in training will be added in the future.

Finally, in this study we ignore the possible effects of calibration errors, which will degrade the cloud mask skill in some way. While there is a rigorous calibration plan in place (see L'Ecuyer et al., 2021), the effect of unknown calibration errors is better suited for a post-launch study since these errors can take a variety of forms.

2.2 Neural Network Environment and Formation

Tensorflow's Keras v2.3.1 is used to create, train and evaluate the NN (Chollet et al., 2015). Keras offers the ability to easily set up and modify the structure of a NN. A neural network 'learns' by minimizing a particular user-defined loss function. The loss function selected for the training of this study's neural networks is balanced binary cross-entropy (Lin et al., 2017, Ramos et al., 2018). Defined as:

$$\text{Loss} = -\frac{1}{|\mathbf{Y}^*|}\sum\left[(1-\beta)\mathbf{Y}^*\log_{10}(\hat{\mathbf{Y}}) + \beta(1-\mathbf{Y}^*)(\log_{10}(1-\hat{\mathbf{Y}}))\right] \tag{2.1}$$

where $\mathbf{Y}^*$ is a binary set representing 'truth'. Clear is defined to be the null condition (0) and cloudy the alternate condition (1). $\hat{\mathbf{Y}}$ is the corresponding set of predicted probabilities by the NN for the cloudy class. Finally, $\beta = \frac{\sum \mathbf{Y}^*}{|\mathbf{Y}^*|}$, adjusts for imbalances between the number of cloudy versus clear scenes in the training set. For each individual prediction, the first term in Eq. 2.1 is equal to 0 if the scene is clear and the second term is equal to 0 if the scene is cloudy. This loss function may be used to understand the predicted probability in a given class. Higher predicted probabilities in an incorrect prediction are penalized more than lower predicted probabilities in an incorrect prediction. The sum of the predicted probabilities for the classes is 1. Whichever class has a probability ¿0.5 (for a binary model) is the predicted class. We define 'confidence' as the predicted class's probability; this is bounded between 0.5 and 1. A confidence of 0.5 represents a very uncertain prediction (and may essentially be thought of as a coin flip). A confidence near 1 represents a prediction that is certain (though that does not necessarily mean the prediction is correct). This may immediately offer a helpful feature for the cloud mask;

confidences associated with predictions are already built into the NN, the user does not have to design a separate procedure to define a proxy for confidence (often a necessity for some of the more traditional methods). One can further understand class probability by using the bulk statistics of a well calibrated model (see Gupta et al., 2020, and the references therein): for example, if the network assigns a confidence of 0.8 to a prediction, the prediction will be correct 80% of the time. The validity of confidence as a metric for accuracy is evaluated below.

Note that the β term in the loss function allows one to overcome class imbalances that may be present in the training data without having to remove data from the surplus class. For this study, the clear and cloudy classes are weighted such that each class has an equal effect on the overall loss. For example, if there is twice as many cloud scenes in the training set ($\beta = 0.66$) , a single cloud prediction has half as much influence on the loss function as compared to that of a single clear prediction. Classes, however, are not weighted for evaluation. That is, for evaluation, the terms β and $(1 - \beta)$ are removed from Eq. 2.1.

The architecture for the network is determined in an iterative way. A combination of various hidden layer depths and widths are tested and the combination that results in the lowest evaluated loss is used. NNs are trained for 200 epochs with a batch size of 10,000 and a validation set that is 25% the size of the training set. The iteration that results in the lowest validation loss during the 200 epochs is saved and used for analysis. Additionally, Dropout layers and BatchNormalization are used to deter overfitting and improve skill during the training process (Santurkar et al., 2018, Srivastava et al., 2014). The detailed structures of the networks used in this study can be found in the appendix material. The NNs used in this study were found to not be very sensitive to changes in model architecture (less than a percent change in accuracy among the various depth and width combinations tested).

The accuracy, loss (Eq. 2.1) and the Matthews Correlation Coefficient (MCC) of the NN are evaluated to guide the discussion of the cloud mask's skill. The MCC, defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.2}$$

is applicable even if classes are of very different size (Boughorbel et al., 2017). True negative (TN) represents the number of times in which the NN correctly predicts clear

scenes. False negative (FN) represents the number of times cloud scenes are incorrectly predicted by the NN as clear. True positive (TP) represents the number of times in which the NN correctly predicts the scene as cloudy, whereas false positive (FP) represents the number of times the NN predicts cloud when the scene is truly clear. This metric is an effective way to summarize the information contained within a confusion matrix and is common for machine learning scoring (Aronoff et al., 1982, Powers, 2020). A value of 1 represents perfect prediction performance, 0 is equal to that of random prediction, and -1 is perfect disagreement between predictions and the true values. Accuracy is defined simply as the number of correct predictions (TP + TN) divided by the total number of predictions.

Since numerical models frequently produce ubiquitous extremely thin (low total water path, TWP) clouds, rather than using GFDL's definition of cloudy (a non-zero value in the cloud water content), scenes that are 'clear' versus 'cloudy' are defined by comparing the sum of the scene's liquid and ice water path to a prescribed threshold that reflects the instrument's sensitivity. A scene which has a TWP above a selected threshold is classified as 'cloud' and a scene below that threshold is classified as 'clear'. The GFDL simulation's non-zero definition is not used since very thin clouds may not produce a spectral signature detectable by TIRS. It is evident, however, that the TWP may vary by quite a few orders of magnitude, even within the polar regions alone (refer to the inhomogeneity within Fig. 2.1(c,d)). While choosing an arbitrary threshold (e.g. 10 g/m^2) to demarcate a cloud may exhibit some amount of skill, this may not fully utilize the potential of the observations or the neural network. Sect. 2.3 introduces a statistical analysis for deriving an 'optimal' threshold for what should be considered 'clear' versus 'cloudy' in TIRS spectra.

2.3 Total Water Path Threshold Optimization

The selection of an 'optimal' TWP threshold is important for several reasons. A threshold too low degrades algorithm performance as the spectral signatures of thin clouds are weaker (due to low opacity) than the noise that may be present from the observed radiances. However, a threshold too great results in missing clouds that carry an obvious spectral signature and have notable radiative effects.

To derive an optimal threshold, distinct networks are trained and evaluated with various TWP thresholds demarcating clear versus cloudy. The data used for evaluation is the

validation set, which is 25% of the scenes derived from the 3 UTC timestamp (roughly $2\text{x}10^6$ different spectra); note that these scenes were not used to adjust weights during training. Only one timestep is used for this procedure due to (a) computational constraints (it is very expensive to run even a single GFDL timestep) and (b) randomly sampling across the northern and southern hemisphere provides a sufficient variety of scenes to appropriately test this procedure. The various NN are then compared to see which TWP threshold results in the most skill according to this large subsample. Since 3-by-3 GFDL patches are being aggregated to simulate a single TIRS FOV, the mean of the TWP for the corresponding 9 GFDL gridboxes that comprise the patch is used as the representative value for the simulated FOV.

Skill for each of the TWP thresholds is defined by the value of the evaluated loss of the NN, where the loss is a function of confidence in the predictions; low values of loss represent high skill. An example of a highly confident prediction may be that of the example prediction in Fig. 2.2(d), where the clear class is predicted with a probability of 0.95. Should this prediction be correct according to the NN's assigned threshold (i.e. this scene has a TWP below the selected TWP threshold), the loss value would be low, and the prediction would contribute to a skillful NN. However, should this prediction be incorrect (the scene is considered truly cloudy since it has a TWP above the selected threshold), the loss would be very large, since the NN has assigned high probability to clear.

Low confidences should be expected near a NN's TWP threshold since slight perturbations in TWP may induce spectral signatures that shift the scene from one class to the other; conversely, high confidences should be expected well away from this threshold. However, there are other subtle characteristics that may also influence confidence and skill. One could imagine that if a threshold is chosen that is above a TWP representative of an already opaque cloud, a scene labeled 'clear' may exhibit the same spectral signature of one labeled 'cloud'. This would then penalize a NN's loss, as two different scenes presented with very similar spectral signatures are labeled as different. That is to say, the loss function should then be minimized ('optimal') at the TWP representative of the thinnest detectable clouds with a sufficiently high signal-to-noise ratio for the NN.

