Probing the Explainability of Neural Network Cloud-Top Pressure Models for LEO and GEO Imagers

CHARLES H. WHITE,[a] ANDREW K. HEIDINGER,[b] AND STEVEN A. ACKERMAN[c]

[a] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[b] *NOAA/NESDIS, Madison, Wisconsin*
[c] *Department of Atmospheric and Oceanic Sciences, University of Wisconsin–Madison, Madison, Wisconsin*

(Manuscript received 4 November 2021, in final form 4 August 2022)

ABSTRACT: Satellite low-Earth-orbiting (LEO) and geostationary (GEO) imager estimates of cloud-top pressure (CTP) have many applications in both operations and in studying long-term variations in cloud properties. Recently, machine learning (ML) approaches have shown improvement upon physically based algorithms. However, ML approaches, and especially neural networks, can suffer from a lack of interpretability, making it difficult to understand what information is most useful for accurate predictions of cloud properties. We trained several neural networks to estimate CTP from the infrared channels of the Visible Infrared Imaging Radiometer Suite (VIIRS) and the Advanced Baseline Imager (ABI). The main focus of this work is assessing the relative importance of each instrument's infrared channels in neural networks trained to estimate CTP. We use several ML explainability methods to offer different perspectives on feature importance. These methods show many differences in the relative feature importance depending on the exact method used, but most agree on a few points. Overall, the 8.4- and 8.6-μm channels appear to be the most useful for CTP estimation on ABI and VIIRS, respectively, with other native infrared window channels and the 13.3-μm channel playing a moderate role. Furthermore, we find that the neural networks learn relationships that may account for properties of clouds such as opacity and cloud-top phase that otherwise complicate the estimation of CTP.

SIGNIFICANCE STATEMENT: Model interpretability is an important consideration for transitioning machine learning models to operations. This work applies several explainability methods in an attempt to understand what information is most important for estimating the pressure level at the top of a cloud from satellite imagers in a neural network model. We observe much disagreement between approaches, which motivates further work in this area but find agreement on the importance of channels in the infrared window region around 8.6 and 10–12 μm, informing future cloud property algorithm development. We also find some evidence suggesting that these neural networks are able to learn physically relevant variability in radiation measurements related to key cloud properties.

KEYWORDS: Cloud retrieval; Remote sensing; Satellite observations; Machine learning; Model interpretation and visualization; Neural networks

1. Introduction

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

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

Inoue (1985) used a split-window (11 and 12 μm) method to obtain cloud-top temperature for cirrus clouds. Heidinger and Pavolonis (2009) used a similar approach for estimating phase, temperature, and COD for cirrus clouds, relying on multiple channels within the 8–13-μm region. Specifically, their approach relies on an optimal estimation method (Rodgers 1976) for the Advanced Very High Resolution Radiometer (AVHRR). This is later formalized as the Algorithm Working Group (AWG) cloud height algorithm (ACHA) for use in operations for the Advanced Baseline Imager (ABI) and Visible Infrared Imaging Radiometer Suite (VIIRS; Heidinger and Li 2017). Optimal cloud analysis (OCA; Poulsen et al. 2012) also uses an optimal estimation method for several cloud properties, including CTP. OCA

Corresponding author: Charles H. White, charles.white@colostate.edu

DOI: 10.1175/AIES-D-21-0001.1 e210001

additionally has the ability to estimate properties of multiple clouds in multilayer scenes (Watts et al. 2011).

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

NNs have also been applied to CTP estimation. Kox et al. (2014) used a simple NN trained with the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) to detect and estimate COD and cloud-top altitude of cirrus clouds with the Spinning Enhanced Visible and Infrared Imager (SEVIRI). Håkansson et al. (2018) also trained an NN with CALIOP to estimate CTP from MODIS measurements. They compared their NN results with operational algorithms and found large improvement even when using a small subset of channels. Pfreundschuh et al. (2018) extended this work to estimate uncertainties in CTP from NN approaches with a quantile loss function.

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

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

Much of the work included in this article is an extension of the second chapter of the first author's Ph.D. dissertation (White 2022).

2. Data

a. *CALIOP*

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

b. *VIIRS*

VIIRS is a LEO imager on the *Suomi National Polar-Orbiting Partnership* (*SNPP*) and *NOAA-20* satellites. VIIRS has a nadir spatial resolution of 750 m for 16 moderate-resolution channels that span visible, near-infrared, and infrared wavelengths. We find coincident observations between *SNPP* VIIRS and CALIOP by nearest-neighbor matching of imager pixels and lidar profiles that occur within 2.5 min. A parallax correction (Holz et al. 2008) is applied, using the VIIRS viewing geometry and the cloud-top altitude reported from CALIOP.

