Developing and interpreting Neural Network alternatives to multi-layer cloud detection in Himawari-8/9

Mathilde E H Ritman

Honours Thesis presented for the B.Sc. (Honours) degree
in the School of Mathematics, Monash University

Supervised by: Steven Siems[1] and Caroline Poulsen[2]
[1]School of Mathematics, Monash University
[2]Australian Bureau of Meteorology

October 2022

Abstract

Clouds are the dominant control of surface radiative energy budgets on Earth. Present-day poor simulation of clouds, particularly over the Southern Ocean (SO), is the predominant source of uncertainty in modelled radiative budgets, therefore limiting the accuracy of equilibrium climate sensitivity estimates and global climate models. As such, SO cloud biases have been the topic of extensive research in recent years. A consequence of this has been the observed connection between the occurrence of multi-layered clouds and the high modelled radiative bias in the SO. However, study of multi-layer cloud climatology has been limited by poor observational accuracy in passive satellite instruments. To overcome this, we develop two deep learning models for the Himawari-8/9 Advanced Himawari Imager, trained on January-June 2019 merged lidar-radar observations over the SO. In particular, we extend the existing literature through the development of a 1-dimensional convolutional neural network, convolved in time, thus incorporating temporal information from successive satellite observations. We place an emphasis on the defensibility and reliability of the developed models, to be assessed through the mathematical model interpretability scheme, Shapley value theory, whose derivation is detailed in full. The convolutional network outperforms the feed forward algorithm, achieving a validation set accuracy of 79.5 % compared to 70.4 %. Both models offer significant accuracy improvements over operational metrics, with multi-layer cloud Probability of Detection scores of $\sim$60 % (existing operational accuracy is $<$50 %). The results of our models offer means to obtain multi-layer cloud occurrence data with unprecedented spatial density and temporal resolution, offering advancements in the exploration of the link between these cloud systems and the large SO radiative bias. Further, the improved performance of the convolutional network highlights the value of incorporating temporal information for future modelling and detection efforts in atmospheric remote sensing.

Acknowledgements

I would like to acknowledge the contributions of Daniel Robbins and Arathy Aneeshkumar Kurup, who assisted in the provision of data and scientific and computational expertise. I would also like to acknowledge the Australian Research Council Centre of Excellence for Climate Extremes, for their financial support.

Publications while enrolled

Published conference proceedings

Ritman, M., Chen, Y., Taylor, J., Yebra, M., Leavesley, A. and Prakash, M. (2022). Towards affordable, dynamic and large-scale fuel hazard assessment in Australia using space-borne LiDAR. Australian Fire and Emergency Services Conference, Adelaide, 23/08/2022.

Hague, B., Jones, D., Jakob, D., McGregor, S., Reef, R., Ritman, M. (2022). The tide is high: new insights into coastal flooding around Australia (and the world). Australian Fire and Emergency Services Conference, Adelaide, 23/08/2022.

Accepted peer reviewed scientific literature

Ritman, M., Hague, B., Katea, T., Vaaia, T., Ngari, A., Smith, G., Jones, D. and Tolu, L. (2022). An assessment of tidal flooding for Pacific small island nations: insights from the Pacific Sea Level and Geodetic Monitoring Project. Journal of Southern Hemisphere Earth System Science.

Accepted peer reviewed scientific reports

McGree, S., G. Smith, E. Chandler, N. Herold, Z. Begg, Y. Kuleshov, P. Malsale and M. Ritman (2022). Climate and Ocean Variability, Extremes and Change in the Western Tropical Pacific: Updated Country Reports. Climate and Oceans Support Program in the Pacific. Pacific Community, Suva, Fiji.

Contents

1 Introduction

1.1 Motivation

Clouds remain the largest source of uncertainty in global climate models (e.g., Masson-Delmotte et al., 2021; Narendra et al., 1974; Webster and Stephens, 1984) and equilibrium climate sensitivity[1] estimates (Sherwood et al., 2020). Global climate models, weather models and reanalysis products have struggled to provide unbiased simulations of the radiative energy budget over the Southern Ocean (SO), and this has been attributed to poor representation of cloud vertical structure (Bodas-Salcedo, 2014; Bodas-Salcedo et al., 2016; Flato, 2013; Trenberth and Fasullo, 2010; Zhang et al., 2005). Trenberth and Fasullo (2010) further found that projected changes in radiative energy budgets were strongly related to present-day cloud simulation errors (also Bony et al., 2006; Williams and Tselioudis, 2007), meaning that improved simulation of modern Southern Hemisphere cloud is necessary to address the significant uncertainties in the modelled SO radiative budget.

The complex influence of clouds on incoming and outgoing radiation is the dominant control on the Earth's radiative balance. Clouds act to both increase and decrease the amount of radiation at the surface through reflection and absorption of radiation (see Wielicki et al., 1995). The balance of radiation reflected or absorbed depends on physical cloud properties, such as cloud phase, temperature, height, coverage and albedo. The occurrence of multiple cloud layers (e.g., Figure 1) strongly influences both the magnitude and vertical profile of radiative balance (Slingo and Slingo, 1988; Stephens, 2002; Wang and Rossow, 1998). Additionally, satellite-based retrievals of cloud physical parameters, which assume the existence of only a single cloud layer, are significantly impacted by the existence of multiple vertical cloud layers (e.g., Marchant et al., 2020; Platnick et al., 2017). This means correctly detecting multi-layer cloud occurrences is critical for ensuring accurate retrieval of radiative properties.

Mace et al. (2009) described the global occurrence of cloud layers using merged Cloud Profiling Radar (CPR) Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) data. They found that the occurrence of multi-layer cloud exhibited strong latitudinal dependencies and distinct structural characteristics over the SO. In particular, they observed that these characteristics were spatially aligned with the key regions of radiative balance uncertainty outlined by Trenberth and Fasullo (2010) (Figures 2 and 3). Indicating that radiative properties associated with multi-layer cloud scenes, which are poorly captured in operational satellite data, may be linked to the large radiative bias persistent in SO simulated climates.

Currently, the optimal means of obtaining cloud vertical structure information is by using merged CPR-CALIOP data, as employed by Mace et al. (2009). Despite their high accuracy, however, these active polar-orbiting instruments are limited in their spatial coverage, temporal resolution and length of data. For instance, the CPR and CALIOP instruments orbit earth every 60 minuets, however revisit a location after a period of 16 days. In contrast, passive geostationary satellites, such as the Advanced Himawari Imager (AHI), offer frequent (every ten minuets) global scale cloud observations, and therefore present an opportune method of expanding the climatological analyses of Mace et al. (2009). While the AHI is our focus here, there is in fact a constellation of geostationary satellites which covers most of the globe (except the poles). We consider the AHI, as it covers the Australian-region of the SO, but our results would be applicable to the other satellites and regions. Despite the clear observational advantage of geostationary satellites, the inability of these instruments to directly provide cloud vertical structure information (see Section 2.3) has limited their implementation in multi-layer cloud analyses.

[1] A key metric used to assess the sensitivity of the earth-atmosphere system to greenhouse gases, defined as the amount of global mean temperature change after a doubling of the atmospheric CO2 concentration.

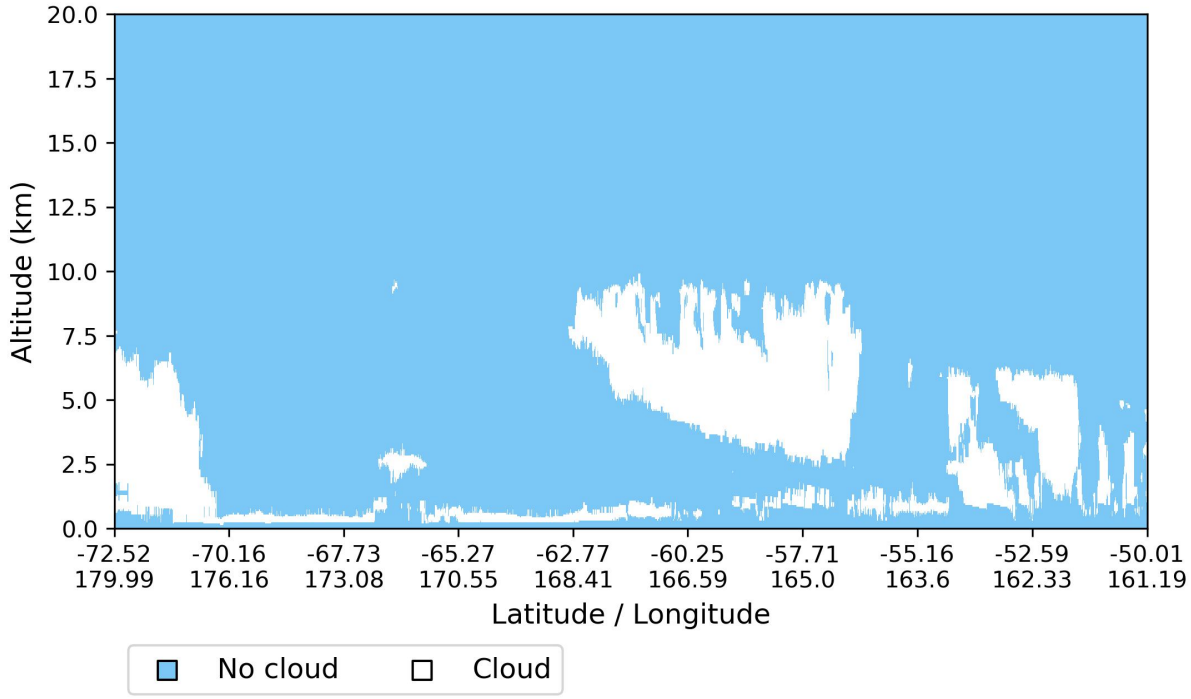

Figure 1: Example of a multi-layer cloud occurrence as observed by the CPR-CALIOP merged product on 2019-01-20 03:06.

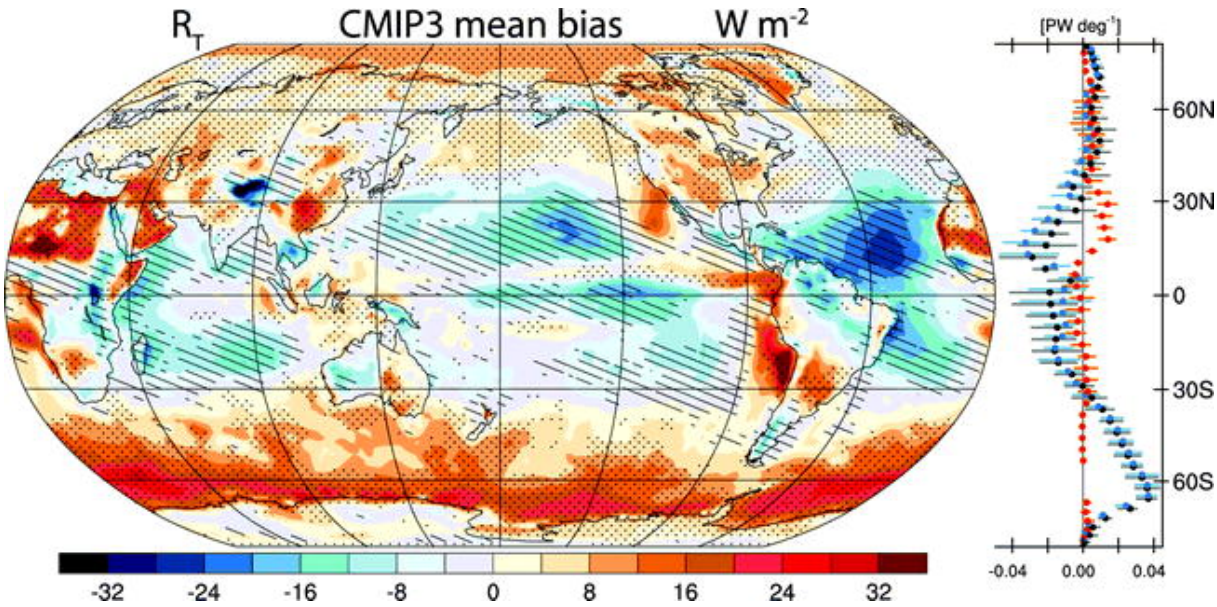

Figure 2: From Trenberth and Fasullo, 2010: top-of-atmosphere net downward radiation biases relative to observations for 1990-99 in Wm^{-2}, where stippled (hatched) regions correspond to regions in which at least three quarters of the models share a common positive (negative) bias. (right) The model zonal mean is given (dots) with the 25th to 75th percentile range (lines) over land (red), ocean (blue), and all (black) surfaces.

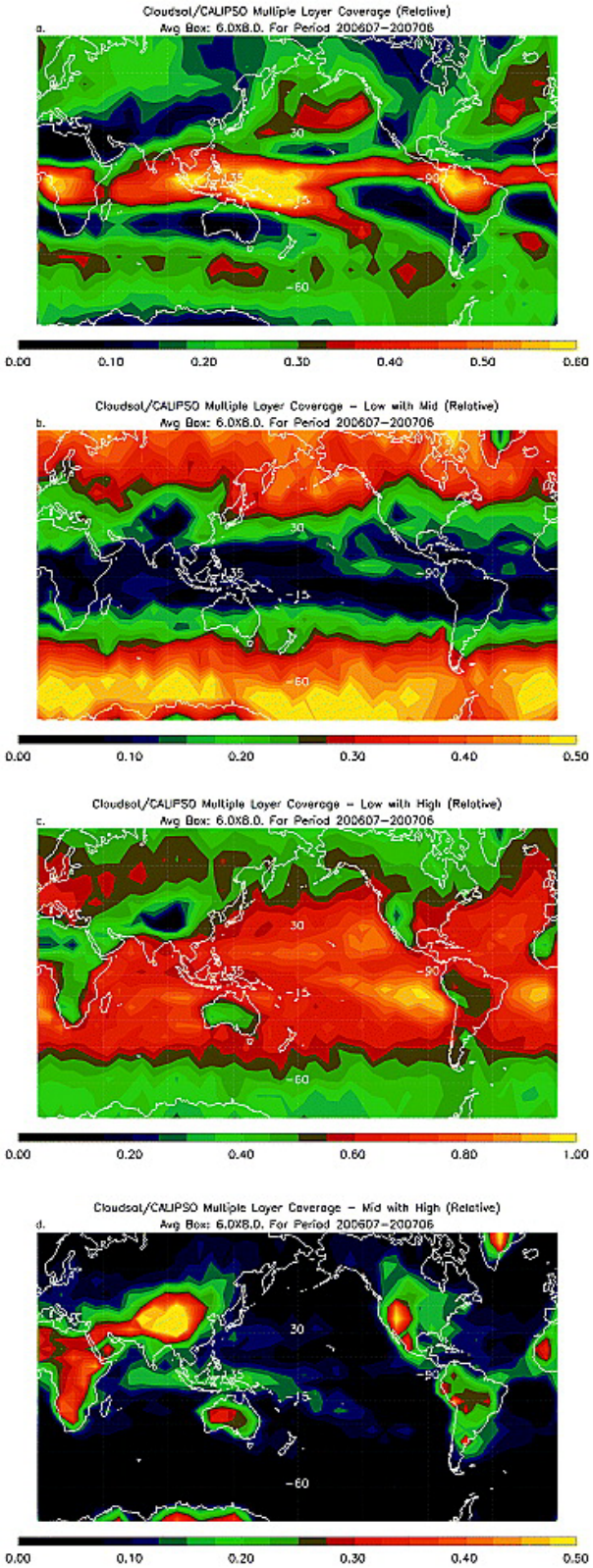

Figure 3: From Mace et al., 2009: multi-layer cloud coverage. (a) Coverage of multi-layer clouds from the merged CPR-CALIOP data relative to the total number of profiles. (b) As in (a) for number of events of low-based layers (base $<$ 3 km) occurring with middle-based layers (base 3-6 km). (c) As in (b) for low-based layers occurring with high-based layers (layer base ¿ 6 km). (d) As in (b) for middle-based layers occurring with high-based layers. The averaging period is from July 2006 through June 2007 in 6 by 8 latitude-longitude averaging regions.

Over the years, a number of threshold-based decision tree algorithms have been proposed to identify multi-layer clouds in passive satellite data (e.g., Desmons et al., 2017; Jin and Rossow, 1997; Joiner et al., 2010; Pavolonis and Heidinger, 2004; Wind et al., 2010). Despite these efforts, operational multi-layer cloud indicators are limited in their implementation and accuracy. For instance, Marchant et al. (2020) found the hit rate of the Moderate Resolution Imaging Spectroradiometer (MODIS) (Oreopoulos et al., 2017) multi-layer cloud flag, compared to CPR-CALIOP data, to be just 34 %. More recently, Artificial Intelligence (AI) methods have been employed, offering significant increases in the accuracy compared to the existing tree-based methods (namely Li et al., 2022; Tan et al., 2021). However, these emerging AI approaches are yet to consider temporal information (i.e., incorporate data from preceding AHI observations), an approach that may offer further increases in model accuracy. Nor have they conducted model interpretability analyses, a mathematical basis from which influential variables can be determined and model reliability can be assessed. Further, these studies have focused on global-scale operational retrievals of cloud physical parameters in multi-layer scenes, and have not assessed the accuracy of their algorithms for the SO region.

1.2 Project aims

The main objective of our research is to develop a defensible (where the physicality of predicted results and method are interrogated) neural network model for the detection of multi-layer cloud scenes in Himawari-8/9 AHI satellite imagery. In particular, we aim to

1. Develop a feed forward neural network to classify cloud scenes over the SO;
2. Determine and develop an alternate neural network algorithm capable of incorporating temporal information from successive satellite scenes; and
3. Prove and use Shapley value theory to interpret and assess the developed models.

2 Scientific background

2.1 Remote sensing of clouds

Geostationary satellite observations offer the only means of obtaining high temporal resolution global surveys of atmospheric conditions including clouds. In the last few decades, geostationary satellites located along the equator have been providing operational weather observations for large spatial regions and at high temporal scales (Eyre et al., 2022; Eyre et al., 2020). Over the years, a suite of satellite-based methods for determining key cloud and precipitation parameters has evolved (see Stephens and Christian, 2007). Many of these methods result from processing the emission and scattering of Electromagnetic (EM) radiation by the atmosphere and Earth's surface. Satellite instruments can either rely on external sources of radiation to illuminate observables (passive sensors) or they can be equipped with their own source of radiation (active sensors).

Passive sensors, such as Himarwari-8/9's Advanced Himawari Imager (Bessho et al., 2016), measure outgoing (from the Earth's surface and atmosphere) EM radiation in the ultraviolet to microwave bands of the EM spectrum. These passive satellite sensors observe radiation in a specified set of spectral domains, chosen for their sensitivity to key cloud and atmospheric parameters (Stubenrauch et al., 2013). Different wavelengths are sensitive to different parameters as each wavelength of radiation is absorbed, transmitted or scattered at different amounts by molecules in the atmosphere and Earth's surface. For example, wavelengths at which atmospheric gases predominantly transmit radiation are called "windows", as radiation emitted by the Earth's surface or cloud layers at these wavelengths can pass through the gaseous atmosphere and be observed by the satellite radiometer.

Following the *detection* of cloud, physical parameters are *retrieved* using a forward radiative transfer model, which converts the input signal to an output constituting the 'retrieved state' (Stephens and Christian, 2007). These functions are the largest source of error in retrieved cloud physical properties such as phase, effective radius and cloud droplet number and concentration.