Figure 2.3(a) depicts the loss as a function of the TWP threshold used for this baseline cloud mask procedure (blue). Each 'x' represents a distinct NN with the respective TWP threshold used to demarcate scenes that are clear versus cloudy during training

and evaluation, and the evaluated loss associated with that demarcation. There are small differences in the evaluated loss for different NNs with the same TWP threshold value due to stochasticity in the training process. However, this 'noise' is typically quite small, meaning any one NN is a good representation of the skill at that threshold value. A cubic function has been fit to the evaluated TWP thresholds and apparent extrema are reached in the loss and accuracy as the TWP approaches what is thought to be 'optimal' values for the TIRS' and algorithm's sensitivity. A minimum is achieved near a TWP value of 4 g/m^2, indicating that this may be the most appropriate threshold to designate 'clear' versus 'cloudy' in order to take full advantage of the NN's returned confidences. However, this analysis treats cloudy sky and clear sky detection as equally important. Since this algorithm will be sorting scenes to pass down to subsequent clear-sky algorithms, which can be quite sensitive to scenes that contain even thin clouds, a slightly lower TWP threshold value is chosen for the cloud mask (1.4 g/m^2) to optimize clear scene detection. This value corresponds to 105% of the fit function's minimum loss (the horizontal and vertical blue dashed lines in Fig. 2.3(a)). The accuracy which corresponds to this threshold value is shown in Fig. 2.3(b). In what follows, then, scenes which are referred to as 'clear' have a TWP value below 1.4 g/m^2 and scenes which are 'cloud' have a TWP value above 1.4 g/m^2. Ultimately this threshold leads to roughly a 2-to-1 cloud-to-clear ratio in terms of the number of scenes that comprise each class in training (and evaluation). We may compare this cloud-to-clear ratio to a study by Kay et al. (2016), which finds the Arctic to be characterized as cloudy approximately 60% of the time in winter and 80% of the time in summer, for a mean of 70%, agreeing relatively well with the optimally derived TWP frequency (about 66%).

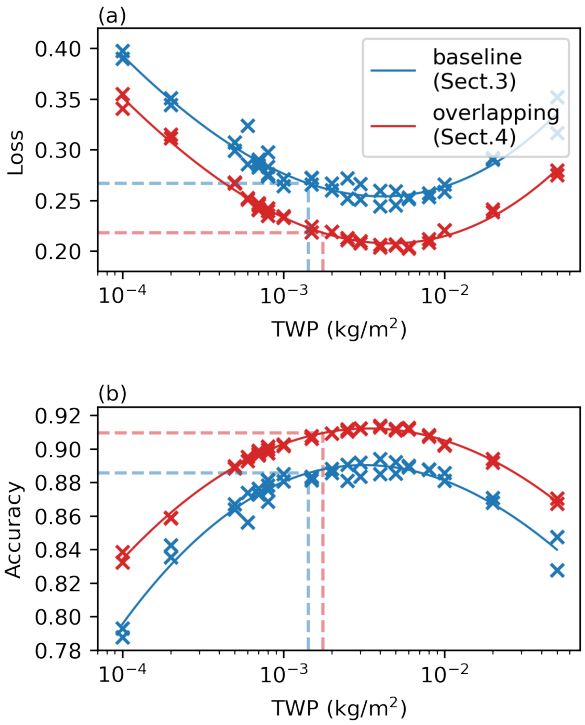

FIGURE 2.3: (a) The binary cross-entropy loss of neural networks with various TWP thresholds demarcating 'clear' versus 'cloud' scenes. The evaluation set for these metrics contains roughly 2x10^6 different spectra. Blue indicates the 'baseline' NN discussed in Sect. 3; red is representative of a more complicated NN which uses PREFIRE's overlapping procedure (discussed in Sect. 4). Solid lines are cubic functions fit to the evaluated points. Vertical dashed lines represent the TWP threshold corresponding to an evaluated loss that is 105% of the fit function's minimum loss (horizontal dashed). (b) The accuracy of the neural networks corresponding to the various TWP thresholds depicted in (a). Horizontal dashed lines correspond to the accuracy for the same TWP threshold depicted in (a).

Chapter 3

Algorithm Performance

In lieu of actual PREFIRE observations, the algorithm is evaluated using simulated spectra from a different timestep (9 UTC) in the GFDL simulation, where these particular spectra have no influence on the NN during the training process. In combination with the Dropout layers, results skewed by overfitting become less likely (Srivastava et al., 2014). Like during training, 3-by-3 GFDL gridboxes are aggregated, averaged, and noise is added to each spectrum for evaluation. This procedure once again results in about $8.4\text{x}10^6$ 'patches', representative of TIRS FOV, for evaluation. The large number of scenes not only allows one to make relatively reliable conclusions, but also contains enough variability to understand where the NN may be deficient.

Table 3.1 summarizes NN performance as a confusion matrix associated with the two classes. This table shows the true and false detection rates associated with clear and cloudy scenes. The upper left corner represents the accurate clear-sky detection rate, the lower right is the accurate cloud-sky detection rate. These rates are approximately equal, due to the fact that the data has been weighted to balance the number of clear and cloudy samples during the training process. Such an approach assumes that it is equally important to identify clouds as it is to identify clear scenes, where each is defined by the TWP threshold chosen. If desired, the NN could be tuned by adjusting the class weights, to further increase the cloud (clear) detection rate at the cost of some increase to the false cloud (clear) detection rate. This tuning could allow one to be more or less 'conservative', where a more conservative NN would be defined as one that predicts the alternate condition (cloudy) only when it is certain that it is cloudy, however, this

TABLE 3.1: Confusion matrix normalized by the true class for 3-by-3 aggregated GFDL gridboxes representing TIRS 'FOV', predicted by a neural network which uses only the spectrum provided by the PREFIRE instrument. Totals for each row and column are provided. loss = 0.264, accuracy = 0.89, MCC = 0.76

		Predicted Class		
		Clear	Cloud	Count
True Class	Clear	0.87	0.13	3,197,725
	Cloud	0.11	0.89	5,181,327
	Count	3,338,678	5,040,374	8,379,052

would come at the expense of increased FN (perhaps missing clouds that are thinner or less easily detected). One may want a less conservative cloud mask if the results are being fed into a surface classification algorithm; this ensures that there are no scenes that are contaminated with clouds. Conversely, an algorithm that is trying to identify cloud microphysical properties may want a more conservative cloud mask so that only cloudy scenes are being fed into the subsequent algorithm. The weighting procedure works in conjunction with the TWP threshold selection, where the weighting procedure is a means of statistically adjusting the detection rate of clouds and the TWP threshold is a means of physically adjusting the definition of those clouds (perhaps in some cases, an optically thinner definition for clouds must be defined, and the TWP threshold should be adjusted appropriately).

Figure 3.1 depicts the cumulative density function (CDF) of confidence for correct predictions (both clear (TN) and cloudy (TP)) and incorrect predictions (FN and FP). There is little difference if the CDFs are broken down into clear and cloudy classes separately (not shown). Encouragingly, the NN tends to have much greater confidence in its correct predictions than it does in its incorrect predictions. Roughly 60% of correct predictions are made with a confidence of 0.9 or greater. In contrast, only 10% of the incorrect predictions have this high degree of confidence. Furthermore, roughly 25% of incorrect predictions have relatively low confidence (below 0.6 for the maximum class probability), whereas nearly none (roughly 5%) of the correct predictions have this little confidence. The difference between the two CDFs is also plotted in black. The greater the area between the red and blue curves (or equivalently, below the black curve) the 'better' the behavior of the NN since this implies incorrect predictions have a greater proportion predicted with low confidence compared to correct predictions.

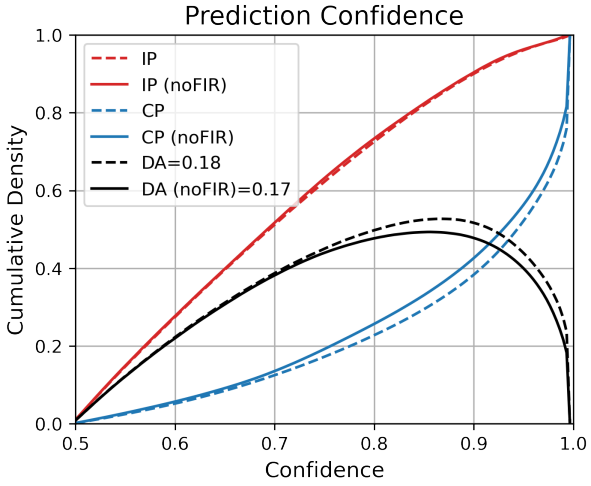

FIGURE 3.1: The normalized CDFs of the baseline neural network's prediction confidence for correct predictions (blue dashed) and incorrect predictions (red dashed). The difference between the CDFs of the incorrect and correct predictions is also provided (black dashed). The area of this curve can be found in the legend (DA). The same metrics but for a neural network excluding the FIR is also plotted (solid lines). A greater DA is thought to be indicative of a more desirable confidence behavior.