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

In addition to the native VIIRS channels, we also include information from the VIIRS–Cross-track Infrared Sounder (CrIS) fusion channels (Weisz et al. 2017). These are estimates of absorbing channels created from coarse-spatial-resolution measurements from CrIS that are convolved to match the spectral response functions of MODIS. The fusion channels are mapped to the same resolution as the

TABLE 1. Central wavelengths of the infrared channels included in the VIIRS models. The left column indicates whether channels are native VIIRS measurements or derived from the CrIS. Note that fusion channels are named after MODIS bands since they are designed to match spectral response functions of that instrument.

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

VIIRS M bands by exploiting the variability in the native VIIRS infrared channels. Others have found improvement in CTP estimates when including the fusion channels (Li et al. 2020) or CrIS information (Heidinger et al. 2019). They are included here since they represent spectral regions not represented in native VIIRS observations. All bands from VIIRS and the fusion channels used in this work are shown in Table 1.

We partition our VIIRS–CALIOP collocations into a training dataset from 2016 to 2018, a validation dataset from 2017, and a testing dataset from 2019. The spatial and seasonal distribution of collocations are shown in Fig. 1. Differences in the spatial distribution between 2019 and the previous years are due to *CloudSat* and CALIOP's exit from the A-Train (Braun et al. 2019). Gaps in these datasets are primarily due to the unavailability of CALIOP data and a gap in some CrIS channels from April 2019 to June 2019.

c. ABI

ABI (Schmit et al. 2017) is an imager on the *Geostationary Operational Environmental Satellite (GOES)-16* and *GOES-17* platforms. The infrared channels considered in this work have a nadir spatial resolution of 2 km. The temporal resolution of ABI full-disk images can vary depending on the scan mode. We use the GOES-16 ABI data from 2019, in which the temporal resolution is mainly 10 min.

The *GOES-16* ABI and CALIOP collocations are found in a similar way to those of VIIRS and CALIOP. One difference is that we relax the time difference requirement to 5 min. We make this change since the nadir resolution of ABI is more than twice as large as the VIIRS M bands, and it is less likely that a cloud observed by CALIOP is advected out of the matched imager pixel when the area observed by the pixel is larger. In our models we include all ABI channels without solar contributions, which includes bands 8–16 (Table 2).

The collocations with CALIOP are partitioned into a training dataset from January 2019 through June 2019, a validation dataset from July 2019 through September 2019, and a testing dataset from October 2019 through December 2019 (Fig. 2).

d. NWP data

Numerical weather prediction (NWP) model output fields are included in our NNs to characterize the environment of observed clouds. We use the 6-h forecast from the 6-hourly Climate Forecast System (CFS) 0.5° output (Saha et al. 2014) and match each set of CALIOP collocations by linearly

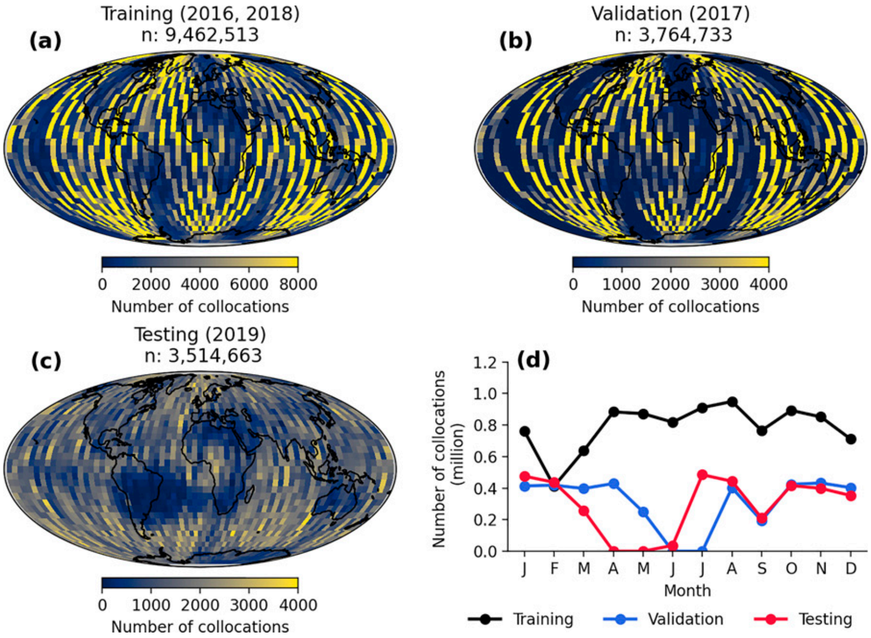

FIG. 1. The distributions of VIIRS collocations with CALIOP for the (a) training, (b) validation, and (c) testing datasets. Also shown are (d) the seasonal distributions of each.