While passive imagers, such as AHI, provide high spatial coverage and temporal resolution, they are limited by the reliance of many of their spectral channels on radiation emitted during the solar day. Further, they can only observe in two dimensions, meaning that gaining insight into cloud vertical structure is challenging (see Section 2.3). Instead, vertical structural information is typically obtained using active sensors, such as lidar (Light Detection and Ranging) or radar (Radio Detection and Ranging) instruments. These technologies emit pulsed radiation at specific wavelengths, the density and timing of returned radiation are then observed, and a vertical profile of atmospheric phenomena can be obtained.

Atmospheric lidar measurements are highly sensitive to water vapour, enabling them to detect optically thin cloud. However, when cloud optical depth[2] is too great the lidar beam is significantly attenuated and lower-level vertical information, such as cloud base height or lower cloud layers, is unattainable. In such instances, radar instruments can be employed to observe the cloud base or penetrate to lower altitudes (Stephens, 2002).

[2] A dimensionless coordinate that refers to the amount of energy passed through a cloud layer without being attenuated.

Table 1: Himawari-8/9 AHI channel descriptions (Bessho et al., 2016; Zhang et al., 2018).

	Band	Central wavelength (μm)	Spatial resolution (km)	Observable
Visible	1	0.47	1	Land surface information, aerosols over land
	2	0.51		Land surface information
	3	0.64	0.5	Land surface information, aerosols over water
Near-infrared	4	0.86	1	Cloud phase, particle size and optical thickness
	5	1.6	2	Cloud phase, particle size and optical thickness
	6	2.3		Cloud phase, particle size and optical thickness
Infrared	7	3.9		Low-level cloud, fog
	8	6.2		High-altitude water vapour
	9	6.9		Mid-altitude water vapour
	10	7.3		Low-altitude water vapour
	11	8.6		Total atmospheric water, cloud phase, dust
	12	9.6		Ozone
	13	10.4		Atmospheric window, low-level clouds
	14	11.2		Atmospheric window
	15	12.4		Atmospheric window
	16	13.3		Air temperature, cloud height

2.2 Instrument description

2.2.1 Advanced Himawari Imager (AHI)

The AHI is a next-generation passive imager aboard the Japanese Meteorology Agency's geostationary satellite, Himawari-8/9 (Bessho et al., 2016). The satellite, which succeeded a previous generation of Japanese satellites, began operational activity in July 2015 and is expected to operate until 2029. The AHI observes 16 observation bands, three visible, three near-infrared, and ten infrared (specifications are given in Table 2.2.1). Full-disk images are taken every 10 minutes and are centred at 140.7° E on the Equator with a radius of view of approximately 80°.

Solar reflectance differences between visible and near-infrared bands observe land surface conditions, such as snow/ice cover and vegetation. Cloud physical parameters are observed by the visible, near-infrared bands and infrared atmospheric window bands observe cloud thermal properties, by capturing long-wave radiation emitted by the surface or clouds and using Planck's Law.

2.2.2 Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP)

CALIOP is an active lidar instrument mounted aboard the joint NASA (United States) and CNES (France) venture satellite, the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) (Winker et al., 2009). CALIOP began observations in 2006 as a member of the A-Train, a constellation of polar-orbiting satellites with synchronised orbits (Stephens, 2002).

CALIOP emits radiation at 532 nm and 1064 nm wavelengths to return nadir-view (directly downward facing) vertical profiles of aerosols and clouds. Three receiver channels observe the returned signals. A 1064 nm 'backscattered intensity' and two 532 nm polarised channels, with perpendicular polarisation. As with other active sensors, the pulsed radiation source limits spatial coverage to the radius of the radiation pulse when it reaches the Earth's surface. The

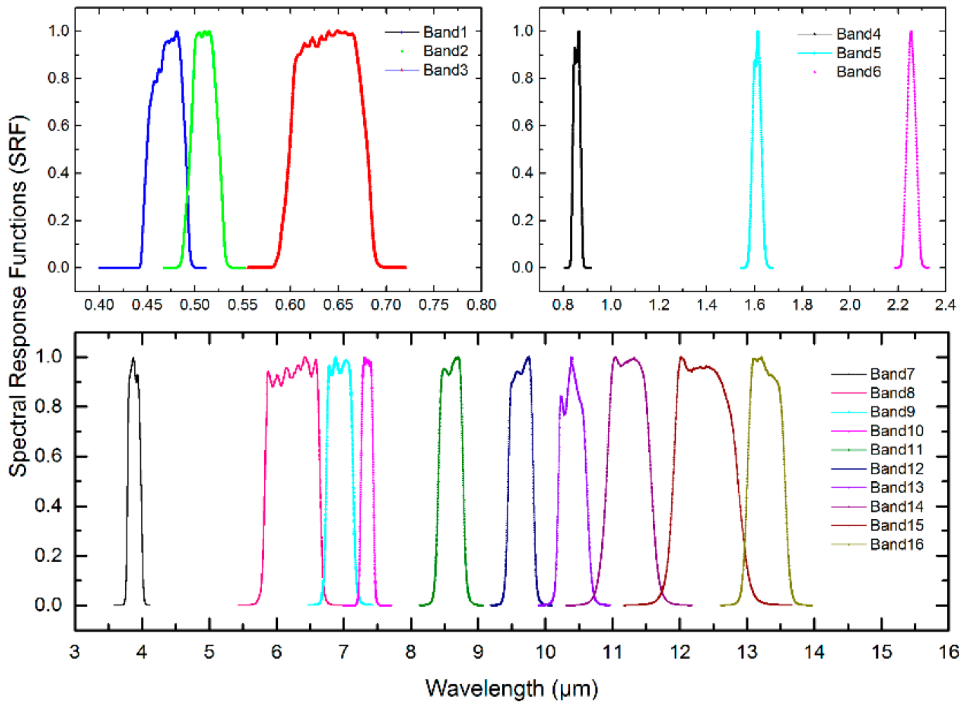

Figure 4: Spectral response functions (radiative response) for each Himarwari-8/9 AHI observed channel (from Zhang et al., 2018).

CALIOP instrument observes a 90 m surface footprint and measurements are taken every 333 m along the surface track; these are then reprocessed to 1 and 5 km along track frequency for improved accuracy, with a horizontal resolution of 1 km and vertical resolution of 60 m below an altitude of 20.1 km.

2.2.3 Cloud Profiling Radar (CPR)

The active radar instrument, CPR, aboard NASA's CloudSat satellite was launched in 2006. CPR is a 3200 μm radar with 500 m vertical resolution and 1.7 km horizontal resolution (Im et al., 2005). Also a part of the A-Train (Stephens, 2002), CPR's nadir-pointing radar observations are typically employed alongside CALIOP to provide comprehensive vertical cloud-layer information (see Stubenrauch et al., 2013). Since 2011, the CPR has been operating in daytime only mode, as such, we consider the daylight period only in our analyses. Together, CPR and CALIOP offer the state-of-the-art of cloud vertical profile observation.

In February 2018, CloudSat exited the A-Train in response to a mechanical failure that posed a collision risk with another A-Train satellite (Atkinson, 2018). Due to the scientific value of their co-located observations, the CALIPSO satellite was subsequently removed from the A-Train orbit to re-join CloudSat at the new lower altitude orbit, forming the 'C-Train' in September 2018.

2.2.4 The CPR-CALIOP merged product

Merged CPR-CALIOP data are taken from the Colorado State University's Version 5 2B-CLDCLASS-LIDAR product (https://www.cloudsat.cira.colostate.edu/data-products/2b-cldclass-lidar). The product combines CloudSat CPR and CALIPSO CALIOP measurements to classify cloud types for each observed cloud layer. The result are observations with a horizontal resolu-

tion of 3.2 km and a vertical resolution of 100 m. For our purposes, we neglect their cloud-type assignments and consider vertical cloud detection profiles alone. This approach is standard in recent literature (e.g., Li et al., 2011; Li et al., 2022; Marchant et al., 2020; Oreopoulos et al., 2017; Wang et al., 2016).

2.3 Review of past work

The problem of the modelled Southern Ocean (SO) radiative bias is long-standing (Wild et al., 1995) and many efforts have been made to provide insight into possible resolutions. Despite this, recent model experiments highlight the persistent uncertainty in this field. For instance, Fiddes et al. (2022) found that correctly simulating cloud type over the SO resulted in no improvements to radiative biases. Also, Kuma et al. (2022) found a negative relationship between modelled cloud type errors and equilibrium climate sensitivity, suggesting that either better representation of clouds results in higher equilibrium climate sensitivity estimates than are currently supported in the literature, or present day cloud simulation errors must be negatively correlated to future errors.

Global cloud physical and modelled properties are far from spatially homogeneous and the SO observes many cloud climatological characteristics that are distinct from elsewhere in the globe. For instance, Huang et al. (2012) found that cloud top phase over the SO was predominantly super-cooled liquid water, in contrast to the ice-top conditions in other regions. Schuddeboom et al. (2019) found a latitudinal dependence of the modelled radiative bias in response to changes in cloud phase parameterisations. Further latitudinal dependencies were uncovered in observational studies undertaken by Mace et al. (2021) and Truong et al. (2022), the latter showing that discrepancies in cloud vertical structural properties at high latitudes were particularly large in multi-layer scenes. The different radiative effects of single and multi layer cloud scenes have been detailed in the literature (e.g., Li et al., 2011; Slingo and Slingo, 1988; Stephens, 2002; Wang and Rossow, 1998). Further, it is well documented that cloud physical parameters derived under the incorrect assumption of single-layer cloud have consistent errors (e.g., Marchant et al., 2020; Platnick et al., 2017). In response, Mace et al. (2009) conducted a global climatology of vertical cloud layers using the first year of CPR-CALIOP data, finding a strong latitudinal dependence in global and SO multilayer cloud occurrence and structural characteristics (Figure 3). Considered in combination, these results have strong implications for the connection between poor multi-layer cloud detection and related cloud property retrieval errors and the SO radiative bias.

Despite early recognition of their importance for accurate cloud retrievals (Stephens and Webster, 1984), and later for the SO radiative bias problem (Mace et al., 2009), the difficulty of observing cloud vertical structure in passive satellite data has limited the progression of multi-layer cloud assessments globally. Early efforts to identify multi-layer cloud occurrence in passive satellite data adopted threshold-based decision tree approaches using multi-spectral information, CO2-slicing methods and ancillary data. In 1994, Baum et al. presented a decision tree algorithm to identify upper semitransparent cirrus overlaying large-scale stratus cloud, using a case study of NOAA 11 advanced very high-resolution radiometer (AVHRR) imagery. Their approach used CO2-slicing and spatial coherence methods with ancillary sounder data to retrieve key cloud properties. Multi-layer clouds were then identified using the “fuzzy logic” classifier model developed by Tovinkere et al. (1993). This classifier was created to distinguish cloud types using AI trained on human-labelled scenes, and the ”fuzzy logic” approach allowed for several simultaneous cloud type classifications. In 1995, Baum et al. applied their model to NOAA Advanced Very High Resolution Radiometer (AVHRR) data, providing the first classification of multi-layer scenes in the AVHRR.

Later, Huang et al. (2005) developed a decision tree algorithm using surface microwave

Table 2: Probability of detection (%) for multi-layer clouds in past work.

	MODIS 6 (Marchant et al., 2020)	Tan et al., 2021	Li et al., 2022
Day-only	-	70	70
Day-night	34	64	60

radiometer measurements, showing consistent improvements in cloud optical depth and ice water path retrievals (Huang et al., 2006). This model was extended by Minnis et al. (2007), showing that overall increases in retrieval accuracy compared to assumed-single layer retrievals reached 42 %. Alternate approaches were presented by Baum and Spinhirne (2000) and later Nasiri and Baum (2004), for the detection of optically thin cirrus overlying low-level cloud (also, Heidinger and Pavolonis, 2005; Pavolonis and Heidinger, 2004).

More recently, the deployment of the CPR and CALIOP instruments along with the A-Train (Stephens et al., 2018) introduced a new generation of cloud vertical structure research. Multi-layer cloud detection algorithms were developed using empirically determined thresholds from CPR data for the MODIS (Joiner et al., 2010; Platnick et al., 2017; Wind et al., 2010) and Polarization and Directionality of the Earth Reflectance (POLDER) (Desmons et al., 2017; Yao et al., 2010) radiometers. In particular, an operational multi-layer cloud indicator was developed by Wind et al. (2010) and updated as described by Platnick et al. (2017). Assessments of the accuracy of these latest decision tree detection algorithms have been undertaken (Desmons et al., 2017; Marchant et al., 2020; Wang et al., 2017). From this, Desmons et al. (2017) found probability of detection (PoD) of their POLDER algorithm for the 2006-2010 period was 47 % where, in the same period, the MODIS 6 algorithm saw PoD of 46 %. Evaluated against CPR-CALIOP results, Wang et al. (2016) found that where CPR-CALIOP reports 47.4/25.5 % single-layer/multi-layer clouds, MODIS reports 26.7/14.0 %. Despite the generally low accuracy, Marchant et al. (2020), who found that the MODIS 6 flag and CPR-CALIOP data agreed with multi-layer classifications 33.73 % of the time and disagree 20.04 % of the time, suggested the MODIS flag was nonetheless reasonably skilled at improving cloud parameter retrievals in MODIS.

With the recent expansion in AI expertise, and the increased accessibility of AI model architectures and programming tools, new means of accurately identifying multi-layer cloud scenes have become possible. Namely, two recent studies have presented deep and machine learning approaches to multi-layer cloud identification in passive AHI scenes using truth labels from coincident CPR-CALIOP shots. In 2021, Tan et al. presented an assessment of four different models, Random Forest, K-Nearest Neighbour, Artificial Neural Network and Support Vector Machines. For each approach, two models were developed, one that includes solar AHI channels and thus limits application to daytime detection, and a second that excludes these channels and offers both day and night detection. Results for the Random Forest model, which saw the best overall accuracy by a small margin, are given in Table 2.3. Li et al. (2022) adopted a Deep Neural Network approach to detect multi-layer clouds and classify cloud phase. They argued that a key limitation of the models developed by Tan et al. (2021) was the neglection of background atmospheric information, for which they included data from the European Centre for Medium-Range Weather Forecasts' ERA5 Reanalysis product. Both studies considered the full AHI field of view (and may therefor perform less well over the SO) and used 2016 collocated AHI and CPR-CALIOP data for training, and 2017 data for testing. Other studies have also explored multi-layer cloud detection using AI for other geostationary instruments (e.g., Haynes et al., 2022; Minnis et al., 2019).

These approaches offer unprecedented accuracy for the identification of multi-layer clouds in passive satellite instruments. Thus, the high spatial density and temporal resolution offered by passive instruments such as the AHI can now be leveraged to provide in-depth insights into

multi-layer cloud climatology and properties over the SO, extending the results of key literature such as Mace et al. (2009).

Rather than employ the models presented by Tan et al. (2021) or Li et al. (2022) over the SO, we propose the development of an additional neural network approach specifically optimised for the SO. This is important as, due to the high viewing angle of geostationary instruments such as the AHI over this region, the accuracy of existing algorithms optimised for broader regions may be compromised. We also explore the addition of temporal information to the neural network, an approach yet to be undertaken in the literature that may offer further advancements. Further, a key outcome this research is the mathematical interpretation of model results, which sheds light on the physical quantities most relevant for accurate multi-layer cloud detection.

3 Mathematical basis

3.1 Feed Forward and Convolutional Neural Networks

Increasingly, AI alternatives to threshold-based cloud detection and retrieval methods are becoming state-of-the-art. These approaches have been employed to improve cloud mask accuracy, cloud top height or phase retrieval and vertical structure information (e.g., Goosse et al., 2018; Li et al., 2022; Poulsen et al., 2020; Robbins et al., 2022; Tan et al., 2021; Wang et al., 2020). A key reason for this is the lack of physics-based mathematical relationships between observed spectral channels and atmospheric quantities of interest. Indeed, if some functional relationship exists between a set of spectral inputs observed by passive satellites and a cloud layer classification, it is highly improbable that it can be explicitly determined from empirical or physical information. This challenge is compounded by the high dimensionality of available passive satellite spectral data and the likely nonlinearity of such a function. Highlighting this, existing efforts to derive empirical relationships that enable accurate classification of multi-layer scenes have been limited in their accuracy to less than 50 % (PoD) (Desmons et al., 2017; Marchant et al., 2020; Wang et al., 2016).

In light of this, deep learning models, such as neural networks offer a compelling alternative. These statistical models require no prior assumptions about the functional relationship between a set of predictor variables and their prediction (i.e., they are non-parametric). By iteratively minimising a loss function between predicted variables and observed conditions, these models approximate a functional relationship that best fits the given data. This provides a best-statistical-guess that can outperform human intuition, depending on the quality and quantity of available data.

A standard network structure involves a series of connected processors, called neurons. Each neuron corresponds to a real-valued "activation" function, that takes input from the preceding neurons, and computes an output that is passed to the next neuron layer. Supervised learning is the process by which weights and biases passed to each neuron are optimised to to minimise a loss function, and provide an accurate prediction.

Neural networks are a type of *deep learning* model, as they consist of "hidden" layers where activation functions are optimised unobserved (Figure 5). The number of nodes in a hidden layer refers to the number of activation functions computed at that layer. For example, if the network contains one hidden layer with five nodes (Figure 5), five activation functions will be computed and the results of each will be given to the output layer for the final activation and prediction, $\hat{y}$.

Each activation function, $f(x)$, maps a linear combination of the predictor variables ("features"), with weights, w_j, and a bias, b, to be optimised (Equation 1). A differentiable, non-linear activation function is chosen for the hidden layers, typically one of hyperbolic tangent, the sigmoid function (Equation 2) or the Rectified Linear Unit (ReLu) (Equation 3). ReLu is perhaps the most common choice, due to it's simple and efficient computation, and reduced susceptibility to convergence issues during optimisation. The choice of activation function for the output layer is determined by the desired output format of the prediction. For example, a linear activation will output a real-valued prediction, whereas a sigmoid will output a value between 0 and 1, which can be interpreted as a probability and is used in classification tasks. The non-linearity of the activation function ensures the model is able to capture nonlinear interactions between variables. Without the nonlinear activation, the neural network would collapse to a simple linear model.

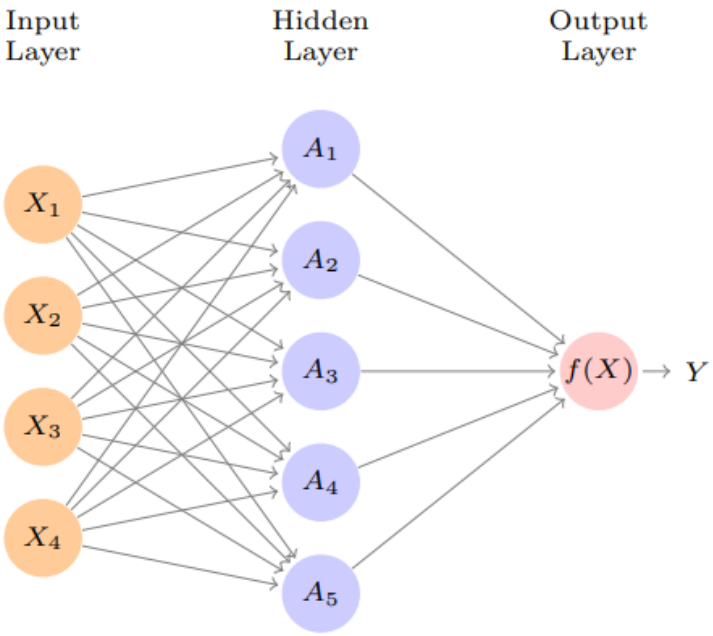

Figure 5: Example of a neural network architecture with one hidden layer that takes four input variables processes data through five activation functions and returns a prediction, $f(X) = \hat{y}$ (from Tibshirani, 2013).

$$y = f(z) = f\left(\sum_{j=1}^{p} w_j x_j + b\right) \tag{1}$$

Where p is the number of predictor variables (features).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

$$f(z) = \begin{cases} 0 & z <= 0 \\ z & z > 0 \end{cases} \tag{3}$$

Another output layer activation function, and our choice here, is the softmax function, defined

$$f_m(z) = Pr(Y = m|X) = \frac{e^{z_m}}{\sum_{k=1}^{N} e^{z_k}}, \forall m \in \{1, \ldots, N\}, \tag{4}$$

Where N is the number of classes possible for classification.