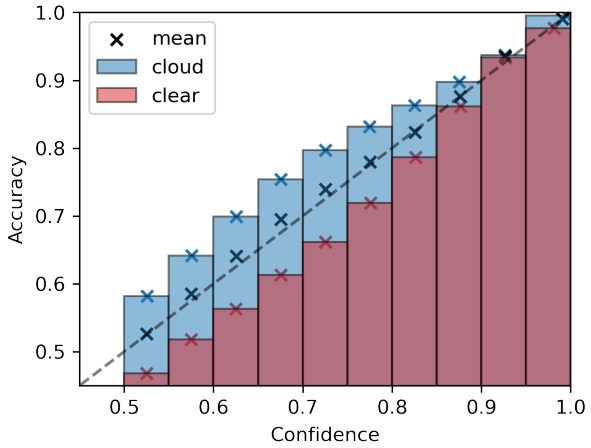

FIGURE 3.2: Binned values of confidence and their respective average accuracy for the baseline neural network. A perfectly calibrated model will lay on the 1:1 line. Bins are broken down by class (red and blue for clear and cloud, respectively). Both classes binned together is also plotted (black scatter).

It is also useful to understand the degree to which the NN is 'calibrated'. A well calibrated NN should assign a confidence of 0.5 to a prediction that will be correct 50% of the time. Figure 3.2 depicts the mean accuracy for various levels of confidence. Points should lie on the 1:1 line if the NN is perfectly calibrated. Confidences are broken down by each

true class (blue and red for cloud and clear, respectively) as well as the combination of both classes (black). The mean between the classes shows that this NN is relatively well calibrated as a whole. However, there is some mis-calibration with respect to the individual classes. True clear scenes tend to be predicted with a greater confidence than should be assigned based on accuracy (confidences between 0.5 and 0.55 are associated with roughly 45% accuracy, for example). However, cloud scenes, tend to be associated with a lower confidence than should be assigned (confidences between 0.5 and 0.55 are associated with roughly 60% accuracy, for example). This discrepancy likely has to do with the class imbalance in the data, where clear scenes have been weighted greater in the training process which has skewed their confidence calibration slightly. The two classes calibrations' converge towards 'perfect' calibration as confidence increases, which as exhibited in Fig. 3.1, occurs for the majority of correct predictions. Together, Fig. 3.1 and Fig. 3.2 suggest that the confidences assigned by the NN provide a useful means for attaching a quality flag to the predictions.

To characterize NN skill for different conditions and cloud types, Fig. 3.3(a) depicts a 2d histogram of the confidence as a function of the estimated cloud top height (eCTH) and the TWP of the evaluated scene. The eCTH is determined by assuming a 1000-meter thick cloud and is thus the lowest pressure level (greatest altitude) in which the total water content (TWC) of the GFDL simulation is greater than one one-thousandth of the TWP threshold. Note that, in the case of geometrically thin model clouds, this constraint may be satisfied in some cases even if the scene is deemed 'clear' based on the selected TWP threshold. These scenes (as well as all scenes below the TWP threshold) have been assigned as 'no eCTH' and are labeled as such in the figure (top row) for consistency with the training assumptions. In the case of geometrically (and optically) thick clouds, the TWC may not peak above the required threshold, so in some cases no eCTH is found (top right of plot). Furthermore, there are some instances in which aggregating 3x3 patches leads to artifacts in this table. Explicitly, to find the eCTH for our 3x3 patches, we first find the eCTH for each of the GFDL gridboxes which comprise it individually, and then take the median of the 3x3 patches to find the representative eCTH for the coarsened 'FOV'. This causes problems along some of the cloud edges, specifically where the patch lies over sharp transitions from clear to very opaque clouds. This leads to nonphysical artifacts, like the top right bin of Fig. 3.3, where very opaque clouds never satisfy the TWC requirement and thus no eCTH is found. However, this occurs in relatively few cases, as indicated by the hatching, and thus these bins with very small sample sizes

should be used with caution. The use of eCTH offers a simple means of binning the data and comparing NN performance at different levels and is not necessarily a rigorous evaluation of the 'true' cloud top height or effective emission height.

A more skillful NN should exhibit more confidence in the cloudy class (darker blue) to the right of the TWP threshold (1.4 g/m^2), demarcated by the vertical line labeled t, as these have been defined as 'true' cloud scenes. Conversely, more confidence in the clear class (darker red) should lie to the left of the threshold line. By binning the data as we have, we are able to mitigate whether the FV3 model can produce the true TWP and vertical distribution of clouds that is exhibited in nature. Rather, we are able to simply see what scene types the NN is more or less skilled at predicting.

As partially exhibited in the derivation of an optimal TWP threshold, the confidence generally increases as the TWP of the scene diverges from the threshold chosen. Additionally, confidence increases as the eCTH increases in altitude. This would be expected as high clouds generally have stronger spectral signals since their emission temperature is typically more different from the surface than that of low clouds. Furthermore, water vapor absorption bands may still carry some information for high altitude clouds, while low altitude clouds may be completely masked by strong absorption in these bands. Note that no clouds exist below pressure levels of 200 mb according to this definition. Figure 3.3(b) compliments Fig. 3.3(a) by showing the cloud detection rate for each pressure bin (blue), as well as the mean clear detection rate (red); these depict binary accuracy rather than confidence. The mean cloud detection rate is also plotted for reference (dashed vertical line). As one would hope, greater confidences generally mean greater accuracy in cloud detection at that level. Clouds near the surface are, on average, the most difficult for the NN to detect, especially near the TWP threshold. The clouds are commonly labeled with low confidence for either class (indicated by white).

Figure 3.3(c) gives the cloud detection rate (blue) for each TWP bin above the threshold line, as well as the clear detection rate (red) for each TWP bin below the threshold line. As reflected in the confidences, accuracy generally increases as the TWP diverges away from the threshold chosen. Very opaque clouds (TWP ¿ 30 g/m^2) are detected nearly every time they occur (an accuracy of 95%), despite the difficulties of low altitude cloud detection.

TABLE 3.2: Confusion matrix normalized by the true class as predicted by a neural network which does not include PREFIRE's FIR portion of the spectrum (¿ 15μm). Totals for each row and column are provided. loss = 0.293, accuracy = 0.87, MCC = 0.73

		Predicted Class		
		Clear	Cloud	Count
True Class	Clear	0.84	0.16	3,197,725
	Cloud	0.11	0.89	5,181,327
	Count	3,245,906	5,133,146	8,379,052

3.1 Value of the FIR

These results suggest there is value to adding the FIR channels measured by PREFIRE for improving polar cloud detection. To further test this, we train a NN without any channels above 15 μm and compare it to the baseline NN which contains the full spectrum. The same training procedure, architecture (aside from the input layer), TWP threshold, and evaluation set is used for both NNs.

Table 3.2 depicts the confusion matrix for a cloud mask omitting the TIRS FIR channels (referred to as noFIR). The error indices for this particular NN are also reported in the figure caption. As compared to the full spectrum NN, the loss increases by roughly 10%, the MCC decreases by 0.03 and the accuracy decreases by 2%; all of which correspond to a less skillful NN. Note that, while this change in accuracy may seem relatively small, the absolute accuracy of the NNs are already very high (as well as the sample size to which they are evaluated on) such that a decreased accuracy of 2% represents a relatively large increase in incorrect classifications of nearly 20%. Interestingly, most of the divergence in accuracy between the two NNs occurs in the TN rate (correct detection of clear scenes); this decreases from 0.87 in the baseline to 0.84 in the noFIR NN. The TP rate (correct detection of cloud scenes), however, is approximately equal between the two NNs.

The effect that discarding the FIR has on the CDFs of confidence is depicted in Fig. 3.1 (solid). There is little difference in the distribution of confidence for incorrect predictions whether the spectrum contains the FIR or not (similar red solid and red dashed). However, not including the FIR results in a greater proportion of less confident correct predictions (blue solid and blue dashed). This is further reflected in the difference curve

(black), which decreases in area if the FIR is not included in the prediction (as mentioned previously, a greater area indicates a NN with more desirable confidence traits).

Figure 3.4(a) compares the 2d histogram for the noFIR NN to that of the full spectrum NN (Fig. 3.3). If the full spectrum is more skillful, TWP values greater than the TWP threshold (right of t) should appear blue while TWP values less than the threshold (left of t) should be red. Including the FIR acts to improve the confidence in correctly detecting clear scenes at all TWP bins below the threshold; thus including the FIR may be acting to 'confirm' already correct detections of clear scenes. As exhibited in the confusion matrix, including the FIR also improves the clear scene detection rate (red in Fig. 3.4(b,c)).

The effect of including the FIR, however, has a slightly more complicated relationship for cloudy scenes. Perhaps surprisingly, the full spectrum NN actually has less confidence in several of the true cloud cases as compared to the noFIR NN, especially for lower level, more opaque clouds (lower right of Fig. 3.4(a)). This is also reflected in the accuracy metrics (bottom of Fig. 3.4(b) and right of Fig. 3.4(c)). Most of the improvement in confidence for true cloud scenes occurs in the mid-to-high altitude bins (300-700 mb), where using the full spectrum leads to more confident predictions with an accompanying increase in cloud detection rate of 1-4%.