TABLE 2. Central wavelengths of the infrared channels included in the ABI models.

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

interpolating in space and time. The fields used are the temperature and pressure at the surface and tropopause, total precipitable water, and the temperatures at pressure levels of 20, 100, 300, 500, 700, and 900 hPa.

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

3. Neural network training and validation

a. *Neural network details*

We use neural network with a quantile loss function that draws from Pfreundschuh et al. (2018), which demonstrated the ability of quantile regression NNs to estimate uncertainties for CTP. The quantile loss is shown in Eq. (1) where L is the loss for a prediction $\hat{y}$ for quantile τ and where y is the CALIOP CTP. When multiple quantiles are estimated, L can be averaged over multiple values of τ. The implications of Eq. (1) are that for larger quantiles overestimates are penalized more than underestimates (and the opposite for lower quantiles):

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

Each neural network has four fully connected layers consisting of 64, 32, 16, and 9 units. These values were determined by starting with the architecture used in Håkansson et al. (2018). We found a decrease in mean absolute error of 3.7 hPa (from 65.3 to 61.6 hPa) on the VIIRS validation dataset after roughly doubling the number of units used in Håkansson et al. (2018) and adding an additional layer. Further increases in the number units increased the computational expense but did not substantially improve performance. For example, an architecture with five layers of 128, 64, 32, 16, and 9 units decreased the mean absolute error only by 0.2 hPa, which is within the range of MAE values obtained by using different random weight initializations.

All layers except the last are followed by rectified linear unit (ReLU) activations. The last layer represents the nine evenly spaced quantiles we estimate (τ ranging from 0.1 to

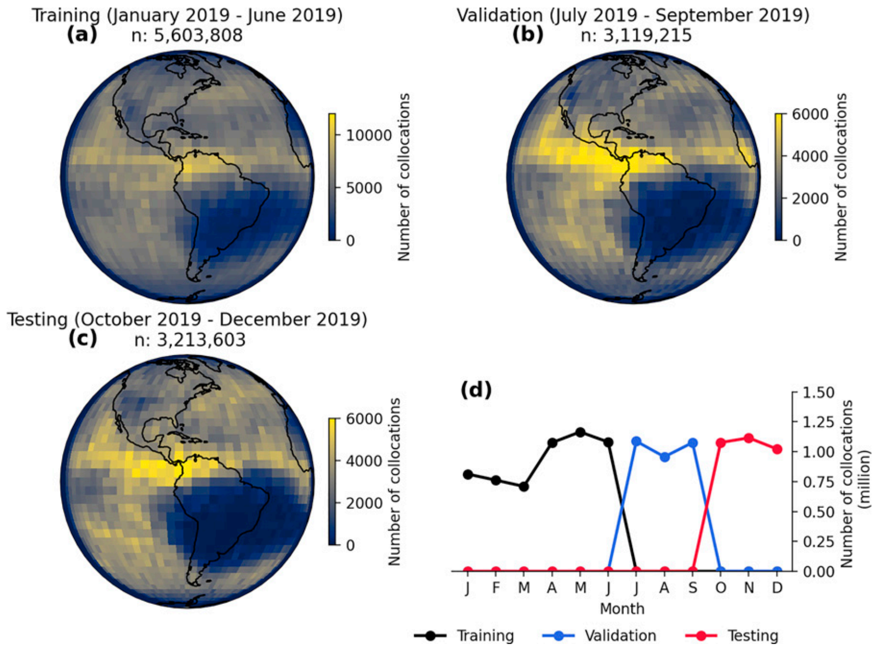

FIG. 2. As in Fig. 1, but for ABI collocations.

0.9 in increments of 0.1) and has no activation function. Predictions for CTP are obtained from the 50th quantile. A frequent problem in quantile regression is the crossing of quantiles since there is no mechanism to ensure that the curves do not overlap. In our datasets, we observe crossing quantiles in less than 2% of predictions that we judged to be small enough to ignore for our applications. Other works have suggested solutions for minimizing the crossing of quantiles (Cannon 2018).