The vector-valued softmax function outputs N values between 0 and 1, one for the probability of occurrence of each class. For the softmax multi-class classification problem, the cross-entropy loss function (also called log loss or logistic loss) is minimised in optimisation. Defined

$$\mathcal{L}_{CE} = -\sum_{k=1}^{N} y_k \ln \hat{y}_k,$$

Where y_k is the true label and $\hat{y}_k$ the predicted probability for class k,

The cross-entropy loss increases as the predicted probability diverges from the actual label. The continuous differentiablity of the activation functions allows the loss to be minimised through a series of "back-propagation" equations with the weights and biases being updated according to the rule

$$w^l = w^l - \alpha * \frac{\partial \mathcal{L}_{CE}}{\partial w^l}$$
$$b^l = b^l - \alpha * \frac{\partial \mathcal{L}_{CE}}{\partial b^l}$$

For each layer l and learning rate α, where $w^l \in \mathcal{R}^{M \times p}$ and $b^l \in \mathcal{R}^{M \times 1}$ and M the number of neurons in layer l.

This minimisation process is called Gradient Descent. In practice, Gradient Descent is exceptionally time consuming and computationally expensive, as it uses all available training data at every iteration. Hence, alternate minimisation procedures are typically employed, such as Stochastic Gradient Descent, which uses a different random subset of the training data at each iteration, and Adaptative Moment Estimation (Adam). Adam is a modification of SGD, which, instead of using the gradient for the update rule, uses the first and second moments of the gradient. This has the effect of stabilising the minimisation process, as each gradient is computed on a different subset of the training data and may vary significantly. We employ and compare both of these two popular optimisers in this research.

3.1.1 Incorporating temporal information

Some datasets are sequential, such as time series data. In such instances it becomes desirable to incorporate the sequential information into the model, which can be achieved by a *recurrent* neural network or *convolutional* neural network, convoloved in the temporal dimension.

Recurrent networks allow for a model to capture temporal interdependencies by feeding data into successive layers sequentially (Pascanu et al., 2012). However, they have historically suffered from the vanishing and exploding gradient problem, motivating the development of the Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRUs) (Cho et al., 2014; Hochreiter and Schmidhuber, 1997; Murugan, 2018) and deep (Ienco et al., 2017; Minh et al., 2018) or bidirectional (Schuster and Paliwal, 1997) models. Such approaches have been employed in remote sensing, particularly in agricultural and land-mapping domains (e.g., Lakhal et al., 2018; Ndikumana et al., 2018; Sun et al., 2019). Convolutional networks have also seen such implementation, as well as atmospheric uses (e.g., Chen et al., 2017; Goosse et al., 2018; Liu et al., 2016). Where applying a 1-dimensional convolution in the temporal dimension has seen accuracy improvements for time-series classification (e.g., Di Mauro et al., 2017; Goosse et al., 2018; Kussul et al., 2017; Rawat et al., 2021; Zhong et al., 2019). In their experimental review paper, Pelletier et al. (2018) contrasted the two popular approaches, concluding that the 1-D convolutional networks offered higher accuracy compared to the recurrent approach.

Given the comparison results, we propose a convolutional model approach may be most appropriate for our purposes. Although future experimentation (out of the scope of this thesis) could be beneficial, as the literature available are largely for land surface applications.

Convolutional neural networks were originally designed to classify images, but have since seen success in much broader domains. These algorithms combine two specialised types of hidden layers, convolution layers and pooling layers. Convolution layers detect small patterns in the data, whereas pooling layers downsample these to select the dominant subset. A convolution layer is comprised of convolution "filters", these are small matrices with different arrangements of 0's and 1's that are used to convolve the input data. The convolved data are essentially scaled such that regions where the input data are structured similarly to filter see large values and other regions see small values. I.e., regions similar to the filter are highlighted.

For example, consider the 1-dimensional temporal convolution filter, for two time steps of 4 input features:

$$\text{Input data} = \begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \end{bmatrix}$$

$$\text{Convolution filter} = \begin{bmatrix} \alpha & \beta \\ \gamma & \sigma \end{bmatrix}$$

$$\text{Convolved data} = \begin{bmatrix} a_1\alpha + b_1\beta + a_2\gamma + b_b\sigma & b_1\alpha + c_1\beta + b_2\gamma + c_2\sigma & c_1\alpha + d_1\beta + c_2\gamma + d_2\sigma \end{bmatrix}$$

The pooling layer is intended to condense the data, and various methods can be used to achieve this. A common means is called max pooling, whereby the maximum value of blocks within the matrix is taken. In this method, any large value resulting from the convolution filter will be registered by the algorithm.

The filters are learned in the training process and the number of filters is tuned by the developer. A fully connected layer typically follows the convolution and pooling layers and is used to receive the results of the convolutional process.

3.2 Model interpretability

Model interpretation is the process of understanding what predictor variables (features) were most influential in the prediction of a response variable. Some models are interpretable by design. For instance, the importance of each feature in a linear model is simply its weight (if data are normalised). However, other more complex models, such as neural networks, require different tools to determine how predictions are made. Over recent decades, a number of interpretation methods have been developed to this end. The literature organises these different methodologies by two key characteristics, "model-agnostic" and "local" or "global". Model-agnostic interpretation methods are those which can be applied to any machine learning model, an important property for model comparison (Ribeiro et al., 2016). Local methods describe how features affect an individual prediction, while global methods describe how features affect a set of predictions on average.

Recently, Lundberg and Lee (2017) proposed Shapley Additive Explanations (SHAP), a local model-agnostic interpretation method that is built on a strong game theoretical foundation and has unified many existing model interpretibility methods. SHAP likens a model and its features to an instance of game theory. Here, the model is the "game", the features are "players" and the resulting prediction is the total "payout".

Game theory is a branch of applied mathematical modelling that presents a theoretical framework for quantifying the behaviour of interacting agents. Game theory was first put forward by John von Neumann in his 1928 proof and developed by von Neumann and Morgenstern in the book *Theory of Games and Economic Behaviour* (1953). The framework constructed by game theory reaches applications far beyond that of players engaged in a game of chance, nor are they limited to the field of their origin, economics. Since its inception, mathematicians have

proposed various means of distributing a games payout between its players, one such solution is the Shapley value.

The Shapley value was first proposed by Lloyd Shapley in Shapley, 1953. Shapley argued that the "fair" distribution of the total payout was that which saw players receiving payouts proportional to their contribution to the outcome of the game. For example, if player A was solely responsible for a game's total payout of \$100, while players B and C made no contributions, then player A would receive the full \$100 and players B and C \$0. Shapley derived a value to distribute this fair payout, showing that it must be equal to the marginal contribution of a player to the total payout, averaged over all possible coalitions of players.

3.2.1 Shapley value proof

A derivation and proof of the Shapley value was first given in *Contributions to the Theory of Games II: A value for n person games* (Shapley, 1953). Here, we present an analogous proof of the Shapley value, which leans on methodology developed in the proof by (Osborne and Rubinstein, 1994).

First, let us define a game, following coalitional game theory, as the supperadditive set-function $v : \mathbb{R}^{2^{|N|}-1} \mapsto \mathbb{R}$ where N is the finite set of $|N|$ players.

As discussed previously, for some coalition of players $S \subseteq N$, the Shapley value defines the *fair* distribution of a total game payout $v(S)$ as the marginal contribution of the i^{th} player. For a given coalition S, $i \notin S$, this is given by

$$\Delta_i(S) = v(S \cup \{i\}) - v(S). \tag{5}$$

Thus, averaged over all sets of possible coalitions S, the i^{th} Shapley value, ϕ_i, is defined in Definition (1). The value, by definition, provides the *efficient* payout distribution

$$v(N) = \sum_{i=1}^{|N|} \phi_i(v). \tag{6}$$

Definition 1. *Let $\tilde{R}$ be the set of all $|N|!$ orderings of the players N, and let $S_i(R_i)$ be the set of players preceding player i in ordering R, then*

$$\phi_i(N, v) = \frac{1}{|N|!} \sum_{R \in \tilde{R}} \Delta_i(S_i(R)) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Delta_i(S).$$

Shapley proposed that this value defined a fair distribution of a total game payout as it satisfied what he termed *"the axioms of fairness"*. These three axioms described conditions where, firstly, if the contributions of two players are interchangeable, they would receive the same payout. Secondly, if a player did not contribute, they would receive no payout. And, thirdly, if a game is comprised by two subgames, the payout to each player from the game is given by the sum of the payouts they received from their contributions to each subgame. Mathematically, we can construct these rules of fairness with the help of two definitions:

Definition 2. *The player i is a dummy (or carrier) player, iff*

$$v(S \cup \{i\}) = v(\{i\}) + v(S) \iff \Delta_i(S) = v(\{i\});$$

Definition 3. *The players i, j are interchangeable, iff*

$$\forall S \subseteq N, i, j \notin S, v(S \cup \{i\}) = v(S \cup \{j\}) \iff \Delta_i(S) = \Delta_j(S).$$

With the aide of Definitions (2) and (3), we can present the axioms described previously, called Symmetry, Dummy (Carrier) and Additivity, respectively.

Axiom 1. *Symmetry If i, j, are interchangeable, then $\phi_i(v) = \phi_j(v)$.*

Axiom 2. *Dummy (Carrier) If i is a dummy, then $\phi_i(v) = v(\{i\})$.*

Axiom 3. *Additivity $\forall i \in N, \forall S \subseteq N$, if the game $< N, v + w >$ can be represented as $(v + w)(S) = v(S) + w(S)$, then $\phi_i(v + w) = \phi_i(v) + \phi_i(w)$.*

Shapley's theory determined that together these axioms alone would uniquely determine the Shapley value as that defined by Definition (1). Moving forward with the constructions developed above we are now ready to prove Shapley's theorem.

[Shapley] For any N a unique value function, ϕ, exists that satisfies Axiom (1), (2) and (3); it is given by (1).

Proof. First, let us show that ϕ satisfies each axiom, beginning with Axiom (1). Let i, j be interchangeable, then, by Definition (3),

$$v(S \cup \{i\}) = v(S \cup \{j\}), i, j \notin S.$$

Let the ordered coalition of players R be equivalent to that of R' excepting that i and j are interchanged. Then in the case where i precedes j in R, neither i nor j are included in the coalition $S_i(R)$, the set of players preceding player i in ordering R. Thus

$$\Delta_i(S_i(R)) = \Delta_j(S_j(R')).$$

Alternatively, in the case where j precedes i in R, the player j is included in the coalition $S_i(R)$. Thus using Equation (5) where $S = S_i(R) \backslash \{j\}$ we have

$$\begin{aligned} \Delta_i(S_i(R)) - \Delta_j(S_j(R')) &= v(S \cup \{i\}) - v(S) - (v(S \cup \{j\}) - v(S)) \\ &= v(S \cup \{i\}) - v(S \cup \{j\}) \end{aligned}$$

But since i, j are interchangeable

$$\Delta_i(S_i(R)) = \Delta_j(S_j(R')).$$

Then, using the above and Definition (1), we can consider the difference

$$\begin{aligned} \phi_i - \phi_j &= \frac{1}{|N|!} \sum_{R \in \tilde{R}} \Delta_i(S_i(R)) - \frac{1}{|N|!} \sum_{R \in \tilde{R}} \Delta_j(S_j(R')) \\ &= \frac{1}{|N|!} \sum_{R \in \tilde{R}} (\Delta_i(S_i(R)) - \Delta_j(S_j(R'))) \\ &= 0. \end{aligned}$$

Thus Axiom (1) holds. Next, let us see that Axiom (2) holds also. Let i be a dummy player in the game v, then, by Definition (2), $\Delta_i(S) = v(\{i\})$. But Definition (1) is given by

$$\begin{aligned} \phi_i &= \frac{1}{|N|!} \sum_{R \in \tilde{R}} \Delta_i(S_i(R)) \\ &= \frac{1}{|N|!} \sum_{R \in \tilde{R}} v(\{i\}) \\ &= v(\{i\}). \end{aligned}$$

Thus Axiom (2) holds. Now, we can show that Axiom (3) holds also. Let $u = v + w$ be a game such that, for arbitrary $S \subseteq N, u(S) = v(S) + w(S)$. Equation (5) implies that

$$u(S \cup \{i\}) - u(S) = v(S \cup \{i\}) - v(S) + w(S \cup \{i\}) - w(S).$$

Then

$$\begin{aligned}
\phi_i(N, u) &= \frac{1}{|N|!} \sum_{R \in \tilde{R}} \Delta_i(S_i(R)) \\
&= \frac{1}{|N|!} \sum_{R \in \tilde{R}} (v(S \cup \{i\}) - v(S) + w(S \cup \{i\}) - w(S)) \\
&= \frac{1}{|N|!} \sum_{R \in \tilde{R}} (v(S \cup \{i\}) - v(S)) + \sum_{R \in \tilde{R}} (w(S \cup \{i\}) - w(S)) \\
&= \phi_i(N, v) + \phi_i(N, w).
\end{aligned}$$

Thus Axiom (3) holds. Finally, all that remains to show is that ϕ constitutes a unique solution. From Definition (1) we observe that $\phi_i(N, v)$ is a function that evaluates to a scalar value for each player i. This means that the value ϕ_i is unique if there exists an algebraic basis from which it can be determined. A basis must be a linearly independent subset of the space C that spans C, where, for our purposes, $C \supseteq N$. Recall that the game $< N, v >$ is a collection of $2^{|N|} - 1$ numbers, then let

$$v = \sum_{T \subseteq C} \alpha_T v_T \tag{7}$$

be a potential basis. Since $(v_T)_{T \subseteq C}$ is a collection of $2^{|N|} - 1$ games, it spans v. Thus (7) is a basis iff $(v_T)_{T \subseteq C}$ are linearly independent. We can show that this is true by assuming the contrary. Let $(v_T)_{T \subseteq C}$ be linearly dependant, then

$$\exists \alpha_i \in \mathbb{R} \text{ such that } \sum_{i=1}^{2^{|N|}-1} \alpha_i v_{T_i} = 0 \text{ for some } \alpha \neq 0. \tag{8}$$

Intuitively, we can see that in the set of possible coalitions T there must be at least one coalition, T_1, with the minimum number of players for which we can set $\alpha_1 \neq 0$. Then (8) becomes

$$\begin{aligned}
\sum_{i=1}^{2^{|N|}-1} \alpha_i v_{T_i} = \alpha_1 v_{T_1} + \sum_{i>1}^{2^{|N|}-1} \alpha_i v_{T_i} = 0 \\
\implies v_{T_1} = \frac{1}{\alpha_1} \sum_{i>1}^{2^{|N|}-1} \alpha_i v_{T_i}.
\end{aligned}$$

But since, for the game with total payout equal to one (the unitary game) which is defined on the universe of players T, we have

$$v_S(T) = \begin{cases} 1 & \text{if } S \subseteq T, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, by counterexample,

$$\begin{aligned}
v_{T_1}(T_1) = \frac{1}{\alpha_1} \sum_{i>1}^{2^{|N|}-1} \alpha_i v_{T_i} T_1 \\
\implies 1 = 0,
\end{aligned}$$

we reach a contradiction. So $(v_T)_{T\subseteq C}$ is linearly independent, and v is a basis that, for some $\beta_j \in \mathbb{R}$, can uniquely define $\phi_j = \sum_{N\subseteq C} \beta_N v_N$.

Thus, the proof is concluded. □

In practice, computing the Shapley value is expensive. In fact, it becomes exponentially more expensive to compute as the number of players increases. This is prohibitive for many real-world applications in deep learning, as a large number of features (players) are often necessary for accurate predictions. In response to this limitation, various methods have been proposed to estimate the Shapley value (e.g., Datta et al., 2016; Owen and Prieur, 2016; Štrumbelj and Kononenko, 2014; Sundararajan et al., 2017).

In their 2017 paper, Lundberg and Lee (2017) showed that many of these estimation approaches could be generalised. Their work also generalised other local model-agnostic interpretability schemes, which were previously unrelated to Shapley game theory. To achieve this, they summarised the Shapley values using a linear model, which is a linear combination of all the Shapley values associated with a single game. This is called the "additive explanation model" and is defined by

Definition 4. *Let $z' = \{0,1\}^M$ be the coalition vector, indicating the presence of each feature in a coalition with maximum size M, then*

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i.$$

They required that the Shapley values defined by this linear model (henceforth referred to as "SHAP" values) satisfy three key properties: local accuracy, missingness and consistency. If the SHAP values satisfy these properties, then it can be shown that they also satisfy uniqueness and Shapley's three axioms of fairness. Meaning that they are indeed Shapley values (Lundberg and Lee, 2017). To see this, we introduce properties 1 - 3. First, Property 1 requires that the explanation model is accurate for locally simplified feature values x',

Property 1. *Local accuracy Let $\hat{f}(x)$ be the prediction for data x, and $h(x')$ be the mapping of simplified data x' to orignal data x, then*

$$\hat{f}(x) = g(x'), x = h(x').$$

Next, Property 2 requires that features missing from the model in all subsets of features (coalitions) receive no payout. This property is exactly Axiom (2) where the lone contribution of the feature (player) in question is null, $v(i) = 0$:

Property 2. *Missingness*

$$x'_i = 0 \implies \phi_i = 0.$$

Lastly, Property 3 requires that an increase in the marginal contribution of a feature, i, corresponds to an increase in the payout to that feature, the SHAP value ϕ_i:

Property 3. *Consistency Let $\hat{f}_x(z') = \hat{f}(h_x(z'))$ and $z'_{\setminus j} \implies z'_j = 0$, further let $\hat{f}'$ correspond to a distinct model relative to $\hat{f}$, then*

$$\hat{f'}_x(z') - \hat{f'}_x(z'_{\setminus j}) >= \hat{f}_x(z') - \hat{f}_x(z'_{\setminus j}) \implies \phi_i(x, \hat{f'}) >= \phi_i(x, \hat{f}).$$

Computation of the SHAP values can be performed using a range of either model-agnostic or model-specific approaches. In their work, Lundberg and Lee (2017) proposed several such

methods of computing the SHAP values. For a model-agnostic method, where feature independence can be assumed, Lundberg and Lee (2017) proposed a KernelSHAP method which saw improved efficiency when compared to the existing Shapley sampling approach by Young (1985).

However, limitations persist in the SHAP framework which have become a topic of current research. Sampling from the marginal distribution, as in KernelSHAP and Ribeiro et al. (2016) and Štrumbelj and Kononenko (2014), ignores feature interdependence when approximating the conditional expectation. Meaning that the independence assumption is an important assurance that the SHAP values are being interpreted appropriately. In cases where this assumption may not hold, this leads to reduced trustworthiness. In such cases, the SHAP values may misrepresent feature importance by assigning a high payout to a highly correlated feature which, due to its high correlation to another feature, may not have contributed as much as suggested to the prediction.