The reason for the improved clear scene confidences and detection rates when including the FIR is likely a result of the unique FIR signatures of ice clouds, water vapor, and different surfaces. However, it is important to state that some of this behavior may in part be a result of the threshold chosen for cloud versus clear, and a different threshold may exhibit different behavior for each class.

Furthermore, it is evident that while the FIR appears to aid in mid-to-high altitude cloud detection, the noFIR NN has a comparable level of skill in overall cloud detection. This suggests that some of the information offered in the FIR portion of the spectrum may be redundant to that of the channels shorter than 15 μm. Furthermore, the TIRS FIR channels have greater instrument noise than some of the shorter wavelength channels. Removing redundant information in these channels that is associated with a lower signal-to-noise may actually improve accuracy in some cases. This would explain why very opaque clouds, which are likely associated with a very obvious mid-infrared spectral signature to begin with, show decreased detection rates; the noise in the FIR essentially only acts as a means of confusing the NN. Similarly with low clouds, which are likely

associated with the least amount of signal to begin with and may have no information in the FIR anyway (due to water vapor absorption), the increased noise in the FIR could prove to be detrimental.

However, like with the clear detection rate, there may be artifacts in the change in skill based on particular model assumptions, specifically how cloud phase is simulated in the GFDL dataset. Clouds that occur at pressure levels less than 440 mb are deemed as ice clouds versus those greater than 440 are liquid clouds. The transition apparent between the 400-500mb bin and the 300-400mb bin, may be explained by this model parameterization. Thus, one should be cautious making conclusions based solely on the estimated cloud top height.

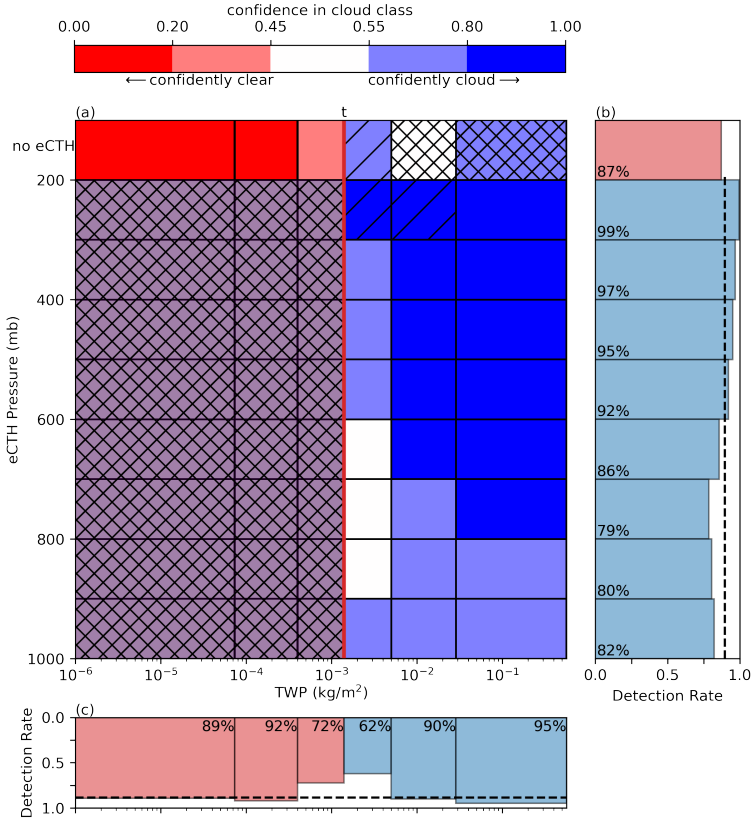

FIGURE 3.3: (a) A 2d histogram of the assigned probability to the 'cloud' class by the baseline neural network for scenes binned by estimated cloud top height (eCTH) and the total water path (TWP) of the scene. Bins which contain less than 0.5% of the dataset have a single slash, bins with less than 0.1% are crossed, empty bins are grey. True clear scenes lie to the left of the designated TWP threshold, t, whereas true cloud scenes lie to the right. All true clear scenes are binned together as having no eCTH (top) and thus are broken down only by a function of TWP. Darker shades of blue indicate that the neural network is on average very confident that scene bin is cloud. Darker shades of red indicate that the neural network is on average very confident that scene bin is clear. White indicates little confidence one way or the other. (b) The detection rate (number of times correctly predicted 'cloud' divided by the number of clouds within that bin) for cloud scenes as a function of eCTH. The mean cloud detection rate, over all eCTH levels, is plotted with a vertical dashed line. The clear detection accuracy is also listed at the top. (c) As with (b) but as a function of the TWP of the scene. The mean accuracy of the NN is plotted with a horizontal dashed line. In (b) and (c) blue bars indicate statistics relating to cloudy scenes and red bars indicate statistics relating to clear scenes.

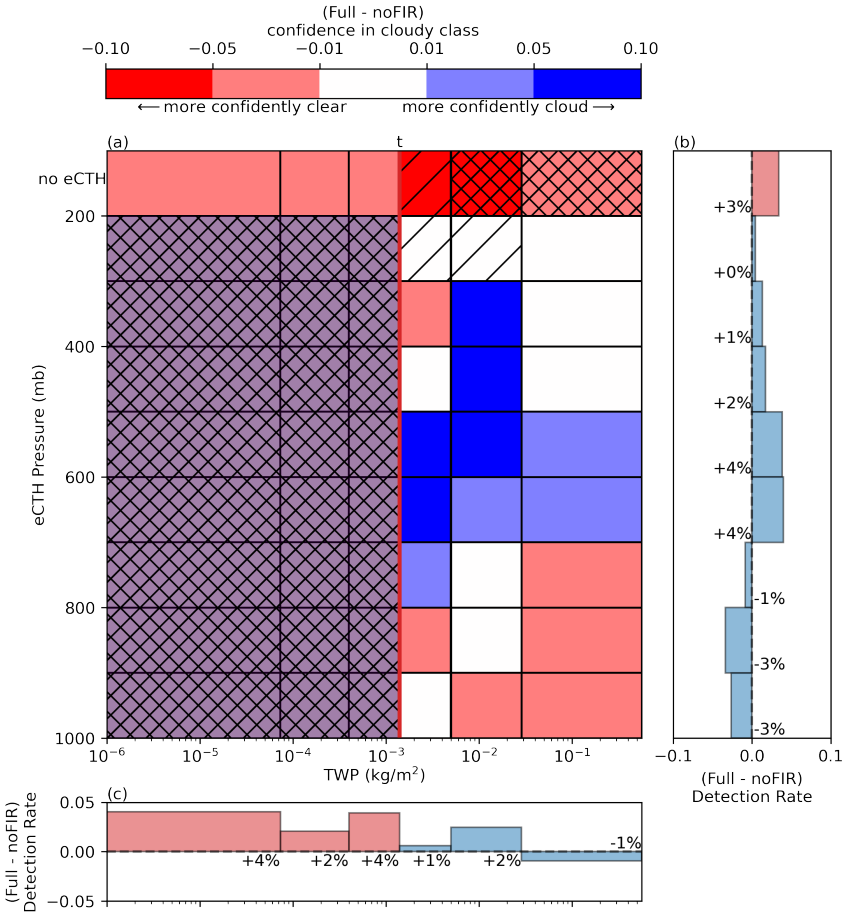

FIGURE 3.4: (a-c) The difference between the 2d histogram and accuracy values depicted in Fig. 3.3 and that of a neural network does not include PREFIRE's FIR portion of the spectrum ($> 15\mu$m). (a) Darker shades of blue indicate that the full spectrum NN assigned a greater probability to the cloud class, darker shades of red indicate that the full spectrum NN assigned a greater probability to the clear class. (b,c) the change in the detection rates for binned scenes as a function of eCTH and TWP, respectively, when comparing the full to the noFIR NN; positive values indicate increased detection by the full spectrum NN.

Chapter 4

Improving the Baseline NN

There are several ways the NN described above may be improved upon. Two approaches are explored here: the use of overlapping FOV and the use of environmental conditions. The basis for each is first discussed independently and then the performance of a NN which incorporates both modifications is analyzed using similar metrics to those discussed thus far.

4.1 Super-resolution Detection

The fact that neighboring TIRS samples overlap by two-thirds in the along-track direction may both enable cloud detection at a super-resolution (smaller than a single FOV) and reduce the mis-classification of scenes as result of instrument noise. The TIRS sampling is illustrated in Fig. B.2. With the appropriately selected overlapping FOV, each one-third of a scene will be overlapped by three separate FOV. A NN can be trained which predicts only on that area which is overlapped by three separate FOV. Explicitly, this means the NN predicts at one-third the FOV's scale in the along-track direction. Improved scene detection may occur as a result of the contrast from neighboring FOV; this contrast may be especially strong when the three FOV reside over clear-sky to cloud-sky transitions. This use of contrast is not available to the baseline NN because FOV are passed through independently.