The Adam optimizer (Kingma and Ba 2017) is used starting with a learning rate of 5×10^{-3}. This learning rate was chosen by testing values between 1×10^{-5} and 1 using a learning rate range test (Smith 2017). The batch size was increased from 250 used in Håkansson et al. (2018) to 5000 where the time taken for each epoch stopped decreasing. The loss is evaluated on the validation dataset after each epoch. The learning rate is reduced by a factor of 10 when the validation loss has not decreased for the last five epochs to a minimum of 1×10^{-6}. Training is stopped when the validation loss has not decreased for the previous nine epochs. These values are subjectively chosen on the basis of the number of epochs needed for the loss to stop decreasing after each learning rate reduction.

The inputs of each NN include the infrared channels specified in Table 1 for VIIRS and Table 2 for ABI. In addition to these values and the NWP information, we include several spatial metrics derived from a 5 × 5 pixel array surrounding the central pixel where the prediction is made. These spatial metrics include differences between the central pixel and both the coldest and warmest pixels and the standard deviation of all 25 pixels calculated for all channels. In total, the VIIRS NN includes 51 inputs, and the ABI NN includes 47 inputs. All inputs are standardized by subtracting the mean and dividing by the standard deviation calculated from the training dataset. The CALIOP observations of CTP are divided by 1000 hPa, meaning predicted CTP values typically lie between 0 and 1. Standardizing the CALIOP observations had no impact on performance, but consistently reduced training time by several epochs. Note that there are several hyperparameter decisions that we did not explicitly optimize for including the intermediate activation functions, the specific sets of estimated quantiles, the use of different learning rate schedules, and the early stopping criteria. It is possible that better results could be achieved if these decisions were included in a more formal hyperparameter search. All hyperparameter decisions are made using the mean absolute error of the estimate for the 50th quantile on the validation dataset.

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

b. *Neural network performance evaluation*

Because of our choice of the 1-km CALIOP CTP product, when we analyze with respect to optical depth, we can only compare instances in which the 1- and 5-km products have identified roughly the same cloud layer. Otherwise, one risks using the optical depth of a cloud to characterize the cloud-top pressure of another cloud lower in the atmosphere. Where we use optical depth, we limit the collocations to where the products agree on CTP within 150 hPa. For both imagers this removes the overall number of collocations by 16%. The optical depth and location of these collocations that are temporarily removed are shown in Fig. 3. Clouds with optical depths less than 0.5 are primarily affected with most of these removed profiles occurring in the tropics. Fewer than 2.5% of clouds with an optical depth near 1 are removed by this requirement.

The performance of the VIIRS NN is evaluated on our testing dataset in Fig. 4. The 99% confidence intervals for mean absolute error (MAE) and bias (Figs. 4a,b) are formed by 1000 bootstrapped samples of our testing dataset. A two-sided *t* test indicates significant differences (p values less than 0.001) between the NN and ACHA at all levels and COD ranges, except the difference between 1000 and 950 hPa at COD values between 3 and 30. MAE is, as expected, larger for clouds with low COD. ACHA appears to struggle with CTP estimation at the midlevels between 700 and 500 hPa. The MAE for the entire testing dataset is 58.1 hPa for the NN and 109.3 hPa for ACHA. The NN shows statistically significant improvement in most regions especially at the mid- and high latitudes (Figs. 4d,e).

Both approaches have issues with biases with respect to CALIOP in their predictions of CTP (Fig. 4b). The NN systematically fails to predict extreme values of CTP near the surface and places them too high in the atmosphere. The opposite problem occurs at the upper levels but is less exaggerated for clouds with high COD. Low COD clouds are most affected, with large positive biases above the 700-hPa level. ACHA has similar, but more extreme behavior, for clouds with low COD. ACHA has different signed biases as a function of CTP for clouds with COD greater than 1. This results in ACHA placing these clouds between 600 and 900 hPa too low in the atmosphere, and clouds with COD greater than 3 between 600 and 300 hPa too high in the atmosphere. This results in a tendency for ACHA to predict a lower frequency of clouds in the midlevels and could be a contributor to the larger MAE at these levels. In terms of location, the bias patterns are similar between the NN and ACHA (Figs. 4f,g), with a negative bias at the low latitudes and a positive bias at higher latitudes, but ACHA's mean bias is typically of larger magnitude.