4 Methodology

The main approach of the study, based on the literature, is the following:

1. Classify "true" cloud vertical structure using merged CPR-CALIOP data
2. Collocate merged CPR-CALIOP data with AHI data for all available 2019 data (6 months)
3. Develop and optimise two neural network models, a feed-forward neural network and a 1-D convolutional neural network (convolved in the temporal dimension)
4. Use KernelSHAP Shapley value estimation software for model interpretation and assessment

4.1 Defining the "truth"

Any supervised deep or machine learning algorithm is only as good as the "truth" data it is given, which are used to derive the functional model. An algorithm trained on poor quality truth labels may achieve high accuracy yet be good at predicting an unreliable quantity. In our case, the active radar-lidar (CPR-CALIOP) data provides the most accurate atmospheric structural measurements with sufficient data quantity. As such, we classify cloud vertical structure using the merged CPR-CALIOP cloud layer classification product.

We define a multi-layer cloud occurrence as when two cloud layers are detected that are at least 100 m apart (vertically), with an allowance of 320 m to account for the coarsest resolution of the product. This approach is consistent with existing literature (e.g., Li et al., 2011; Li et al., 2022; Marchant et al., 2020; Oreopoulos et al., 2017; Wang et al., 2016). All cloud profiles are classified as one of three classes, no cloud, single-layer or multi-layer. The process for designating the classifications is detailed in Algorithm 1.

A total of approximately 5,500,000 good quality cloud profile observations were classified in this way between January and June 2019 for latitudes higher than 50° South, with approximately 1,336,000 defined as no cloud, 2,910,000 single-layer and 1,257,000 multi-layer. Two examples are shown in Figure 6, which show that the algorithm is behaving as expected for both low-mid layer cloud co-occurrences and low-upper layer cloud co-occurrences.

4.2 Collocation of data

To train the neural network models accurately requires appropriate temporal and spatial collocation of the training (AHI) and "truth" (CPR-CALIOP) data. While the CPR-CALIOP truth data are recorded from a nadir-pointing viewing angle at all locations, the AHI's geostationary orbit results in different viewing angles for each pixel observed, depending on the distance from the satellite. To ensure the two sets of observations are indeed observing the same cloud layer, we need to account for the different viewing angles by performing a parallax correction.

Level 1B AHI data are read in Himawari Standard Data (HSD) format using the Satpy Python package (Raspaud et al., 2022) with default calibration. Higher resolution channels (channels 1-4) are downsampled to 2 km resolution using software published by Robbins and Proud (2022) and the mean and the standard deviation are used as additional inputs to the neural networks. To correct for the parallax when collocating these data with the merged CPR-CALIOP data, we use results from Robbins et al. (2021) who parallax-corrected AHI data with approximately 7300 CALIOP overpasses for 2019. The merged CPR-CALIOP data are then collocated with the parallax-corrected AHI-CALIOP results as outlined in Figure 7. Where

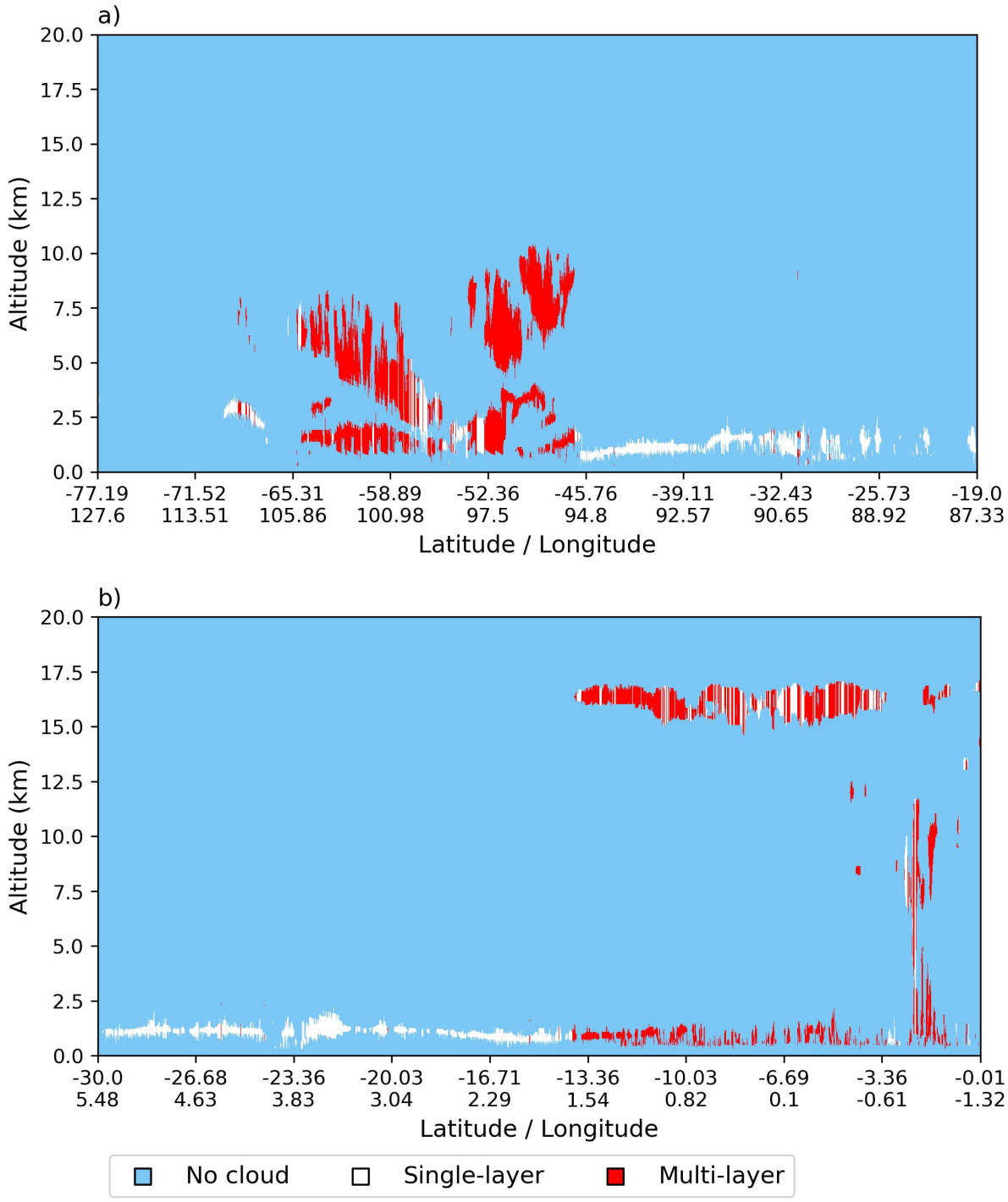

Figure 6: Examples of the cloud layer classification results for the merged CPR-CALIOP (radar-lidar) product for (a) 2019-02-01 07:25 and (b) 2019-03-26 13:04. Shows successful discrimination between cloud profile types.

Algorithm 1 Classify cloud vertical structure

```
Read merged CPR-CALIOP cloud layer product
Set d = 0.4                                   ▷ Minimum vertical gap between layers
Set x_i = 0 ∀i
for each i (cloud profile) do
    if no good quality cloud layers then
        x_i ← −9
    else if at least one good quality cloud layer then
        x_i ← 1
        if more than one good quality cloud layer then
            for each j,k consecutive layers do
                if layer altitudes have a quality issue then
                    x_i ← −9
                else if dist(top_j, bottom_k) > d then
                    x_i ← 2
                    break
                end if
            end for
        end if
    end if
end for
```

multiple cloud layers have undergone parallax correction for the AHI-CALIOP data, the latitude and longitude of the top layer are taken and used for collocation with the CPR-CALIOP data. This means that while the top cloud layer has been corrected for in our study, lower layers may still suffer from some parallax effect, particularly where the AHI viewing angle is large (at high latitudes).

We require each CPR-CALIOP profile to be within the AHI scene ten minute time window, with an error allowance of 5 minute (Robbins et al., 2022). If the profile is also within 2km (coarsest horizontal resolution of the merged product) of the parallax correct CALIOP observation, the profile is collocated with the corresponding AHI pixel. This ensures the chosen profile will also satisfy the parallax correction of Robbins et al. (2022) for upper level cloud.

In total, approximately 312,000 AHI pixels were collocated with merged CPR-CALIOP data between January and June 2019 over the SO. Of these, 297,000 were daylight observations, and 15,000 were during the night/twilight period and were excluded from further analyses. Observations span the full latitude of the SO, taken to be from 50° South, and are contained to the AHI field of view, concentrated south of Australia between 65° and 220° East (Figure 8). Night/twilight observations were excluded as CloudSat's CPR has been operating in daytime only mode since 2011. More observations were classed as single-layer than any other class, due to the cloud climatology of the SO region.

The main cause of an unsuccessful collocation was a lack of CPR-CALIOP observation within the required distance of the parallax corrected CALIOP data. If this requirement was satisfied, only quality issues would cause the collocation to terminate. I.e., if a CPR-CALIOP observation was found within the required distance, the observation was also within the necessary time window. The CPR-CALIOP data quality was best at the beginning of the period, with significantly fewer good quality observations after May. This is due to the reduced daylight hours of the austral autumn and winter period, particularly given the high latitude domain, explaining the temporal bias observed in Figure 8.

The midday period (between 12:00 - 13:00 H, local time) was used to assess correlations

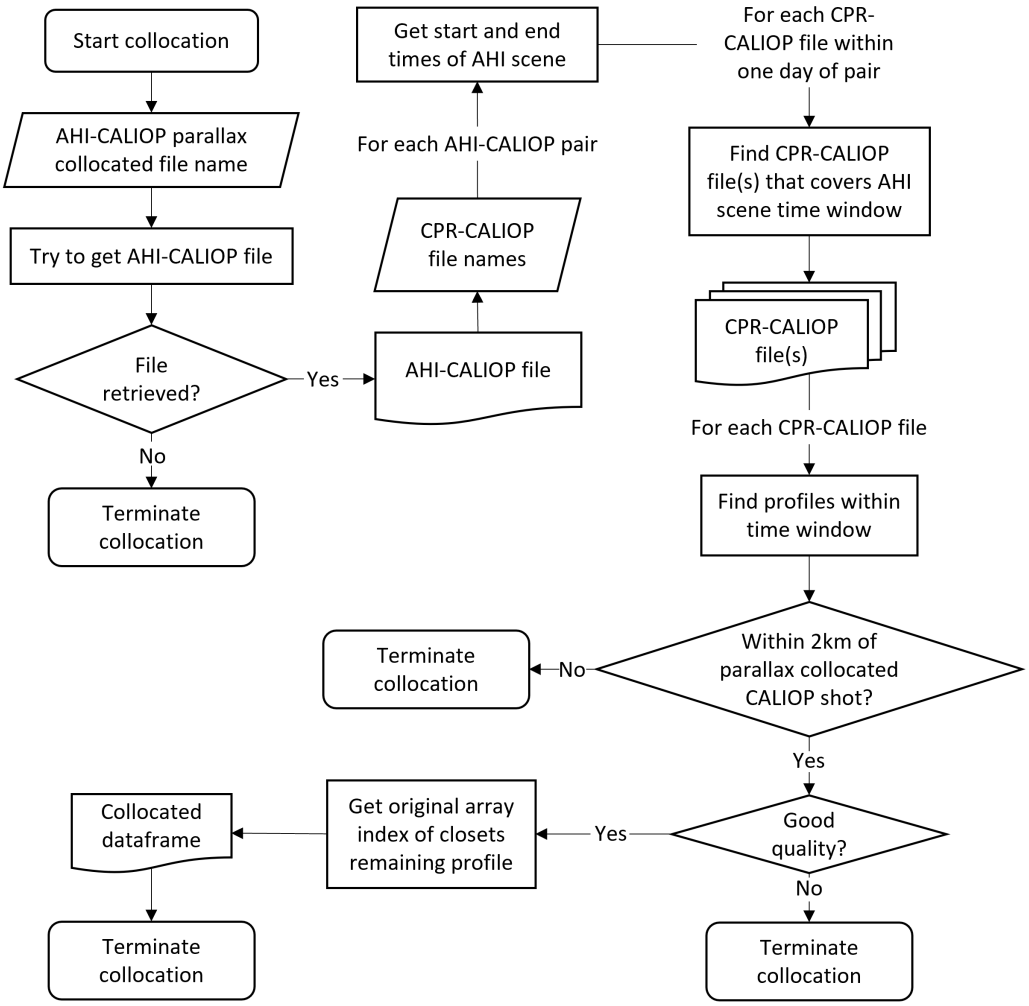

Figure 7: Collocation procedure for the CPR-CALIOP "truth" data and the AHI pixels to be used in the neural network models.

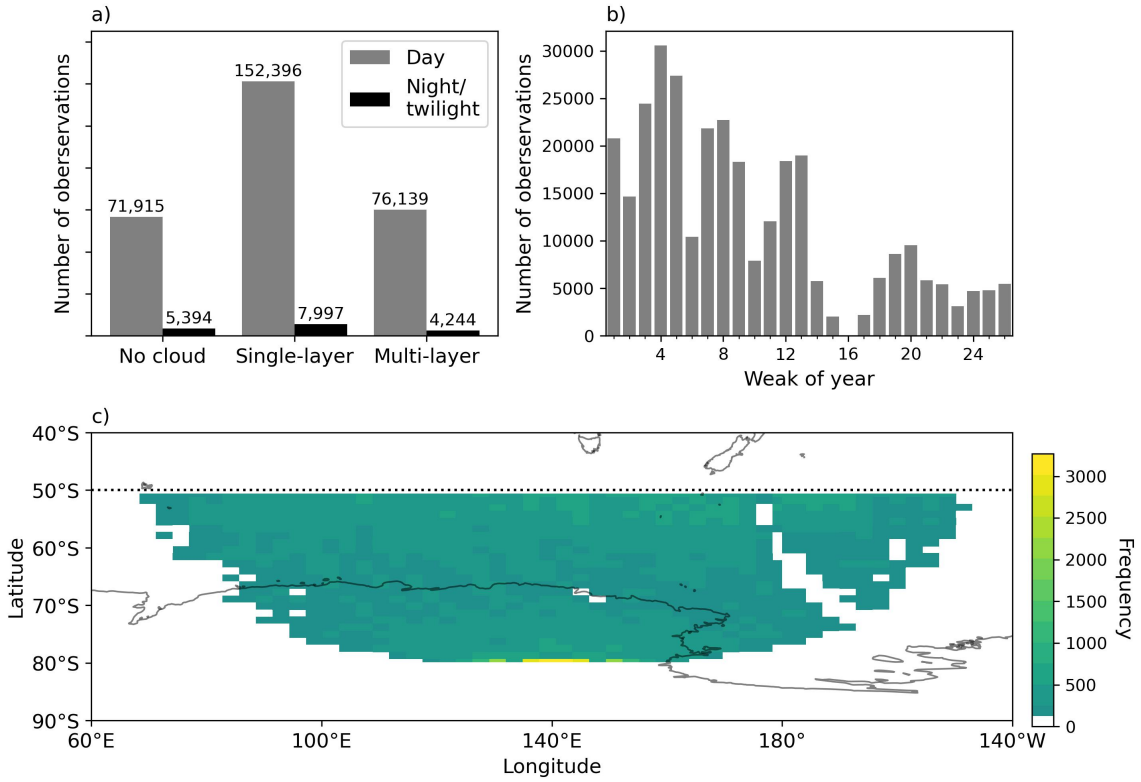

Figure 8: Distribution of collocated data available for training and validation of the neural networks.

between observation bands to remove the temporal trend associated with the passage of sunlight. Correlations between the AHI observation bands during this time varied with band wavelength (Figure 9). The mean values of the visible bands were well (positively) correlated, as were high infrared bands. Weak negative correlations were observed between the visible and infrared bands. Band 12 was an exclusion to this pattern, being positively correlated to the visible bands and weakly correlated to the infrared bands. This may result from the high latitude of the region considered, as at this large viewing angle (the "limb") Rayleigh aerosol scattering in the visible bands and ozone absorption in band 12 both cause limb brightening.

High correlation between features can introduce multicollinearity issues to many machine and deep learning algorithms. Typically, multicollinearity between variables is understood to arise where the Variance Inflation Factor (VIF) (Equation 9) is greater than five. Our data observe VIF greater than five for 24 out of the 170 variable pairs (Figure 10). However, due to the high complexity of deep neural networks, multicollinearity issues are not a concern. Despite this, multicollinearity can cause reliability issues in model interpretation using SHAP (see Section 3.2), hence our assessment.

$$VIF = \frac{1}{1 - R^2}. \tag{9}$$

For the convolutional neural network, 92.4 % of the original approximately 312,000 collocated samples used for the feed forward algorithm were successfully paired with good quality data from the previous AHI observation (Figure 11. Of those unsuccessfully paired the majority were in January, with smaller losses seen in the other months.

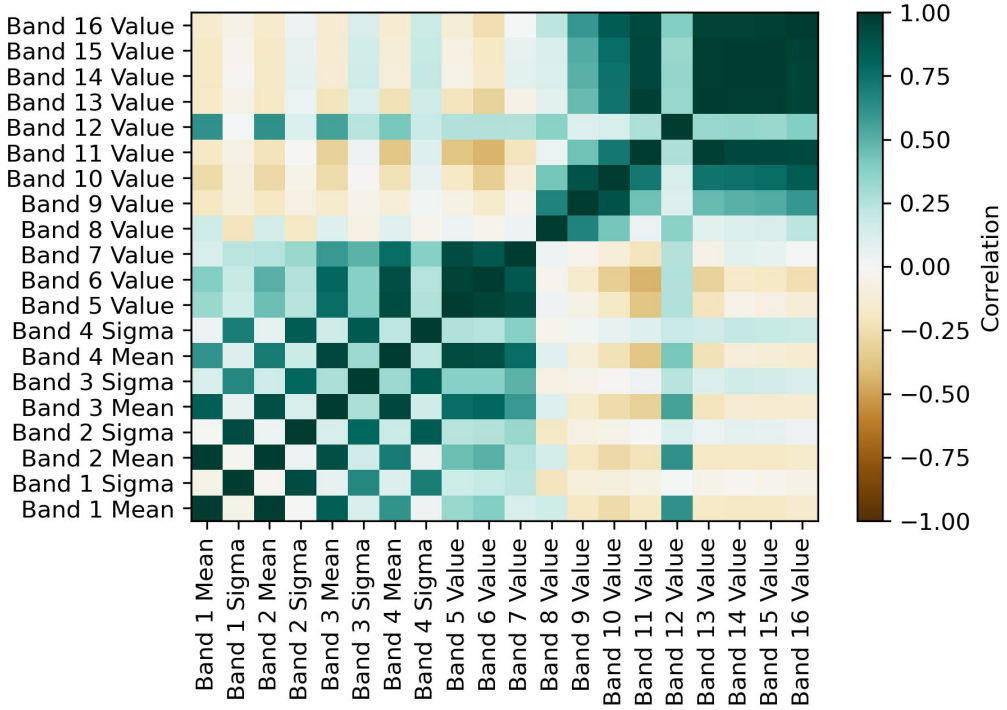

Figure 9: Pearson correlation between AHI observation bands between 12:00 - 13:00 H (local time).