To accommodate this overlap information, the structure of the neural network is altered to read in three neighboring along-track FOV at once (for example, the three colored FOV

in Fig. B.2(b)) rather than a single FOV. Simulating overlapping FOV in the GFDL simulation is relatively simple since the sliding window of 'patches' that simulate FOV inhomogeneity already overlap. However, predictions are now only for that portion which is overlapped by all three FOV (golden outline in Fig. B.2(b)) and thus labels are aggregates of 3-by-1 GFDL gridboxes. As before, if the average TWP of this 3-by-1 patch is larger than the TWP threshold, that scene is labeled 'cloud', otherwise it is labeled 'clear'.

It is reasonable to expect that the sensitivity of TIRS to cloud cover may change when overlap is explicitly modeled, so the overlapping procedure may have a different 'optimal' TWP threshold for demarcating 'clear' versus 'cloudy'. Thus the procedure performed to derive an optimal threshold for the baseline NN is repeated with the new overlapping NN. The skill and accuracy for various TWPs using the overlapping NN is depicted in red in Fig. 2.3. The shape of the fit function is very similar to that of the baseline NN (blue), however, the overlapping NN is more skillful (and more accurate) at all TWP thresholds. There is a slight shift in the 'peak' skillfulness to a larger TWP threshold for the overlapping procedure. The corresponding 105% skill TWP value is also slightly greater (roughly 1.75 g/m^2 as compared to 1.4 g/m^2 for the baseline procedure). The reason for this shift is not fully understood, however, it may be related to differences in the 'optimal' amount of spectral contrast between the clear and cloud scenes that the overlapping procedure uses. However, given the similarity between the thresholds for the baseline and overlapping NNs, we choose to retain 1.4 g/m^2 as the TWP threshold demarcating clear versus cloud for the overlapping procedure. This avoids confounding effects of threshold changes for comparisons between the two NNs.

The effects that instrument noise have on the overlapping NN as compared to the baseline NN are evaluated by training separate versions of each NN with and without noise and finding the difference in skill. If the overlapping NN is less sensitive to noise, we would expect its skill to deteriorate less than the baseline NN. With the little variation in evaluation loss for different NNs trained with the same TWP threshold (Fig. 2.3), especially near the threshold that has been chosen, any one NN that is trained is a good representation of the skill at that TWP threshold value.

It is found that the baseline procedure's loss degrades from 0.203 to 0.263 (accuracy from 91.5% to 88.5%) with the addition of spectral noise. The overlapping procedure's loss degrades from 0.169 to 0.220 (accuracy from 93.4% to 90.9%) with the addition of spectral

noise. The overlapping procedure is, therefore, degraded by about 15% less than that of the baseline procedure, however the relative magnitude of degradation is about equal. This indicates that the overlapping procedure may not have a large effect on reducing the error caused by spectral noise alone. That being said, there is improved skill using the overlapping FOV compared to the baseline NN regardless of whether noise is added or not.

4.2 Overlapping FOV and Environmental Conditioning

Since cloud occurrence depends on the local environment, another potential improvement to the baseline NN's skill may be offered by conditioning the NN using meteorological variables as inputs. The PREFIRE auxiliary meteorology product (Aux-Met) will provide various meteorological conditions mapped to the FOV captured by the PREFIRE satellites. To exploit this information, the new overlapping NN is expanded upon with the use of two of these model variables: surface temperature and total column water vapor (TCWV). In the context of the simulated data, we directly use the FV3 model output as the 'NWP'.

The Aux-Met outputs that are derived from numerical weather prediction (NWP) model output will have some uncertainty associated with them, thus, the effect of uncertainty in these two model variables must also be included. Uncertainty is modeled as a Gaussian distribution whose mean is 0 and standard deviation is equal to that of $n\sigma_i$. Where σ_i represents the 'expected uncertainty' of model variable i. For surface temperature $\sigma_{ts} = 1\ ^o\text{C}$ is used and for TCWV a relative value of $\sigma_{tcwv} = \frac{1}{10}TCWV$ is used. All models are trained with $n = 1$, however, NNs with other factors of n are tested for evaluating the impact of NWP errors on NN performance.

Figure 4.1 depicts the loss and accuracy for models that include overlapping FOVs and the two model variables with varying uncertainty factors (black). The loss and accuracy for the baseline NN (red dashed) and overlapping NN alone (blue dashed) are depicted for reference. Note that including meteorological conditions improves skill compared to using the overlapping procedure's radiances alone until the model variable uncertainty exceeds twice the expected values ($2\sigma_i$). It remains more skillful than the baseline NN until triple the expected values ($3\sigma_i$). The fact that model variables only contribute skill

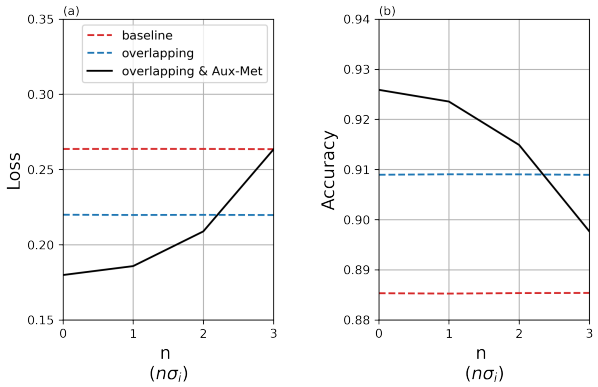

FIGURE 4.1: (a) Four different evaluated losses for a single model which uses the overlapping procedure and model variables (surface temperature and total column water vapor). The evaluations differ by the amount of uncertainty that is added to the model variables; uncertainty is drawn from a Gaussian distribution whose mean is 0 and standard deviation is equal to $n\sigma_i$. The x-axis represents various values of n. All models are trained with $n = 1$. The evaluated loss for a model which does not use the model variables (blue) and a model which does not use the overlapping procedure or the model variables (red) is also provided. (b) As with (a) but for the evaluated accuracy.

until uncertainty reaches values of $2\sigma_i$ may speak to the NN's ability to extract similar information from just the radiances themselves, possibly through some combination of channels or nonlinear characteristics. For further evaluation, the expected values of model uncertainty (i.e. $n = 1$) are used.

Using the same procedure as before, the benefit that the FIR has to a NN which also contains the model variables is briefly evaluated. It is found that including the FIR increases accuracy by approximately 1% and decreases the loss by 10%. Again, while the accuracy difference may seem small, the number of incorrect predictions is relatively few to begin with (and in this case, the absolute accuracy of the NNs is greater than before).

Figure 4.2 depicts the CDFs of confidence for incorrect and correct predictions for a NN which uses overlapping FOV and the two model variables (henceforth referred to as the 'O-AM' NN for Overlapping-Aux-Met). The baseline NN's CDFs are plotted for comparison (dashed). The O-AM NN produces an even greater proportion of correct predictions with high confidence than the baseline. Roughly 80% of correct predictions have a confidence of 0.9 or greater for the O-AM NN versus about 60% for the baseline NN. The incorrect prediction distribution flattens out slightly across all confidences for

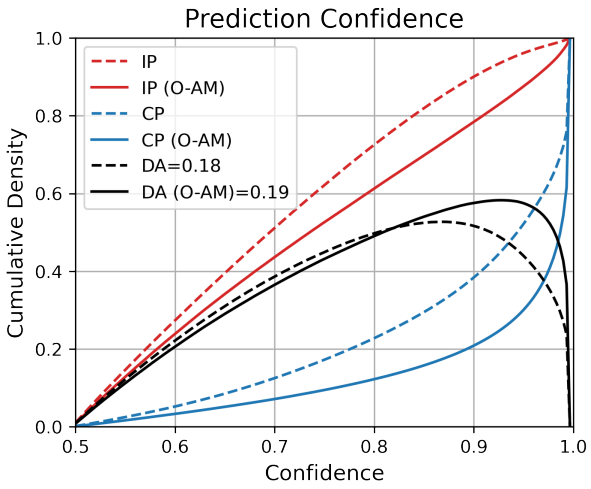

FIGURE 4.2: As with Fig. 3.1 but with a model which uses the overlapping procedure and two Aux-Met variables with their expected uncertainty values (solid) compared to that of the baseline full spectrum model (dashed).

TABLE 4.1: Confusion matrix normalized by the true class as predicted by a neural network which uses the TIRS overlapping FOV and model variables with expected uncertainty values. Totals for each row and column are provided. loss = 0.186, accuracy = 0.92, MCC = 0.84

		Predicted Class		
		Clear	Cloud	Count
True Class	Clear	0.93	0.07	3,204,120
	Cloud	0.08	0.92	5,174,930
	Count	3,374,742	5,004,308	8,379,050

the O-AM NN compared to the baseline. This could be as a result of some of the 'on-the-edge' predictions (probabilities near 0.5) that the baseline NN would make, shifting over to the correct prediction set with the O-AM NN (leading to an increased accuracy). In any case, even with the flattening of the incorrect prediction CDF, the difference area for the O-AM NN still increases relative to that of the baseline.

Table 4.1 presents the confusion matrix associated with the O-AM NN. Once again, the accuracy for each class is balanced. The correct detection rate for both classes shows noticeable improvement as compared to the baseline NN, with an increase of 0.87 to 0.93 for clear scenes and 0.89 to 0.92 for cloud scenes. The MCC increases from 0.76 to 0.84 when using the O-AM NN. Finally, the loss decreases from 0.264 to 0.186 when using

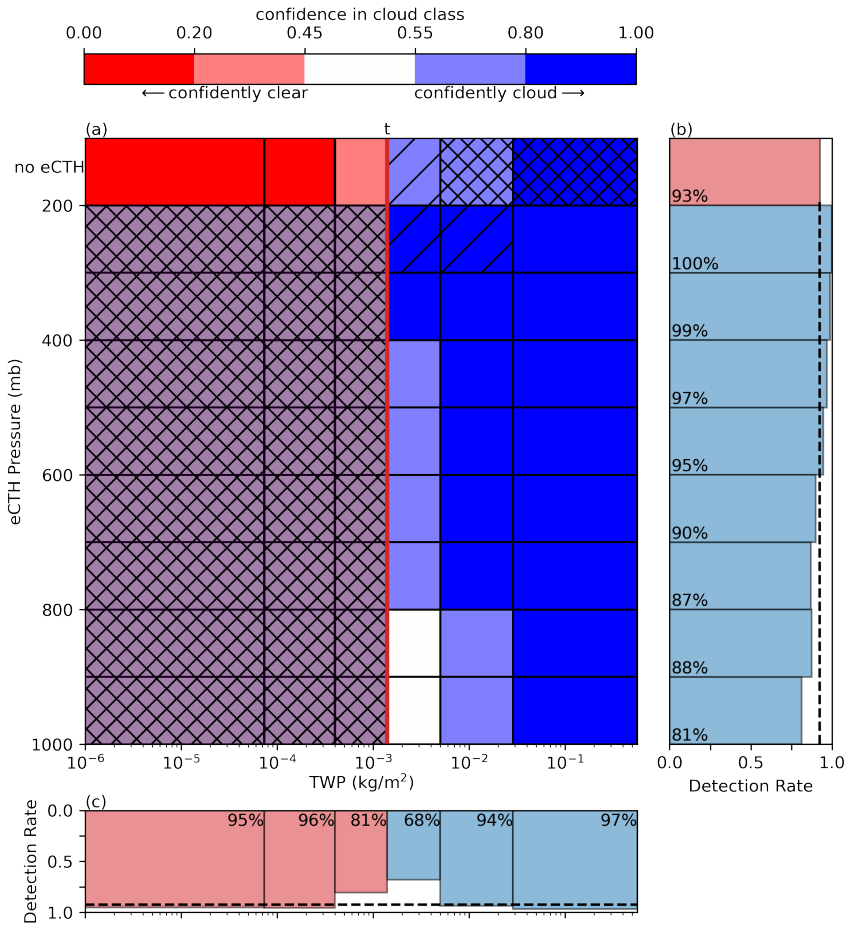

FIGURE 4.3: As with Fig. 3.3 but for a neural network which uses the TIRS overlapping FOV and model variables (O-AM) with expected uncertainty values.

the O-AM NN as compared to the baseline. All three of these metrics suggest that O-AM is a far more skillful NN than the baseline. Additionally, recall that the O-AM NN is predicting on a sub-FOV scale, providing higher resolution output than the baseline. Clearly these new additions not only increase the number of correct predictions, but also increase the confidence in those predictions as compared to the baseline NN.

Figure 4.3(a) depicts the 2d histogram of confidence as a function of eCTH and TWP for the O-AM NN. Note that bin sizes may differ slightly from that of Fig. 3.3 and Fig. 3.4 because class labels represent 3-by-1 aggregates instead of 3-by-3 aggregates. As with the baseline NN, confidence generally increases as the TWP of the scene is further away from the threshold t. These confidences, however, are typically of greater magnitude for the correct class using the O-AM NN compared to the baseline. This is especially the case for

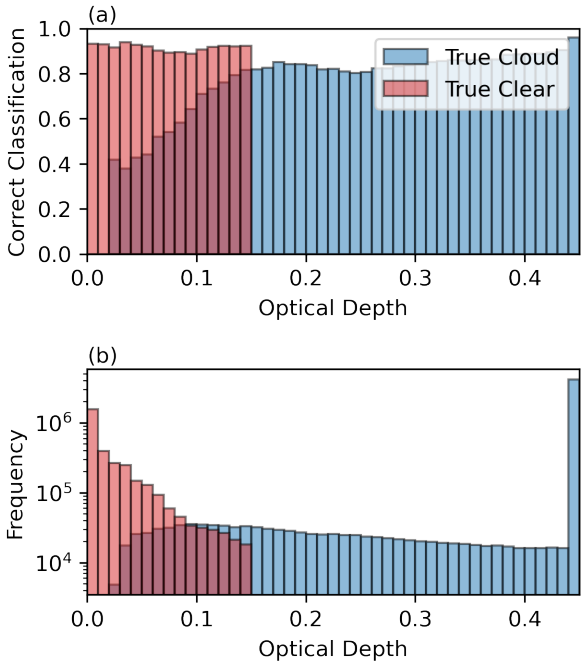

FIGURE 4.4: (a) Binned values of scenes' visible optical depth compared to the fraction in which the O-AM NN classifies the scenes as clear. Scenes which are truly clear are depicted in red. Scenes which are truly cloud are depicted in blue. Red bins are cutoff near the maximum optical depth that the true clear scenes have, and blue bins are cutoff near the minimum optical depth that the true cloud scenes have. (b) The number of true clear (red) and true cloud (blue) scenes that are contained within each of the optical depth bins. Note that the final bin of the 'true cloud' histogram is clipped in the sense that it contains all counts for optical depths greater than 0.45 as well.

mid-to-high altitude clouds, perhaps aided by the O-AM NN's ability to use contrasting radiances. Nearly every cloud pressure level bin shows improved accuracy using the O-AM NN. Similarly, every TWP bin (Fig. 4.3(c)) shows improvement as compared to the baseline NN; even the clear bin very near the threshold chosen achieves an accuracy of 81%. Cloud scenes whose TWP exceeds 5 g/m^2 are detected nearly perfectly (greater than 94% detection rate). Clear scenes which have a TWP less than 0.4 g/m^2 are detected approximately 95% of the time.

Figure 4.4(a) depicts the fraction of scenes classified clear by the O-AM NN as a function of the visible optical depth of the scene. The computation of the optical depth assumes a monodisperse cloud with spherical particles which have an effective diameter of 20 μm for liquid and a temperature dependent value for ice. Explicitly this is defined as $\tau_{liq} = \frac{3}{2}\frac{LWC}{(\rho_w)(r_e)}$ for liquid clouds, where $\rho_w = 997$ kg/m^3 and $r_e = 10\text{x}10^{-6}$ m and $\tau_{ice} =$

$\frac{3}{2}\frac{IWC}{(\rho_i)(r_e)}$ for ice clouds, where $\rho_i = 917$ kg/m^3 and $r_e = \frac{1}{2}(326.3 + 12.42T_c + 0.197T_c^2 + 0.0012T_c^3)x10^{-6}$ (Ou and Liou, 1995). The optical depth for each layer is determined by the greater of τ_{ice} and τ_{liq}. The optical depth of the scene is the sum of these layers. Figure 4.4(b) provides the frequency of each optical depth bin for reference (log scale). Red is used to represent the optical depth for scenes deemed 'clear' and blue is used to represent the optical depth for scenes deemed 'cloud'. The final bin in the histogram is clipped in the sense that it also contains those scenes which have an optical depth greater than 0.45.

The correct classification of cloud scenes steadily increases as the optical depth of the scene increases. Scenes with optical depths beyond approximately 0.4 are nearly always detected correctly. Scenes with TWP above the designated threshold (defined as clouds) but which have low optical depths– for example, ice clouds with relatively large effective radii, are the most difficult for the NN to detect. Interestingly, the correct identification of clear scenes is relatively constant as a function of the optical depth of the scene; however, the majority of clear scenes have optical depths less than 0.01 and so effects of varying sample sizes (for training and evaluation) may come into play. Encouragingly, the NN is relatively predictable in terms of what proportion of scenes will be classified cloudy, relative to the optical depth of the scene.