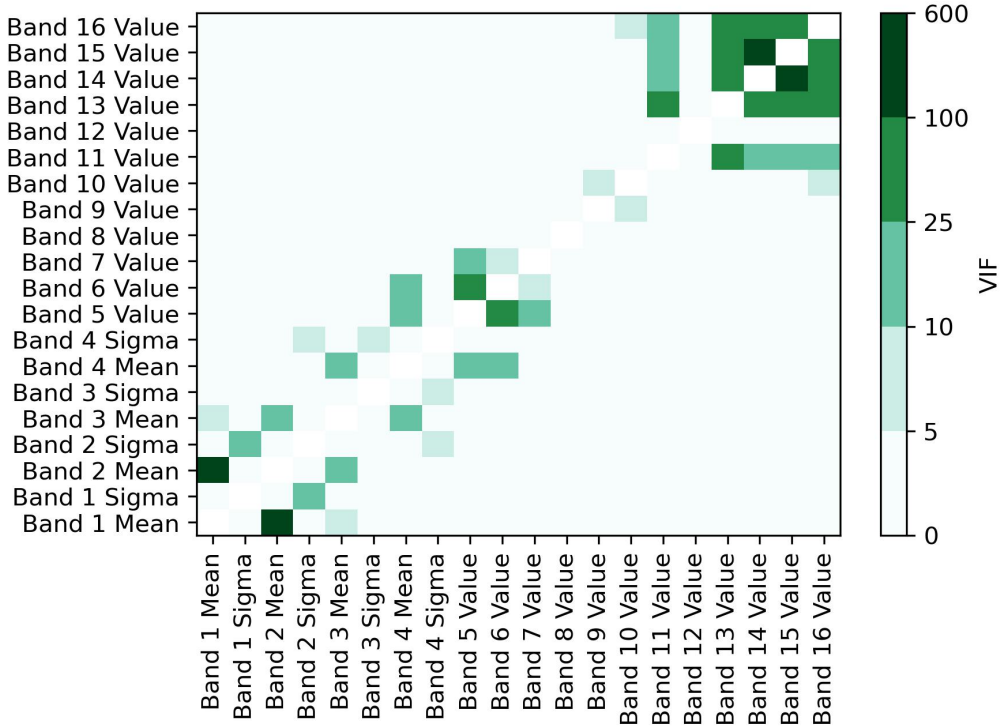

Figure 10: VIF scores for AHI observations bands during between 12:00 - 13:00 H (local time).

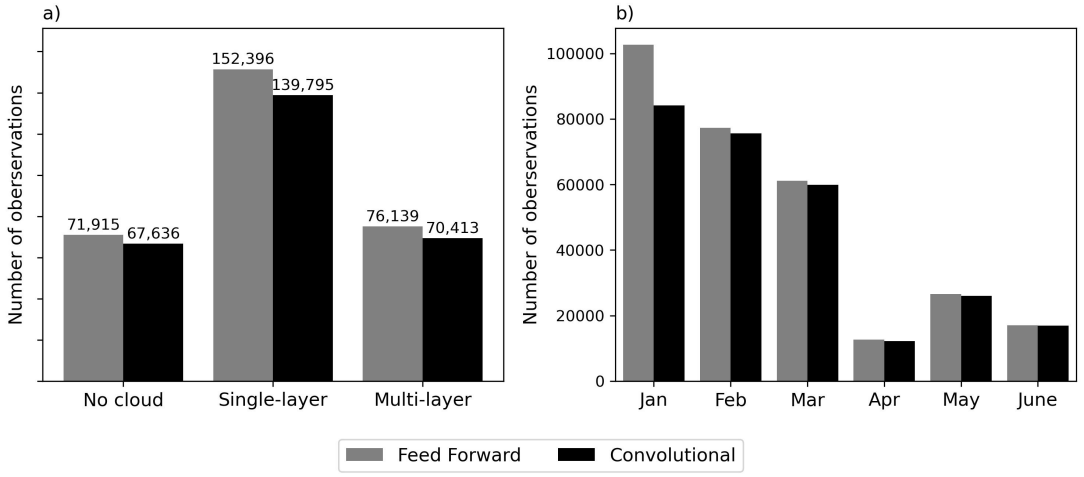

Figure 11: Proportional of original collocated dataset available for the convolutional neural network.

4.3 Training and validation of neural networks

Feed forward and convolutional neural networks were implemented using the TensorFlow (version 2) package for Python (Abadi et al., 2016). Day time models were created for each algorithm following the approach of existing literature, where time of day was determined by solar zenith angle (SZA) with $0° <= SZA < 85°$ for day and $SZA >= 85°$ for night/twilight. The 27 inputs to the neural networks are the following:

- channels 1-16 (mean and variance values for downscaled channels 1-4);
- Himawari observation angles; and
- pixel latitudes and longitudes and SZA;

The input data are normalised between approximately -1 and 1 to ensure optimal training of the neural networks by improving minimisation of the loss function.

The target data are the three classification classes defined in Section 4.1, no cloud, single-layer and multi-layer. Multi-class classification algorithms require a softmax activation function (Equation 4) in the output layer, which gives a probability value for the occurrence of each class. The maximum of these in any prediction is taken as the final class predicted by the model. ReLu activation is used for the hidden layers, and a dropout layer is included with dropout rate of 0.2 to prevent overfitting (Srivastava et al., 2014).

A random sample of 20 % of the collocated data was selected for validation purposes and was unseen in by the model in the training and optimisation processes. An out-of-sample set was not chosen, as data at both the beginning and end of the time period saw an uneven spread in the classification labels, meaning only in-sample periods would be suitable for robust validation.

The convolutional neural network was convolved in the temporal dimension using AHI data from one preceding time step. One convolution layer was used, with a pooling layer and fully connected layer following. A subset of the same validation set was held out for the convolutional algorithm. Meaning that although this network was trained and tested on fewer observations than the feed forward, the seen and unseen data sets are identical.

The neural networks are optimised in a two-stage process. During optimisation, 5-fold Cross-Validation (CV) is employed to reduce the risk of overfitting. 5-fold CV splits the training data

into two 80:20 train:test splits five times, such that each CV test set is unique, five models are then trained and tested for each CV data split and mean test accuracy and PoD scores are obtained. Here, accuracy is the rate at which predictions equal the truth data and PoD is the proportion of multi-layer observations that were predicted.

First, the number of epochs are optimised for two simple network structures, a 1-layer network with 10 neurons per layer and a 10-layer network with 50 neurons per layer for the Adam optimiser (Kingma and Ba, 2014). Epochs of sizes 25, 50, 100, 150 and 200 are considered. Only incremental improvements in accuracy were observed for more than 100-150 epochs in each model. Hence, we select an epoch size of 150 for the final models. For the convolutional network, the number of filters in the convolution layer is also optimised for in this stage. Filters of number 16, 32, 64 and 128 are considered for the two simple network structures. While some increases in accuracy were observed, no significant improvements in accuracy were observed for very large filter numbers, hence 64 filters are used for the final model.

The second stage of optimisation performs a parameter search using 150 epochs and 64 filters (for the convolutional network). A grid of 1, 2, 5, 10 and 20 layer structures with 10, 20, 50, 100 and 200 neurons is considered for SGD and Adam optimisers.

The simplest networks with the highest CV accuracy and multi-layer PoD were selected as the final algorithms for both the feed forward and convolutional models. For the feed forward neural network this is found to be 5 layers of 100 neurons when using the SGD optimiser (Figure 12). The SGD optimiser saw improved performance compared to the Adam optimiser in PoD, observing roughly a 5% increase in PoD. Greater accuracy was observed for a greater number of layers, however the PoD scores significantly dropped for more than 5 layers, suggesting some overfitting may arise in the multi-layer class at this level of complexity. Therefore, the final structure of the feed forward network consists of

- an input layer of N neurons, where N is the number of inputs used, with ReLu activation;
- a dropout layer with dropout rate of 0.2;
- 5 layers of 100 neurons each with ReLu activation for each layer; and
- a three neuron output layer with softmax activation.

Whereas for the convolutional neural network this is found to be 5 layers of 200 neurons when using the SGD optimiser (Figure 13). Therefore, the final structure of the convolutional network consists of

- an input layer of N neurons, where N is the number of inputs used, with ReLu activation;
- a convolution layer with 64 filters;
- a pooling layer with pooling size 2;
- a dropout layer with dropout rate of 0.2;
- a flattening layer to allow for feed forward operations;
- a fully connected layer with ReLu activation;
- 5 layers of 200 neurons each with ReLu activation for each layer; and
- a three neuron output layer with softmax activation.

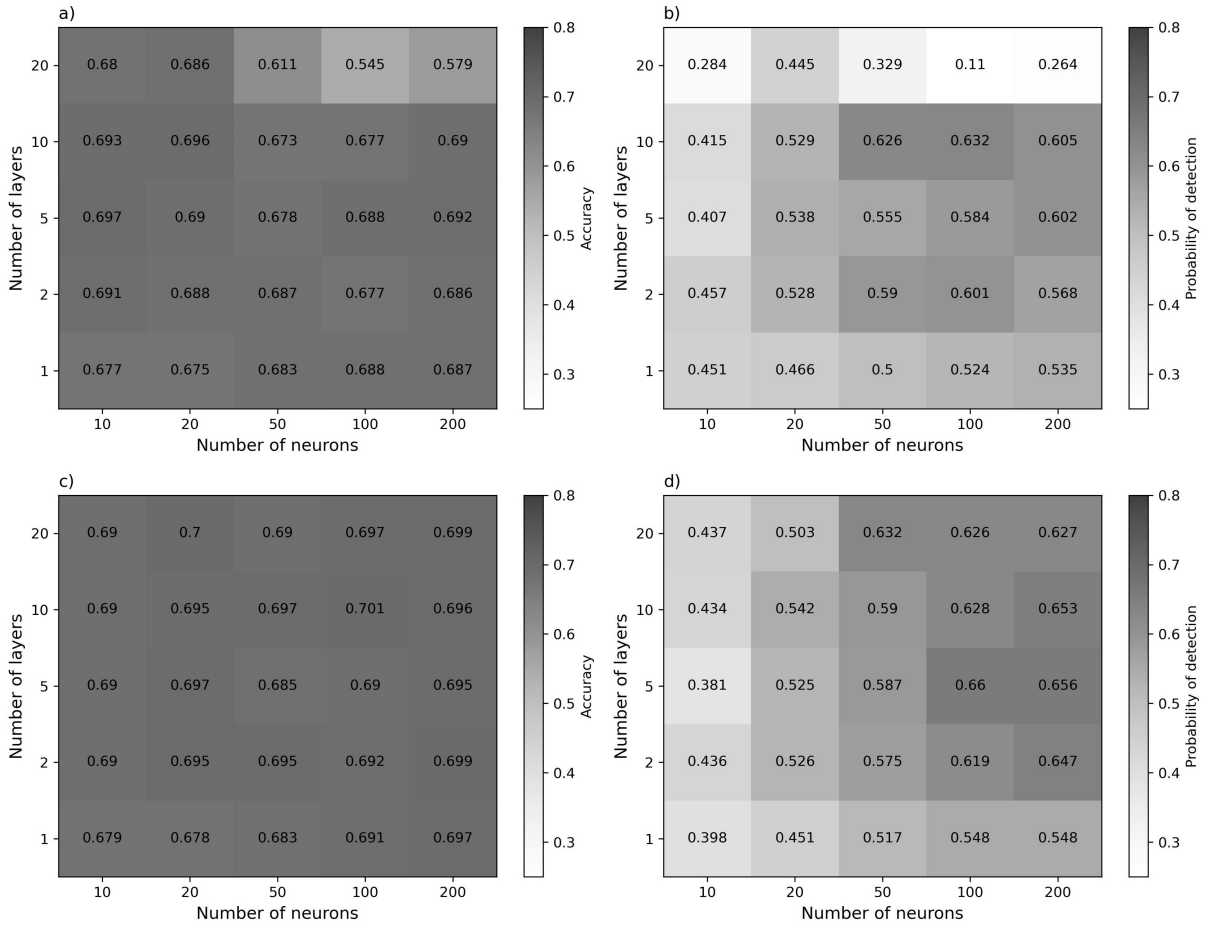

Figure 12: Feed forward accuracy (left) and Probability of Detection (PoD) (right) results for the hyperparameter grid search for number of layers and neurons for the Adam (top) and Stochastic Gradient Descent (SGD) (bottom) optimisers.

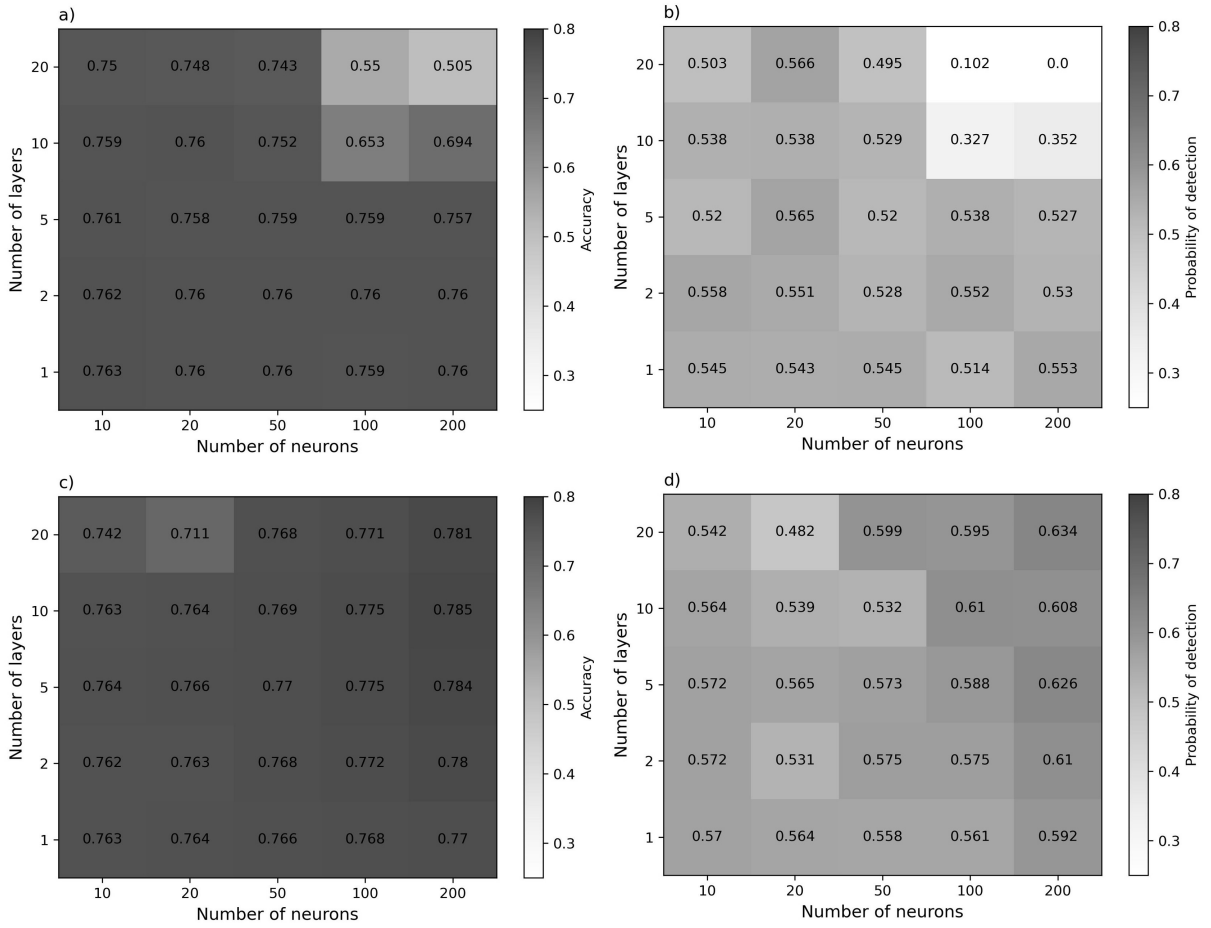

Figure 13: Convolutional accuracy (left) and Probability of Detection (PoD) (right) results for the hyperparameter grid search for number of layers and neurons for the Adam (top) and Stochastic Gradient Descent (SGD) (bottom) optimisers.

5 Model results

5.1 Statistical comparison of model results

The two models are statistically compared using the in-sample validation set (which is unseen by the models during training and cross-validation). The validation set accuracy of the feed forward and convolutional networks were 70.4 % and 79.5 % respectively. The convolutional network observing nearly a 10 % increase in accuracy compared to the feed forward. This improvement is not as well represented in the PoD of multi-layer clouds, which was 58 % and 61 % respectively.

Further assessments are undertaken using the True Positive Rate (TPR) (Equation 10) and False Positive Rate (FPR) for each model (Equation 11). These are calculated for each class (no cloud, single-layer and multi-layer) using the probability value which is output by each network (between 0 and 1), which corresponds to the probability that the given observation belongs to each category. The maximum of these is used to determine the predicted class. Here, rather than selecting the maximum, we compute the receiver operating characteristic (ROC) curve, which presents the TPR vs FPR for series of different probability thresholds. ROC curves are given for the multi-layer classification and overall classification and surface type, ocean and land. Greater area under the ROC curve (AUC) indicates better performance, as the TPR is increased and the FPR reduced.

$$\text{True positive rate} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{10}$$

$$\text{False positive rate} = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}} \tag{11}$$

In the validation set, roughly 2/3 of the data were over the ocean and 1/3 over land. Over ocean, 32 % of observations were multi-layer, while this was only 9 % over land (Table 3), meaning significantly fewer observations were available for statistical analyses of the multi-layer classification over land than over the ocean. Similarly, fewer no cloud observations occurred over ocean, although this is a smaller reduction than seen for multi-layer clouds as more data were available over the ocean than over land.

Table 3: Percentage of observations of each classification type over land and ocean in the validation set.

	No cloud	Single-layer	Multi-layer
Ocean	9.4	58.3	32.2
Land	57.9	33.0	9.2

According to the ROC curves, the convolutional network outperforms the feed forward in both the overall and multi-layer classification results. In fact, the convolutional network performs as well for multi-layer classification as the feed forward network does overall. Whether the increased performance is primarily due to the additional data or the additional complexity associated with the convolution layers is unclear. This could be further drawn out through the addition of more previous time steps of data, which would give indication to the level of improved accuracy associated with the data for networks with identical architectures. Little difference is apparent in the AUC between the land and ocean, suggesting a comparable balance between false positives and negatives over each surface type as the classification probability threshold is modified.

Confusion matrices are also presented, showing the classification results for each model and each of the three classes (no cloud, single-layer and multi-layer) (Figure 15). These are normalised by the true label (by row), meaning the diagonal entries correspond to the PoD of each

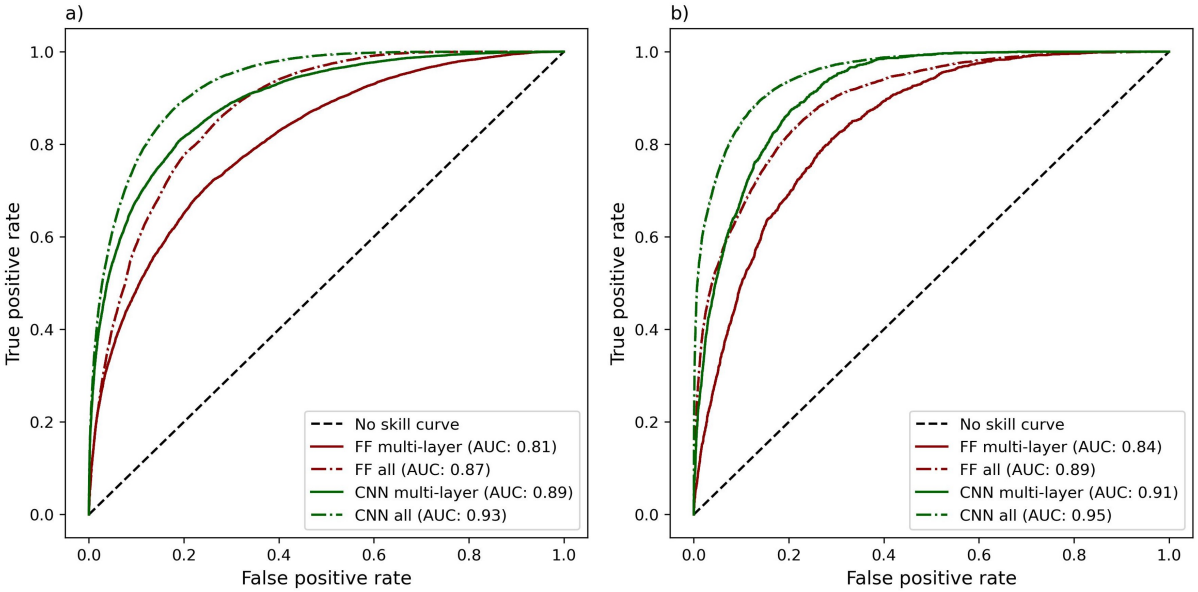

Figure 14: ROC curve for the final feed forward and convolutional neural networks over (a) ocean and (b) land. Results for both the overall TPR and FPR and multi-layer class TPR and FPR are given.

class. Overall, the convolutional network outperforms the feed forward in all classes, seeing an ∼10 % improvement for no cloud and single-layer. Only a slight improvement is observed for multi-layer clouds however, with PoD increasing from 58 % to 61 %. In general, little effect over overfitting is observed, with the multi-layer PoD being within ∼ 5 % of that observed during the CV training process.