Chapter 5

Conclusions

The polar regions have many unanswered questions in regard to current and future climate. The Polar Radiant in the Far InfraRed Experiment (PREFIRE) hopes to answer some of these questions, especially those surrounding radiative balance. A major step in this process is the development of a cloud mask for the region that effectively utilizes PREFIRE's unique FIR observations. This study explores the potential skill, as well as limitations, of a neural network-based cloud mask using simulated TIRS spectra. While interpretations of skill should be performed cautiously since the simulated spectra have several limitations (such as the evaluation set being from a single timestep, assumptions made about the inputs to PCRTM, the simulated FOV being smaller than the true TIRS FOV, and ignoring instrument calibration errors), the NN exhibits similar skill for both poles (which are in opposite seasons) and with differing simulated FOV size. The lack of sensitivity to the domain and the size of the simulated FOV suggests that the NN exhibits skill due to its ability to detect patterns within the spectra rather than biases due to some type of overfitting. Furthermore, the variety of scenes offered by the FV3 simulation allows us to understand what particular cloud configurations will be most challenging for our cloud mask to detect.

The single FOV baseline NN which uses the captured TIRS spectra alone, is able to correctly detect approximately 89% of clouds within the polar regions. This baseline, however, can be improved upon using the TIRS along-track overlapping FOV and auxiliary meteorological variables (surface temperature and total column water vapor). This 'O-AM' NN predicts on a sub-FOV scale with an even greater accuracy (93%) than the

baseline NN. Model variables themselves, however, have uncertainty associated with them and are found to augment NN performance until uncertainty values exceed about 2^{o}C and 20% for surface temperature and total column water vapor, respectively. The use of the overlapping procedure alone results in a NN with an accuracy of roughly 91%, which still surpasses the anticipated cloud mask skill defined in L'Ecuyer et al. (2021).

The inclusion of the far-infrared (FIR) is found to significantly improve clear scene detection and better resolve mid-to-high altitude clouds. However, there is slight degradation in skill for low-altitude clouds, likely due to the lower signal-to-noise and the increased absorption by water vapor associated with these channels. That being said, including the FIR increased overall NN skill. A scene-dependent decision tree approach for whether to include the FIR as an input to the cloud mask may be considered in the future.

The designed neural networks have the added benefit of returning a class probability associated with each prediction, a property that most 'traditional' cloud detection methods do not have already built in. These assigned probabilities (or confidences) are found to be well calibrated and trustworthy. Additionally, correctly predicted scenes are found to generally have a much greater prediction confidence than incorrectly predicted scenes. These properties make the use of returned confidences helpful in an operational sense, perhaps as a means of assigning a quality flag with each evaluated scene.

Confidence and accuracy generally increase as the estimated cloud top height increases and as the TWP of the scene diverged away from the selected threshold demarcating 'clear' and 'cloud' (1.4 g/m^2). Clear scenes with low TWP values (less than 0.4 g/m^2) or cloud scenes with high TWP values (greater than 5 g/m^2) are nearly always correctly identified (over a 95% detection rate). Furthermore, the NN has very stable behavior with respect to the frequency it classifies a scene as 'clear' and the optical depth of the scene (Fig. 4.4).

Although a 'true' evaluation of skill, especially in comparison to other operational cloud masks, should be saved until real (as opposed to simulated) spectra from the PREFIRE mission are captured and an independent means of validation is available (mainly through co-location with ground-based instruments or whatever other instruments are operational), the skill exhibited by the presented neural networks may outperform other cloud masks in the polar regions for several reasons: namely the specificity (polar-targeted), the sensitivity (based on an 'optimal' TWP threshold) and the unique observations (such

as the FIR) that the networks have been designed around. The flexibility surrounding neural network creation (the loss function used and what information that may return) as well as targeted scene identification (weighting the importance of each class' contribution to loss) allows networks to be easily adapted to different problems at hand and may be especially useful for a mission's algorithm chain, where a variety of techniques to extract various geophysical information are used. The development of a cloud mask that exhibits the same skill in the polar regions but can be applied to the global domain with minimal degradation will be explored for the operational form of the mask.

Appendix A

NN Structures

This section contains an overview of all the neural network structures used in this study. The name of the network as it is referred to in the text is listed at the top of each table. 'Layer levels' are a rough guide for the order in which inputs progress through the network. More information for the particular layer types can be found in Keras' documentation (Chollet et al., 2015).

TABLE A.1: The neural network structure for the baseline NN (Sect. 3). Layer levels may be used as a reference for how inputs progress through the NN. Layer types and output shape are listed for each layer.

Baseline NN Structure		
Layer Level	Type	Output Shape
1	Input	(None, 52)
	BatchNormalization	(None, 52)
2	Dense	(None, 256)
	Dropout(0.2)	(None, 256)
3	Dense	(None, 256)
	Dropout(0.3)	(None, 256)
4	Dense (softmax activation)	(None, 2)

TABLE A.2: As with Tab. A.1 but for the noFIR NN (Sect. 3a).

noFIR NN Structure		
Layer Level	Type	Output Shape
1	Input	(None, 9)
	BatchNormalization	(None, 9)
2	Dense	(None, 256)
	Dropout(0.2)	(None, 256)
3	Dense	(None, 256)
	Dropout(0.3)	(None, 256)
4	Dense (softmax activation)	(None, 2)

TABLE A.3: As with Tab. A.1 but for the O-AM NN (Sect. 4b).

O-AM NN Structure		
Layer Level	Type	Output Shape
1	Input	(None, 54)
1	Input	(None, 54)
1	Input	(None, 54)
1	Concatenate	(None, 156)
	BatchNormalization	(None, 156)
2	Dense	(None, 256)
	Dropout(0.2)	(None, 256)
3	Dense	(None, 256)
	Dropout(0.3)	(None, 256)
4	Dense (softmax activation)	(None, 2)

Appendix B

Additional Figures

This section contains additional figures for the spectral response functions, estimated noise values of the TIRS instrument, and a diagram for the TIRS sampling procedure.

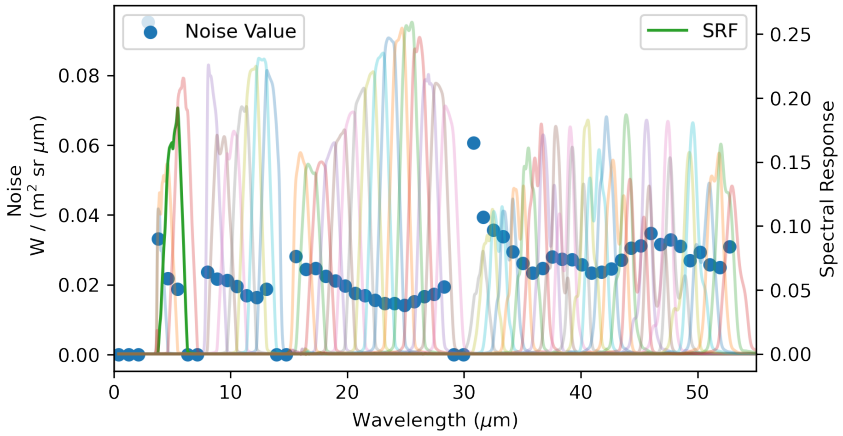

FIGURE B.1: Expected values of each channel's noise plotted against the channel's center wavelength (blue scatter) for the TIRS instrument. Masked channels due to instrument limitations are listed with noise values of 0. The spectral response function (SRF) for each channel is shown with a colored line.

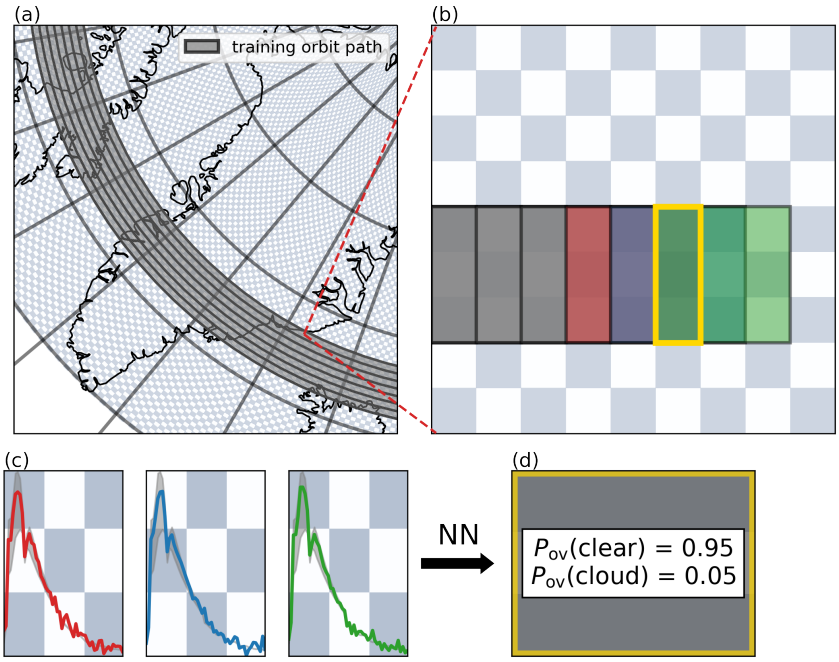

FIGURE B.2: The overlapping procedure which is described in Sect. 4. (a) As with Fig. 2.2, training orbit paths run along lines of latitude. Blue-white checkboard pattern represents the GFDL gridbox size. (b) A zoomed in along-track path. Boxes encompassing 3-by-3 patches of GFDL gridboxes represent simulated TIRS FOV. The red, blue, and green boxes represent the three most recent FOV captured in the orbit path, showing how they overlap with their neighbors. The golden surrounding box represent that area which is overlapped by all three FOV. (c) The three mean spectra, representative of their respectively colored FOV, that will be fed into the neural network to predict on the golden area. (d) The neural network's clear versus cloud prediction for the region which is one-third the size of a single captured FOV.