Over the ocean, the feed forward network achieves only 40 % PoD of no cloud classifications. This is improved to 67 % by the convolutional network, but remains significantly poorer than the PoD observed in the class over land. This may be related to the availability of training data, which was limited to ∼ 9 % over the ocean for this class (Table 3). Over land, the PoD of no cloud is 93 % for the convolutional network, the highest PoD recorded for any class over either land or ocean. Conversely, the PoD of multi-layer clouds over land is at a minimum for both networks, with the feed forward observing 23 % PoD and the convolutional network a marginal improvement to 28 %. This is likely the result of two key factors. Firstly, the significantly fewer multi-layer clouds were observed over land in the training data, which reduces the ability for the network to learn an accurate set of weights and biases. And second, the AHI's high viewing angle in this high latitude zone results in a more significant parallax effect for the lower layers of any multi-layer cloud. This is a result of the parallax correction method employed, which corrected only for the top cloud layer, and not the entire atmospheric column. Further, the high viewing angle in this "limb" region of the AHI field of view means the radiation observed has passed through a greater portion of the atmosphere, impacting the values recorded in some observation channels.

5.2 Model interpretation

Global Shapley values are estimated using SHAP's model agnostic KernelSHAP (KernelExplainer) method detailed in Section 3.2. The background expectation is defined using 5000 observations from the training data and SHAP values were estimated for a random sample of 1000 unseen validation set observations. For each prediction assessed, a perturbation sample

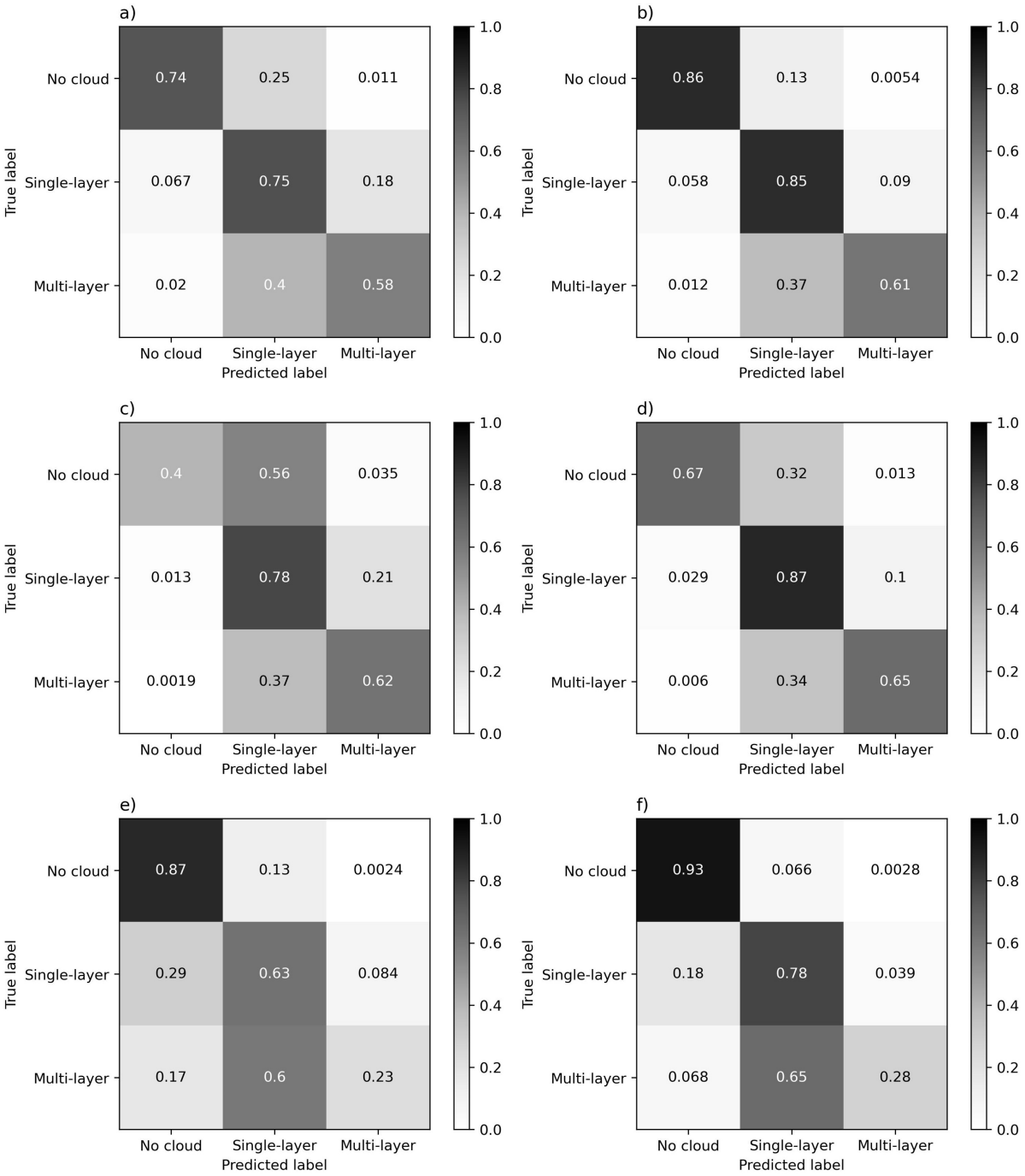

Figure 15: Confusion matrix for validation set classification results, normalised for each true label, i.e., diagonal entries correspond to the Probability of Detection for each class. Results are for the feed forward (left) and convolutional (right) networks for all data (top), over ocean (middle) and over land (bottom).

size of 50 was used. The result across all features for a given prediction sums to the difference between that prediction and the background expectation. The global 'importance' is then defined as the mean of the absolute value of each SHAP result. Results are given for each network in Figures 16 and 17.

Band 5 was an important predictor of both no cloud and single-layer cloud for both algorithms. Band 5 is a near-infrared 1.6μm band sensitive to water phase, a good indicator of cloud presence. Elevation angle, band 7 (sensitive to low cloud) and band 13 (atmospheric window) were also important predictors of no cloud in both models. Band 6 (phase), surface type and solar azimuth angle were important for the feed forward network but not for the convolutional. Whereas bands 11 (phase), 14 and 15 (windows) were important for the convolutional network at both time steps. This suggests that the convolutional network is using changes in the atmospheric windows to detect cloud presence in the column, as well as changes in temperature and phase. In contrast, the importance of the surface type and solar angle for the feed forward network suggests the algorithm may be using climatological information.

The networks saw more overlap in important features for single-layer predictions, with bands 5 (phase), 10 (low level vapour), 8 (high level vapour) and 11 (phase) important predictors in both algorithms. Band 7 (low level cloud) was also important for the feed forward, but not for the convolutional network. Conversely, bands 13 and 15 (windows) were important for the convolutional only.

For multi layer cloud, both networks saw different predictors as most important. In common, contributions were strong for band 8 (high level vapour), 7 (low level cloud), 10 (low level vapour) and 15 (window). The feed forward also saw elevation angle, latitude and 6 (phase) as important predictors which were not important for the convolutional network. Conversely, band 5 (phase), 11 (phase) and 13 (window) were important for the convolutional but not the feed forward. This indicates that the convolutional network is using changes in water vapour presence, temperature and phase to detect multi-layer cloud occurrences. Temperature differences may be a good indicator in successive temporal observations, as a warm followed by cool section of cloud indicates clouds at different altitudes, a likely indicator of a multi-layer system. Again, the feed forward network is using more geolocation variables, which suggests it is learning multi-layer cloud climatologies, or attempting to correct for the AHI viewing angles, which vary with latitude and elevation angle.

In general, the convolutional network makes more use of the atmospheric window bands and bands sensitive to phase, while the feed forward saw more importance given to geolocation variables, such as latitude, surface type and observation or solar angles. The co-occurrence of both time steps for each of the most important bands in the convolutional SHAP results indicates that the algorithm is indeed making use of changes in spectral conditions to perform all classifications. Further supporting the statistical conclusion that the addition of temporal information to the model presents an advancement on existing literature. Given the difference in viewing angle between the AHI and the CPR-CALIOP truth labels at high latitudes, the addition of AHI data from the preceding 10 minute observation may account for skewed labelling while the feed forward requires the geolocation predictors to achieve this correction.

Least important in both algorithms were the visible channels and the downscaled standard deviations (sigma's). The unimportance of the visible channels aligns with expectations for the multi-layer and single-layer classifications, as these channels are typically good at detecting land surface parameters, not distinguishing cloud properties. For the no cloud classification, the ice surface of the land within the spatial zone (Antarctica) reduces the usefulness of these channels for detecting no cloud.

The relative importance of bands 5 and 6 (both sensitive to phase), and bands 13, 14 and 15 (window bands) should not be distinguished, as these groups were highly correlated, observing

high VIF scores (Figure 10). For example, this means that while band 5 was scored more important than band 6 for no cloud classification in the feed forward network. However, due to their high collinearity band 6 may in truth be more important, or band 6 may not have been important at all, merely overblown in value by the high importance of its correlated pair. Regardless, both bands are sensitive to phase, indicating that the near-infrared phase channels are a good indicator of a clear atmosphere.

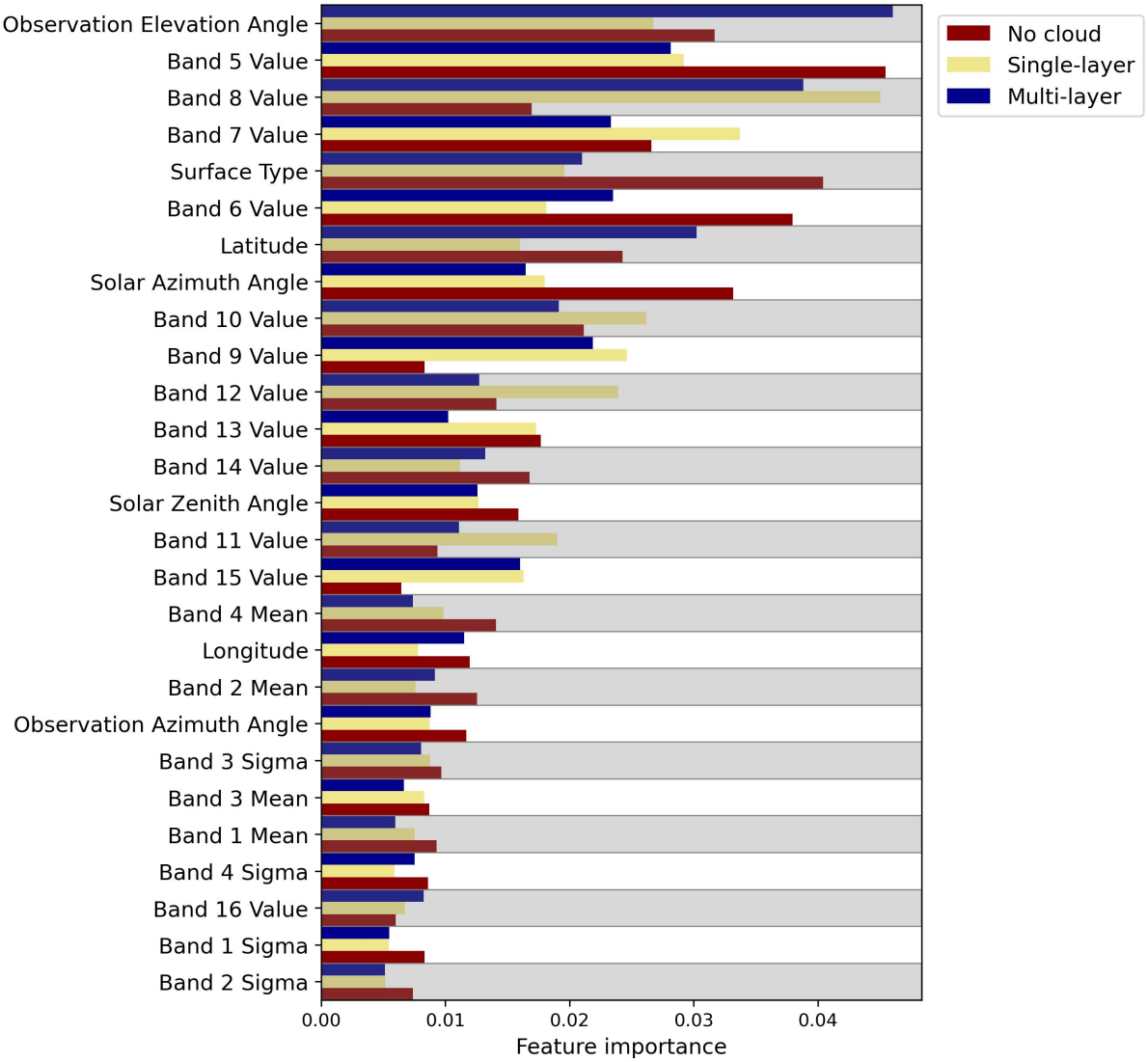

Figure 16: SHAP feature importance results for the feed forward neural network.

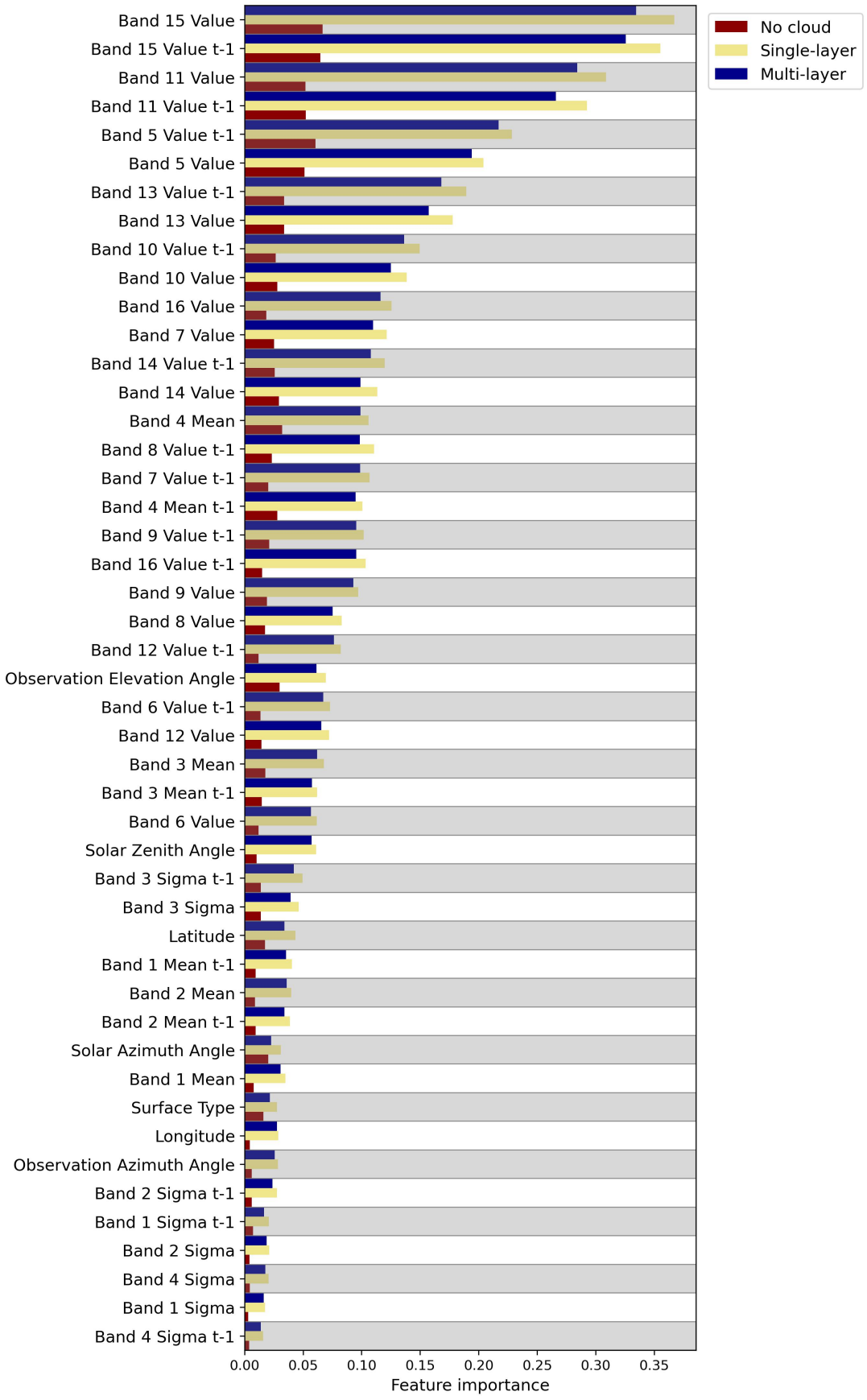

Figure 17: SHAP feature importance results for the convolutional neural network.

6 Discussion

6.1 Comparison with past work

The feed forward and convolutional networks achieved PoD scores of 58 % and 61 % respectively. Both offering significant improvements to the MODIS 6 operational decision-tree based algorithm, which achieved 34 % over the full disk. However, neither achieve the accuracy of the day-only Tan et al. (2021) or Li et al. (2022) feed forward networks which were also trained for the Himawari-8/9 AHI, and both reached PoD of 70 %.

A key reason for this may be the difference in spatial domains. While our algorithms were developed for the SO, defined as greater than 50° South, both the Tan et al. and Li et al. algorithms were developed for the full AHI field of view. Aside from more data available for training, this has the advantage of a smaller portion of observations recorded with large viewing angles, as occurs at high latitudes. While the effect of disagreeing viewing angles between AHI and the truth data is accounted for through parallax correction, this correction was performed for the top cloud layer, not the entire atmospheric column. Hence, some parallax effect likely persists in scenes where distance between multiple cloud layers is moderate to high. Further, surface ice poses greater difficulty in cloud detection than both oceanic and other land surface types. A significantly greater portion of our spatial domain was comprised of surface ice, likely contributing to the reduced PoD of our algorithms. Also, some observational issues associated with the "limb" region may have further reduced PoD scores.

Future refinements of the developed algorithms could include extending the definition of the SO to cover lower latitude regions (such as from -35◦ South) neglecting the high latitude Antarctic land region, neglecting unimportant predictor variables (such as the sigma's and visible bands) and including additional time steps of past AHI observations for convolution. Further improvements would also be gained by performing a full-column parallax correction, as detailed in Robbins et al. (2022). This would likely increase overall multi-layer detection accuracy, but would also increase confidence in SHAP interpretations, uncovering whether the latitude and observation angle variables were important in the feed forward network for climatological reasons or were used to account for the parallax effect.

6.2 Case studies and key limitations

Reduced multi-layer PoD over land than over ocean was a key characteristic and limitation of both models. This is likely due to the ice surface, which challenges many cloud property detection algorithms. Figure 18 shows two scenes which extend from 50° South to 80° South, i.e., cover both oceanic and land surfaces. In Figure 18a, both algorithms identify multi-layer cloud from ∼50-57° South and again at ∼62° South. At higher latitudes however (over Antarctica), ∼71° South and 75° South, both algorithms mislabel an instance of multi-layer and single-layer cloud respectively. Similarly, Figure 18b shows both algorithms identifying multi-layer cloud at ∼54° South and mislabel single-layer cloud as multi-layer at the higher latitude observation at ∼67° South.

In both these examples, the convolutional network appears to outperform the feed forward, seeing a closer representation of the "truth" at high latitudes. To explore this further, Figure 19 shows two scenes in which the feed forward algorithm had mislabelled single-layer cloud as multi-layer while the convolutional network maintained realistic predictions. In Figure 19a, we observe a geometrically thick cloud with interrupted regions of multi-layer structure from ∼57-64° South. The feed forward network incorrectly predicts multi-layer structure for the entire span of the cloud, where the convolutional network is able to represent the broken structure

appropriately. In Figure 19b, geometrically thick cloud with variations in structure spans from ∼58-74° South, reaching over Antarctica. Here, the feed forward network predicts large spans of multi-layer cloud where only single-layer are present, and only accurately captures three of the five interrupted multi-layer cloud occurrence (for which predictions are available). In contrast, the convolutional network only predicts true positives, and detects four out of the five multi-layer occurrences (all excluding that at ∼70° South). These two examples are representative of the 100+ considered in this analysis, which saw frequent false positives (commonly occurring for geometrically thick single-layer cloud) and undetected multi-layer occurrences by the feed forward network where the convolutional network remained realistic.

A typical limitation of multi-layer cloud detection in general is the challenge detecting low level layers under optically or geometrically thick upper layer cloud. This is because a passive sensor will only be able to detect a multi-layer cloud when there is some contribution of the lower layer to the top-of-atmosphere radiance, which is only possible when the upper layer is reasonably thin i.e < 3-5 optical depths. We observe some examples of this, although many scenes of this type are well predicted by both models (e.g., between ∼50-53° South in Figure 18a). The most common misclassification of multi-layer clouds appears to be for geometrically thin upper level cloud. For instance, Figure 19b shows the convolutional network misclassifying multi-layer cloud at ∼69° South where the upper layer is geometrically thin.

In general, the case studies considered reaffirm the statistical and SHAP results which indicated greater performance by the convolutional network. The number of false positives by the feed forward network in conjunction with the SHAP results, which revealed greater reliance on geolocation rather than optical predictors, also suggest the feed forward algorithm has learnt unphysical indicators of multi-layer occurrence.

6.3 Implications for multi-layer cloud analyses

An important advancement offered by accurate multi-layer cloud detection in passive satellite data is the availability of temporally frequent and spatially dense climatological data. In particular, Himawari-8/9's AHI observes the Australasian region at a spatial resolution of 2 km every 10 minuets. In comparison, the merged CPR-CALIOP data sample a track within the Himawari-8/9 field of view every ∼60 minutes and return to the same geographic region only once every 16 days. Most significant is the temporal continuity of observation for each Himawari-8/9 pixel. Where previously limited to the revisit time of the merged active sensors, frequent detection my the geostationary instrument facilitates climatological analyses previously untenable, such as assessments of the diurnal cycle.

The modest success of our feed forward algorithm and improved accuracy of our convolutional algorithm over this high latitude zone offer the means to produce large data sets of multi-layer cloud occurrences in the Himawari-8/9 field of view. While our models are developed for the Himawari-8/9 satellite, our results are relevant for all geostationary instruments, meaning more complete coverage of the SO could be acquired. This increased observational data provision facilitates new detail in the interrogation of multi-layer cloud occurrence and radiative properties over the SO.

Additionally, our results advance progress toward accurate operational multi-layer cloud detection and improvements in cloud property retrievals. The next stages of which involve developing appropriate means of incorporating multi-layer cloud flags into the radiative transfer models, to improve retrievals of cloud properties, such as cloud top height, phase and temperature.

Future analysis would also benefit from a quantitative assessment of accuracy and PoD scores for optically thick upper or thin upper layer cloud. This would further interrogate the relative

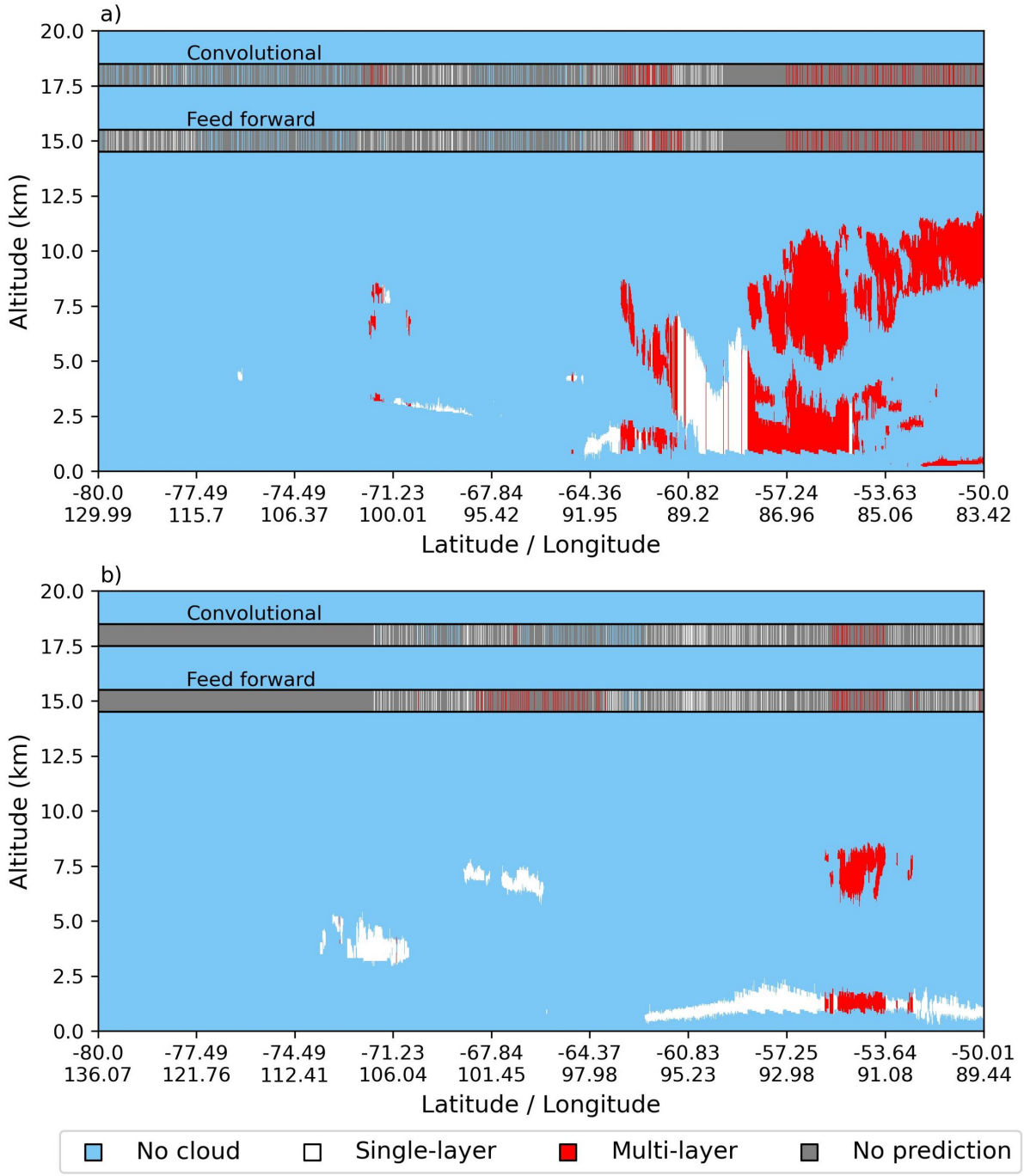

Figure 18: Example curtain plot of the CPR-CALIOP cloud layer "truth" classification with model predictions overlaid at 15 km (feed forward results) and 17.5 km (convolutional results). Where 'no prediction' results from a failed collocation process. Both (a) 2019-01-23 08:17 and (b) 2019-04-05 07:54 show a mixture of good and poor performance by both models.

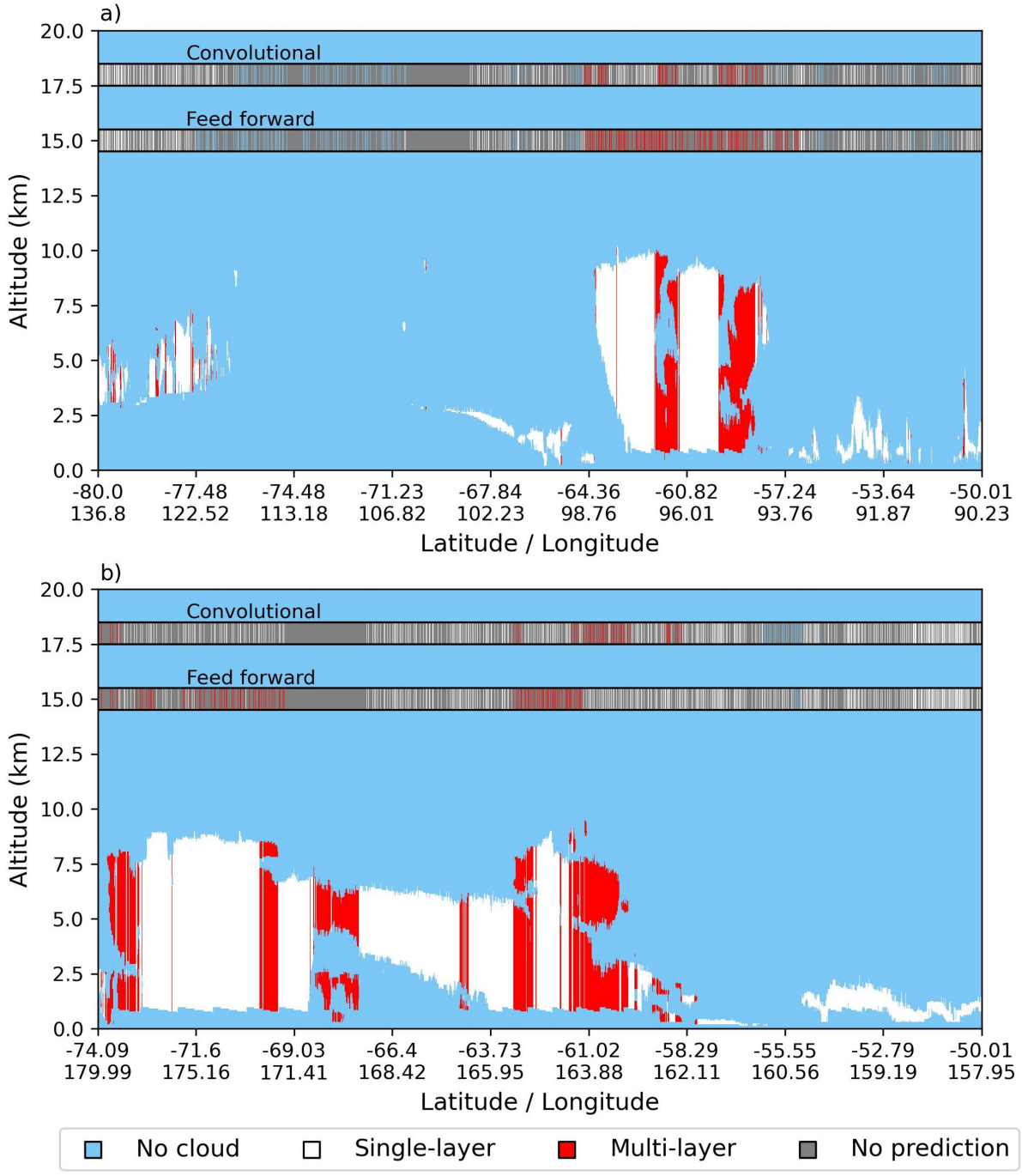

Figure 19: Example curtain plot of the CPR-CALIOP cloud layer "truth" classification with model predictions overlaid at 15 km (feed forward results) and 17.5 km (convolutional results). Where 'no prediction' results from a failed collocation process. Both (a) 2019-02-25 07:50 and (b) 2019-03-21 03:20 show scenes in which the convolutional network outperforms the feed forward.

performance of the algorithms for key cloud structures. An important consideration to inform reliability of climatological assessments of multi-layer cloud physical parameters and radiative properties over the SO.

7 Concluding remarks

We have developed two neural network alternatives to existing multi-layer cloud classification algorithms. Finding that, despite the relatively limited number of training data used and high viewing angles over the SO, significant improvements in multi-layer cloud detection can be achieved with these deep learning alternatives.

The convolutional network offered further increases in accuracy compared to the feed forward algorithm. Whether this was largely due to the additional data from the preceding AHI scene or the convolution process is unclear. Further developments could interrogate this further by including more historic data from preceding scenes with network architecture unchanged.

We have also conducted a careful interpretation of the two models, using the SHAP implementation of Shapley value theory to assign importance scores to the predictor variables. These indicated the the feed forward algorithm relied more on geolocation variables and less on spectral radiance's than the convolutional network. This suggests that the convolutional algorithm may have learnt more physical indicators of multi-layer cloud occurrence, while the feed forward may be biased to climatological information associated with predictors such as latitude. I.e., the feed forward algorithm may suffer reliability issues when used in different time periods or spatial zones. This is further supported by the assessment of multiple case studies, in which the prediction results of each algorithm are compared to the CPR-CALIOP "truth" data. These showed frequent false positives and undetected multi-layer cloud occurrences by the feed forward network where the convolutional algorithm performed well.

The adoption of SHAP model interpretation methods offers many advantages to deep learning applications. In particular, it aides in debugging of model performance, improving confidence in the selection and reliability of demonstrably robust algorithms. For example, while the PoD of our feed forward and convolutional networks were comparable (58 % and 61 % respectively), the SHAP analyses demonstrated that the convolutional network was more physically sound. This implies that the convolutional algorithm offers a strong improvement to the feed forward, even whilst neglecting the large overall accuracy improvements observed. In contrast, without considering SHAP, the small improvement in PoD may have led the developer to the conclusion that the additional data and computational cost of the convolutional network were not worthwhile.

Additionally, SHAP analyses aide in understanding the spectral channels important for cloud vertical structure classification more generally, or the detection of any atmospheric variable targeted in deep learning. These results can be used in model development, to select key input variables and refine an algorithm. They can also inform design of future satellite instruments, supporting the long-term advancement of the field in question.

The success of our algorithms, in particular the convolutional network, facilitates a significant advancement in our understanding of SO cloud processes in the Australasian region. The accuracy of our algorithms increases the availability of observed multi-layer cloud occurrences in AHI data, enabling new and more detailed climatological analyses to be performed in passive satellite data. In particular, as observations by geostationary satellites are high frequency in time at each pixel, our results facilitate temporal assessments of multi-layer cloud climatology that were previously limited by the long revisit time of polar-orbiting sensors.

Further, the improved accuracy of the convolutional network highlights the benefit of incorporating temporal information and convolution layers to deep learning algorithms in future remote sensing detection efforts. The consideration of temporal information had thus far not been undertaken in efforts to detect multi-layer cloud occurrences (Li et al., 2022; Tan et al., 2021), and has seen limited uptake in other related fields. In particular, the no cloud PoD

was increased to 86 % overall from 74 % between the convolutional and feed forward networks. This significant improvement, achieved over ice surface, suggests that temporal convolutional networks may offer advances to cloud masking algorithms, especially at high latitudes.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden, P., ... Zhang, X. (2016). Tensorflow: A system for large-scale machine learning. *CoRR*, *abs/1605.08695*. http://dblp.uni-trier.de/db/journals/corr/corr1605.html%5C%AbadiBCCDDDGIIK16

Atkinson, J. (2018). *Sister satellites, briefly separated, working together again.* https://www.nasa.gov/feature/langley/sister-satellites-briefly-separated-working-together-again

Baum, B. A., Uttal, T., Poellot, M., Ackerman, T. P., Alvarez, J. M., Intrieri, J., Starr, D., Titlow, J., Tovinkere, V., & Clothiaux, E. (1995). Satellite remote sensing of multiple cloud layers. *Journal of Atmospheric Sciences*, *52*, 4210–4230. https://doi.org/10.1175/1520-0469(1995)052⟨4210:SRSOMC⟩2.0.CO;2

Baum, B. A., Arduini, R. F., Wielicki, B. A., Minnis, P., & Tsay, S.-C. (1994). Multilevel cloud retrieval using multispectral hirs and avhrr data: Nighttime oceanic analysis. *Journal of Geophysical Research: Atmospheres*, *99*, 5499–5514. https://doi.org/https://doi.org/10.1029/93JD02856

Baum, B. A., & Spinhirne, J. D. (2000). Remote sensing of cloud properties using modis airborne simulator imagery during success: 3. cloud overlap. *Journal of Geophysical Research: Atmospheres*, *105*, 11793–11804. https://doi.org/https://doi.org/10.1029/1999JD901091

Bessho, K., Date, K., Hayashi, M., Ikeda, A., Imai, T., Inoue, H., Kumagi, Y., Miyakawa, T., Murata, H., Ohno, T., Okuyama, A., Oyama, R., Sasaki, Y., Shimazu, Y., Shimoji, K., Sumida, Y., Suzuki, M., Taniguchi, H., Tsuchiyama, H., ... Yoshida, R. (2016). An introduction to himawari-8/9andmdash; japanandrsquo;s new-generation geostationary meteorological satellites. *Journal of the Meteorological Society of Japan. Ser. II*, *94*, 151–183. https://doi.org/10.2151/jmsj.2016-009

Bodas-Salcedo, A. (2014). Origins of the solar radiation biases over the southern ocean in cfmip2 models. *J. Climate*, *27*, 41–56. https://doi.org/10.1175/JCLI-D-13-00169.1

Bodas-Salcedo, A., Andrews, T., Karmalkar, A. V., & Ringer, M. A. (2016). Cloud liquid water path and radiative feedbacks over the southern ocean. *Geophys. Res. Lett.*, *43*, 10938–10946. https://doi.org/10.1002/2016GL070770

Bony, S., Colman, R., Kattsov, V. M., Allan, R. P., Bretherton, C. S., Dufresne, J.-L., Hall, A., Hallegatte, S., Holland, M. M., Ingram, W., Randall, D. A., Soden, B. J., Tselioudis, G., & Webb, M. J. (2006). How well do we understand and evaluate climate change feedback processes? *Journal of Climate*, *19*, 3445–3482. https://doi.org/10.1175/JCLI3819.1

Chen, M., Davis, J. M., Liu, C., Sun, Z., Zempila, M. M., & Gao, W. (2017). Using deep recurrent neural network for direct beam solar irradiance cloud screening. *Proc.SPIE*, *10405*. https://doi.org/10.1117/12.2273364

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. https://doi.org/10.48550/ARXIV.1409.1259

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

Desmons, M., Ferlay, N., Parol, F., Riédi, J., & Thieuleux, F. (2017). A global multilayer cloud identification with polder/parasol. *Journal of Applied Meteorology and Climatology*, *56*, 1121–1139. https://doi.org/10.1175/JAMC-D-16-0159.1

Di Mauro, N., Vergari, A., Basile, T., Ventola, F., & Esposito, F. (2017). End-to-end learning of deep spatio-temporal representations for satellite image time series classification. In R. Corizzo & D. Ienco (Eds.), *Proceedings of the ecml/pkdd discovery challenges co-located with european conference on machine learning - principle and practice*

of knowledge discovery in database (ecml pkdd 2017). CEUR Workshop Proceedings. http://ecmlpkdd2017.ijs.si/index.html

Eyre, J. R., Bell, W., Cotton, J., English, S. J., Forsythe, M., Healy, S. B., & Pavelin, E. G. (2022). Assimilation of satellite data in numerical weather prediction. part ii: Recent years. *Quarterly Journal of the Royal Meteorological Society*, *148*, 521–556. https://doi.org/https://doi.org/10.1002/qj.4228

Eyre, J. R., English, S. J., & Forsythe, M. (2020). Assimilation of satellite data in numerical weather prediction. part i: The early years. *Quarterly Journal of the Royal Meteorological Society*, *146*, 49–68. https://doi.org/https://doi.org/10.1002/qj.3654

Fiddes, S. L., Protat, A., Mallet, M. D., Alexander, S. P., & Woodhouse, M. T. (2022). Southern ocean cloud and shortwave radiation biases in a nudged climate model simulation: Does the model ever get it right? *Atmos. Chem. Phys. Discuss.*, *2022*, 1–34. https://doi.org/10.5194/acp-2022-259

Flato, G. (2013). *Climate change 2013: The physical science basis*. Cambridge University Press.