Bibliography

Ackerman, S., R. Holz, R. Frey, E. Eloranta, B. Maddux, and M. McGill, 2008: Cloud detection with modis. part ii: validation. *Journal of Atmospheric and Oceanic Technology*, **25 (7)**, 1073–1086.

Aronoff, S., and Coauthors, 1982: Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing*, **48 (8)**, 1299–1307.

Aumann, H. H., and Coauthors, 2018: Evaluation of radiative transfer models with clouds. *Journal of Geophysical Research: Atmospheres*, **123 (11)**, 6142–6157.

Bantges, R., H. Brindley, X. Chen, X. Huang, J. Harries, and J. Murray, 2016: On the detection of robust multidecadal changes in earth's outgoing longwave radiation spectrum. *Journal of Climate*, **29 (13)**, 4939–4947.

Boughorbel, S., F. Jarray, and M. El-Anbari, 2017: Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, **12 (6)**, e0177 678.

Chen, X., X. Huang, and M. G. Flanner, 2014: Sensitivity of modeled far-ir radiation budgets in polar continents to treatments of snow surface and ice cloud radiative properties. *Geophysical Research Letters*, **41 (18)**, 6530–6537.

Chen, X., X. Huang, and X. Liu, 2013: Non-negligible effects of cloud vertical overlapping assumptions on longwave spectral fingerprinting studies. *Journal of Geophysical Research: Atmospheres*, **118 (13)**, 7309–7320.

Chollet, F., and Coauthors, 2015: Keras. https://keras.io.

Cossich, W., T. Maestri, D. Magurno, M. Martinazzo, G. Di Natale, L. Palchetti, G. Bianchini, and M. Del Guasta, 2021: Ice and mixed-phase cloud statistics on the antarctic plateau. *Atmospheric Chemistry and Physics*, **21 (18)**, 13 811–13 833.

Crane, K., and R. Barry, 1984: The influence of clouds on climate with a focus on high latitude interactions. *Journal of climatology*, **4 (1)**, 71–93.

Ebert, E. E., and J. A. Curry, 1992: A parameterization of ice cloud optical properties for climate models. *Journal of Geophysical Research: Atmospheres*, **97 (D4)**, 3831–3836.

Fouquart, Y., 1988: Radiative transfer in climate models. *Physically-Based Modelling and Simulation of Climate and Climatic Change*, 223–283.

Frey, R. A., S. A. Ackerman, Y. Liu, K. I. Strabala, H. Zhang, J. R. Key, and X. Wang, 2008: Cloud detection with modis. part i: Improvements in the modis cloud mask for collection 5. *Journal of atmospheric and oceanic technology*, **25 (7)**, 1057–1072.

Gupta, K., A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, and R. Hartley, 2020: Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800*.

Huang, X., X. Chen, B. J. Soden, and X. Liu, 2014: The spectral dimension of longwave feedback in the cmip3 and cmip5 experiments. *Geophysical Research Letters*, **41 (22)**, 7830–7837.

Huang, X., X. Chen, D. K. Zhou, and X. Liu, 2016: An observationally based global band-by-band surface emissivity dataset for climate and weather simulations. *Journal of the Atmospheric Sciences*, **73 (9)**, 3541–3555.

Jeppesen, J. H., R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, 2019: A cloud detection algorithm for satellite imagery based on deep learning. *Remote sensing of environment*, **229**, 247–259.

Kay, J. E., T. L'Ecuyer, H. Chepfer, N. Loeb, A. Morrison, and G. Cesana, 2016: Recent advances in arctic cloud and climate research. *Current Climate Change Reports*, **2 (4)**, 159–169.

Krogh, A., 2008: What are artificial neural networks? *Nature biotechnology*, **26 (2)**, 195–197.

Lin, S.-J., 2004: A "vertically lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, **132 (10)**, 2293–2307.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár, 2017: Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, C., S. Yang, D. Di, Y. Yang, C. Zhou, X. Hu, and B.-J. Sohn, 2021: A machine learning-based cloud detection algorithm for the himawari-8 spectral image. *Advances in Atmospheric Sciences*, 1–14.

Liu, X., W. L. Smith, D. K. Zhou, and A. Larar, 2006: Principal component-based radiative transfer model for hyperspectral sensors: Theoretical concept. *Applied Optics*, **45 (1)**, 201–209.

Liu, Y., S. A. Ackerman, B. C. Maddux, J. R. Key, and R. A. Frey, 2010: Errors in cloud detection over the arctic using a satellite imager and implications for observing feedback mechanisms. *Journal of Climate*, **23 (7)**, 1894–1907.

Lubin, D., and E. Morrow, 1998: Evaluation of an avhrr cloud detection and classification method over the central arctic ocean. *Journal of Applied Meteorology*, **37 (2)**, 166–183.

L'Ecuyer, T. S., and Coauthors, 2021: The polar radiant energy in the far infrared experiment: A new perspective on polar longwave energy exchanges. *Bulletin of the American Meteorological Society*, 1–46.

Maestri, T., W. Cossich, and I. Sbrolli, 2019: Cloud identification and classification from high spectral resolution data in the far infrared and mid-infrared. *Atmospheric Measurement Techniques*, **12 (7)**, 3521–3540.

McClatchey, R. A., 1972: *Optical properties of the atmosphere*. 411, Air Force Cambridge Research Laboratories, Office of Aerospace Research ....

Mekanik, F., M. Imteaz, S. Gato-Trinidad, and A. Elmahdi, 2013: Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology*, **503**, 11–21.

Niebler, S., A. Miltenberger, B. Schmidt, and P. Spichtinger, 2021: Automated detection and classification of synoptic scale fronts from atmospheric data grids. *Weather and Climate Dynamics Discussions*, 1–28.

NSIDC, 2022: Sea ice fraction [dataset]. URL https://masie_web.apps.nsidc.org/pub//DATASETS/NOAA/G02135/, (date of last access: 2022-02-11).

Ou, S.-c., and K.-N. Liou, 1995: Ice microphysics and climatic temperature feedback. *Atmospheric Research*, **35 (2-4)**, 127–138.

Palchetti, L., G. Bianchini, G. Di Natale, and M. Del Guasta, 2015: Far-infrared radiative properties of water vapor and clouds in antarctica. *Bulletin of the American Meteorological Society*, **96 (9)**, 1505–1518.

Paul, S., and M. Huntemann, 2020: Improved machine-learning based open-water/sea-ice/cloud discrimination over wintertime antarctic sea ice using modis thermal-infrared imagery. *The Cryosphere Discuss., https://doi. org/10.5194/tc-2020-159, in review*.

Peterson, C. A., X. Chen, Q. Yue, and X. Huang, 2019: The spectral dimension of arctic outgoing longwave radiation and greenhouse efficiency trends from 2003 to 2016. *Journal of Geophysical Research: Atmospheres*, **124 (15)**, 8467–8480.

Powers, D. M., 2020: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Ramos, D., J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, 2018: Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, **20 (3)**, 208.

Raschke, E., 1987: International satellite cloud climatology project (isccp) workshop on cloud algorithms in the polar regions. Tech. rep., Report WMO/TD-170). World Meteorological Organization, Geneva.

Saito, M., P. Yang, X. Huang, H. E. Brindley, M. G. Mlynczak, and B. H. Kahn, 2020: Spaceborne middle-and far-infrared observations improving nighttime ice cloud property retrievals. *Geophysical Research Letters*, **47 (18)**, e2020GL087 491.

Santurkar, S., D. Tsipras, A. Ilyas, and A. Madry, 2018: How does batch normalization help optimization? *Proceedings of the 32nd international conference on neural information processing systems*, 2488–2498.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15 (1)**, 1929–1958.

Stevens, B., and Coauthors, 2019: Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, **6 (1)**, 1–17.

Tans, P., and R. Keeling, 2022: Atmospheric co2 [dataset]. URL https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html, (date of last access: 2022-02-11).

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.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, **147 (6)**, 2261–2282.

Yang, P., and Coauthors, 2003: Spectral signature of ice clouds in the far-infrared region: Single-scattering calculations and radiative sensitivity study. *Journal of Geophysical Research: Atmospheres*, **108 (D18)**.