Goosse, H., Kay, J. E., Armour, K. C., Bodas-Salcedo, A., Chepfer, H., Docquier, D., Jonko, A., Kushner, P. J., Lecomte, O., Massonnet, F., Park, H.-S., Pithan, F., Svensson, G., & Vancoppenolle, M. (2018). Quantifying climate feedbacks in polar regions. *Nature Communications*, *9*, 1919. https://doi.org/10.1038/s41467-018-04173-0

Haynes, J. M., Noh, Y.-J., Miller, S. D., Haynes, K. D., Ebert-Uphoff, I., & Heidinger, A. (2022). Low cloud detection in multilayer scenes using satellite imagery with machine learning methods. *Journal of Atmospheric and Oceanic Technology*, *39*, 319–334. https://doi.org/10.1175/JTECH-D-21-0084.1

Heidinger, A. K., & Pavolonis, M. J. (2005). Global daytime distribution of overlapping cirrus cloud from noaa's advanced very high resolution radiometer. *Journal of Climate*, *18*, 4772–4784. https://doi.org/10.1175/JCLI3535.1

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–1780.

Huang, J., Minnis, P., Lin, B., Yi, Y., Fan, T.-F., Sun-Mack, S., & Ayers, J. K. (2006). Determination of ice water path in ice-over-water cloud systems using combined modis and amsr-e measurements. *Geophysical Research Letters*, *33*. https://doi.org/https://doi.org/10.1029/2006GL027038

Huang, J., Minnis, P., Lin, B., Yi, Y., Khaiyer, M. M., Arduini, R. F., Fan, A., & Mace, G. G. (2005). Advanced retrievals of multilayered cloud properties using multispectral measurements. *Journal of Geophysical Research: Atmospheres*, *110*. https://doi.org/https://doi.org/10.1029/2004JD005101

Huang, Y., Siems, S. T., Manton, M. J., Protat, A., & Delanoë, J. (2012). A study on the low-altitude clouds over the southern ocean using the dardar-mask. *Journal of Geophysical Research: Atmospheres*, *117*. https://doi.org/https://doi.org/10.1029/2012JD017800

Ienco, D., Gaetano, R., Dupaquier, C., & Maurel, P. (2017). Land cover classification via multi-temporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, *14*, 1685–1689.

Im, E., Wu, C., & Durden, S. L. (2005). Cloud profiling radar for the cloudsat mission. *IEEE International Radar Conference, 2005.*, 483–486. https://doi.org/10.1109/RADAR.2005.1435874

Jin, Y., & Rossow, W. B. (1997). Detection of cirrus overlapping low-level clouds. *Journal of Geophysical Research: Atmospheres*, *102*, 1727–1737. https://doi.org/https://doi.org/10.1029/96JD02996

Joiner, J., Vasilkov, A. P., Bhartia, P. K., Wind, G., Platnick, S., & Menzel, W. P. (2010). Detection of multi-layer and vertically-extended clouds using a-train sensors. *Atmos. Meas. Tech.*, *3*, 233–247. https://doi.org/10.5194/amt-3-233-2010

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980

Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., & Seland, Ø. (2022). Machine learning of cloud types shows higher climate sensitivity is associated with lower cloud biases. *Atmos. Chem. Phys. Discuss.*, *2022*, 1–32. https://doi.org/10.5194/acp-2022-184

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *PP*, 1–5. https://doi.org/10.1109/LGRS.2017.2681128

Lakhal, M. I., Çevikalp, H., Escalera, S., & Ofli, F. (2018). Recurrent neural networks for remote sensing image classification [https://doi.org/10.1049/iet-cvi.2017.0420]. *IET Computer Vision*, *12*, 1040–1045. https://doi.org/https://doi.org/10.1049/iet-cvi.2017.0420

Li, J., Yi, Y., Minnis, P., Huang, J., Yan, H., Ma, Y., Wang, W., & Ayers, J. K. (2011). Radiative effect differences between multi-layered and single-layer clouds derived from ceres, calipso, and cloudsat data. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *112*, 361–375. https://doi.org/https://doi.org/10.1016/j.jqsrt.2010.10.006

Li, W., Zhang, F., Lin, H., Chen, X., Li, J., & Han, W. (2022). Cloud detection and classification algorithms for himawari-8 imager measurements based on deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–17. https://doi.org/10.1109/TGRS.2022.3153129

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., & Collins, W. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/arxiv.1705.07874

Mace, G. G., Protat, A., Humphries, R. S., Alexander, S. P., McRobert, I. M., Ward, J., Selleck, P., Keywood, M., & McFarquhar, G. M. (2021). Southern ocean cloud properties derived from capricorn and marcus data. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033368. https://doi.org/https://doi.org/10.1029/2020JD033368

Mace, G. G., Zhang, Q., Vaughan, M., Marchand, R., Stephens, G., Trepte, C., & Winker, D. (2009). A description of hydrometeor layer occurrence statistics derived from the first year of merged cloudsat and calipso data. *Journal of Geophysical Research: Atmospheres*, *114*. https://doi.org/https://doi.org/10.1029/2007JD009755

Marchant, B., Platnick, S., Meyer, K., & Wind, G. (2020). Evaluation of the modis collection 6 multilayer cloud detection algorithm through comparisons with cloudsat cloud profiling radar and calipso caliop products. *Atmospheric Measurement Techniques*, *13*, 3263–3275. https://doi.org/10.5194/amt-13-3263-2020

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., & Gomis, M. I. (2021). Climate change 2021: The physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2.

Minh, D. H. T., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., & Maurel, P. (2018). Deep recurrent neural networks for winter vegetation quality mapping via multitemporal sar sentinel-1. *IEEE Geoscience and Remote Sensing Letters*, *15*, 464–468.

Minnis, P., Huang, J., Lin, B., Yi, Y., Arduini, R. F., Fan, T.-F., Ayers, J. K., & Mace, G. G. (2007). Ice cloud properties in ice-over-water cloud systems using tropical rainfall measuring mission (trmm) visible and infrared scanner and trmm microwave imager data. *Journal of Geophysical Research: Atmospheres*, *112*. https://doi.org/https://doi.org/10.1029/2006JD007626

Minnis, P., Sun-Mack, S., Jr., W. L. S., Hong, G., & Chen, Y. (2019). Advances in neural network detection and retrieval of multilayer clouds for ceres using multispectral satellite data. *Proc.SPIE*, *11152*. https://doi.org/10.1117/12.2532931

Morgenstern, O., & Neumann, J. V. (1953). *Theory of games and economic behavior*. Princeton university press.

Murugan, P. (2018). Learning the sequential temporal information with recurrent neural networks. https://doi.org/10.48550/ARXIV.1807.02857

Narendra, K. S., Member, S., & Thathachar, M. A. L. (1974). Learning automata - a survey. *IEEE Trans. Systems, Man., Cybernetics*, 323–334.

Nasiri, S. L., & Baum, B. A. (2004). Daytime multilayered cloud detection using multispectral imager data. *Journal of Atmospheric and Oceanic Technology*, *21*, 1145–1155. https://doi.org/10.1175/1520-0426(2004)021⟨1145:DMCDUM⟩2.0.CO;2

Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., & Hossard, L. (2018). Deep recurrent neural network for agricultural classification using multitemporal sar sentinel-1 for camargue, france. *Remote Sensing*, *10*(8). https://doi.org/10.3390/rs10081217

Oreopoulos, L., Cho, N., & Lee, D. (2017). New insights about cloud vertical structure from cloudsat and calipso observations. *Journal of Geophysical Research: Atmospheres*, *122*, 9280–9300. https://doi.org/https://doi.org/10.1002/2017JD026629

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

Owen, A. B., & Prieur, C. (2016). On shapley value for measuring importance of dependent inputs. https://doi.org/10.48550/arxiv.1610.02080

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, *2*, 1.

Pavolonis, M. J., & Heidinger, A. K. (2004). Daytime cloud overlap detection from avhrr and viirs. *Journal of Applied Meteorology*, *43*, 762–778. https://doi.org/10.1175/2099.1

Pelletier, C., Webb, G. I., & Petitjean, F. (2018). Temporal convolutional neural network for the classification of satellite image time series. https://doi.org/10.48550/ARXIV.1811.10166

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., & Riedi, J. (2017). The modis cloud optical and microphysical products: Collection 6 updates and examples from terra and aqua (2016/10/26). *IEEE transactions on geoscience and remote sensing : a publication of the IEEE Geoscience and Remote Sensing Society*, *55*, 502–525. https://doi.org/10.1109/TGRS.2016.2610522

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

Raspaud, M., Hoese, D., Dybbroe, A., Lahtinen, P., Devasthale, A., Itkin, M., Hamann, U., Rasmussen, L., Nielsen, E. S., Leppelt, T., Maul, A., Kliche, C., & Thorsteinsson, H. (2022). Pytroll: An open-source, community-driven python framework to process earth observation satellite data. *Bulletin of the American Meteorological Society*, *99*(7), 1329–1336. https://doi.org/10.1175/BAMS-D-17-0277.1

Rawat, A., Kumar, A., Upadhyay, P., & Kumar, S. (2021). Deep learning-based models for temporal satellite data processing: Classification of paddy transplanted fields. *Ecological Informatics*, *61*, 101214. https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101214

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. https://doi.org/10.48550/arxiv.1606.05386

Robbins, D., Poulsen, C., Siems, S., & Proud, S. (2022). Improving discrimination between clouds and optically thick aerosol plumes in geostationary satellite data. *Atmos. Meas. Tech.*, *15*, 3031–3051. https://doi.org/10.5194/amt-15-3031-2022

Robbins, D., Poulsen, C., Proud, S., & Siems, S. (2021). *Ahi-caliop collocated data for training and validation of cloud masking neural networks* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5773420

Robbins, D., & Proud, S. (2022). *Dr1315/ahinn: Ahinn initial release* (Version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.6538854

Schuddeboom, A., Varma, V., McDonald, A. J., Morgenstern, O., Harvey, M., Parsons, S., Field, P., & Furtado, K. (2019). Cluster-based evaluation of model compensating errors: A case study of cloud radiative effect in the southern ocean. *Geophysical Research Letters*, *46*, 3446–3453. https://doi.org/https://doi.org/10.1029/2018GL081686

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*, 2673–2681.

Shapley, L. S. (1953). *17. a value for n-person games* (H. W. Kuhn & A. W. Tucker, Eds.). Princeton University Press. https://doi.org/doi:10.1515/9781400881970-018

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., ... Zelinka, M. D. (2020). An assessment of earth's climate sensitivity using multiple lines of evidence [https://doi.org/10.1029/2019RG000678]. *Reviews of Geophysics*, *58*, e2019RG000678. https://doi.org/https://doi.org/10.1029/2019RG000678

Slingo, A., & Slingo, J. M. (1988). The response of a general circulation model to cloud longwave radiative forcing. i: Introduction and initial experiments. *Quarterly Journal of the Royal Meteorological Society*, *114*, 1027–1062. https://doi.org/https://doi.org/10.1002/qj.49711448209

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Stephens, G., & Christian, K. (2007). The remote sensing of clouds and precipitation from space: A review. *Journal of The Atmospheric Sciences - J ATMOS SCI*, *64*, 3742–3765. https://doi.org/10.1175/2006JAS2375.1

Stephens, G. (2002). *Cirrus, climate, and global change*. https://doi.org/10.1093/oso/9780195130720.003.0024

Stephens, G., Winker, D., Pelon, J., Trepte, C., Vane, D., Yuhas, C., L'Ecuyer, T., & Lebsock, M. (2018). Cloudsat and calipso within the a-train: Ten years of actively observing the earth system. *Bulletin of the American Meteorological Society*, *99*, 569–581. https://doi.org/10.1175/BAMS-D-16-0324.1

Stephens, G. L., & Webster, P. J. (1984). Cloud decoupling of the surface and planetary radiative budgets. *Journal of Atmospheric Sciences*, *41*, 681–686. https://doi.org/10.1175/1520-0469(1984)041⟨0681:CDOTSA⟩2.0.CO;2

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*, 647–665. https://doi.org/10.1007/s10115-013-0679-x

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Girolamo, L. D., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., ... Zhao, G. (2013). Assessment of global cloud datasets from satellites: Project and database initiated by the gewex radiation panel. *Bulletin of the American Meteorological Society*, *94*, 1031–1049. https://doi.org/10.1175/BAMS-D-12-00117.1

Sun, Z., Di, L., & Fang, H. (2019). Using long short-term memory recurrent neural network in land cover classification on landsat and cropland data layer time series. *International*

Journal of Remote Sensing, *40*, 593–614. https://doi.org/10.1080/01431161.2018.1516313

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. https://doi.org/10.48550/arxiv.1703.01365

Tan, Z., Liu, C., Ma, S., Wang, X., Shang, J., Wang, J., Ai, W., & Yan, W. (2021). Detecting multilayer clouds from the geostationary advanced himawari imager using machine learning techniques. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–12. https://doi.org/10.1109/TGRS.2021.3087714

Tibshirani, D. W. G. J. T. H. R. (2013). *An introduction to statistical learning : With applications in r* [Includes bibliographical references and index.]. New York : Springer, [2013] ©2013. https://search.library.wisc.edu/catalog/9910207152902121

Tovinkere, V. R., Penaloza, M., Logar, A., Lee, J., Weger, R. C., Berendes, T. A., & Welch, R. M. (1993). An intercomparison of artificial intelligence approaches for polar scene identification. *Journal of Geophysical Research: Atmospheres*, *98*, 5001–5016.

Trenberth, K. E., & Fasullo, J. T. (2010). Simulation of present-day and twenty-first-century energy budgets of the southern oceans. *Journal of Climate*, *23*, 440–454. https://doi.org/10.1175/2009JCLI3152.1

Truong, S. C. H., Huang, Y., Siems, S. T., Manton, M. J., & Lang, F. (2022). Biases in the thermodynamic structure over the southern ocean in era5 and their radiative implications. *International Journal of Climatology*, *n/a*. https://doi.org/https://doi.org/10.1002/joc.7672

v. Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, *100*, 295–320. https://doi.org/10.1007/BF01448847

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

Wang, J., & Rossow, W. B. (1998). Effects of cloud vertical structure on atmospheric circulation in the giss gcm. *Journal of Climate*, *11*, 3010–3029.

Wang, T., Fetzer, E. J., Wong, S., Kahn, B. H., & Yue, Q. (2016). Validation of modis cloud mask and multilayer flag using cloudsat-calipso cloud profiles and a cross-reference of their cloud classifications. *Journal of Geophysical Research: Atmospheres*, *121*, 11, 611–620, 635. https://doi.org/https://doi.org/10.1002/2016JD025239

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. *2017 International Joint Conference on Neural Networks (IJCNN)*, 1578–1585. https://doi.org/10.1109/IJCNN.2017.7966039

Webster, P. J., & Stephens, G. L. (1984). Cloud-radiation interaction and the climate problem.

Wielicki, B. A., Cess, R. D., King, M. D., Randall, D. A., & Harrison, E. F. (1995). Mission to planet earth: Role of clouds and radiation in climate. *Bulletin of the American Meteorological Society*, *76*(11), 2125–2154. https://doi.org/10.1175/1520-0477(1995)076⟨2125:MTPERO⟩2.0.CO;2

Wild, M., Ohmura, A., Gilgen, H., & Roeckner, E. (1995). Validation of general circulation model radiative fluxes using surface observations. *Journal of Climate*, *8*, 1309–1324. https://doi.org/10.1175/1520-0442(1995)008⟨1309:VOGCMR⟩2.0.CO;2

Williams, K. D., & Tselioudis, G. (2007). Gcm intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Climate Dynamics*, *29*, 231–250. https://doi.org/10.1007/s00382-007-0232-2

Wind, G., Platnick, S., King, M. D., Hubanks, P. A., Pavolonis, M. J., Heidinger, A. K., Yang, P., & Baum, B. A. (2010). Multilayer cloud detection with the modis near-infrared water vapor absorption band. *Journal of Applied Meteorology and Climatology*, *49*, 2315–2333. https://doi.org/10.1175/2010JAMC2364.1

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., Hunt, W. H., & Young, S. A. (2009). Overview of the calipso mission and caliop data processing algorithms. *Journal of Atmospheric and Oceanic Technology*, *26*, 2310–2323. https://doi.org/10.1175/2009JTECHA1281.1

Yao, Z., Han, Z., Zhao, Z., Lin, L., & Fan, X. (2010). Synergetic use of polder and modis for multilayered cloud identification. *Remote Sensing of Environment*, *114*, 1910–1923. https://doi.org/https://doi.org/10.1016/j.rse.2010.03.014

Young, H. P. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory*, *14*, 65–72. https://doi.org/10.1007/BF01769885

Zhang, M. H., Lin, W. Y., Klein, S. A., Bacmeister, J. T., Bony, S., Cederwall, R. T., Genio, A. D. D., Hack, J. J., Loeb, N. G., Lohmann, U., Minnis, P., Musat, I., Pincus, R., Stier, P., Suarez, M. J., Webb, M. J., Wu, J. B., Xie, S. C., Yao, M.-S., & Zhang, J. H. (2005). Comparing clouds and their seasonal variations in 10 atmospheric general circulation models with satellite measurements. *Journal of Geophysical Research: Atmospheres*, *110*. https://doi.org/https://doi.org/10.1029/2004JD005021

Zhang, W., Xu, H., & Zheng, F. (2018). Aerosol optical depth retrieval over east asia using himawari-8/ahi data. *Remote Sensing*, *10*(1). https://doi.org/10.3390/rs10010137

Zhong, L., Hu, L., & Zhou, H. (2019). Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, *221*, 430–443. https://doi.org/https://doi.org/10.1016/j.rse.2018.11.032