A similar analysis is done for the ABI NN (Fig. 5). The evaluation of the ABI NN shares many characteristics with that of the VIIRS NN. One difference is that optically thin clouds at the highest levels (<150 hPa) have larger errors than does the VIIRS NN. Similar issues with the biases occur for ABI with a larger positive bias for optically thin clouds at the upper levels. The spatial patterns of MAE are similar to those of VIIRS. The spatial pattern of the mean bias differs greatly, as the ABI NN typically has a positive mean bias regardless of location. The MAE for all ABI and CALIOP collocations is 61.6 hPa. While this is similar to VIIRS, the two MAE values are not directly comparable because of the differences in the areas and meteorological conditions viewed.

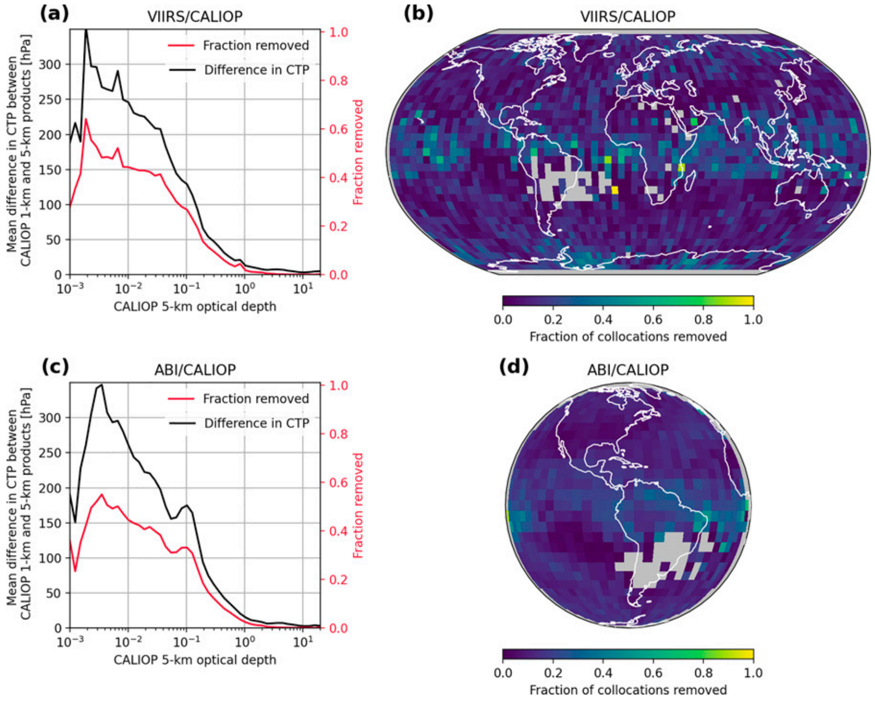

FIG. 3. For (top) VIIRS with CALIOP and (bottom) ABI with CALIOP: (a),(c) The fraction (red lines) of collocations that are removed by applying the requirement that the 1-km and 5-km CALIOP products agree within 150 hPa. Also shown are the mean differences (black lines) between the two products as a function of optical depth. (b),(d) The fraction of collocations removed on a 5° × 5° latitude/longitude grid.

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

c. Prediction uncertainty assessment

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

Figures 6a and 6d show a probability integral transform (PIT; Dawid 1984) histogram that indicates the frequency of observations as a function of the predictive CDF. The PIT histogram and can be used to assess the calibration of the predicted distributions and is created by calculating the quantile that the observation from CALIOP attains using the CDF predicted by each neural network. A perfectly calibrated model has a uniform frequency of observations across the CDF. An overconfident model has higher proportion of observations occurring near 0 and 1 (indicating that the predicted distributions are too narrow), and an underconfident model has a higher proportion of observations occurring near a value of 0.5 in the CDF (predicted distributions are too wide). The VIIRS NN appears to estimate CDFs that are too narrow evidenced by the higher frequencies at the tails and the lower frequencies at the middle of the CDFs. The ABI NN appears well calibrated with only small differences in observed frequencies throughout. Figures 6b and 6e show similar information presented differently and again confirm that the VIIRS NN predicts distributions that are slightly narrow, but the ABI distributions accurately capture the range of observed values. Figures 6c and 6f illustrate how the width of the predicted distribution (illustrated by the standard deviation of predicted quantiles) corresponds to a wider range of errors observed when comparing with CALIOP. Altogether, Fig. 6 shows that the predicted distributions from each of these neural networks are typically well calibrated and correspond to the observed errors in an intuitive manner.

4. Explainability assessment

Many methods for explaining predictions from ML models have been proposed. Some of these approaches offer the ability to provide local explanations, which attempt to describe how individual features contribute to a single model prediction. This is in contrast to global explanations, which are computed over a

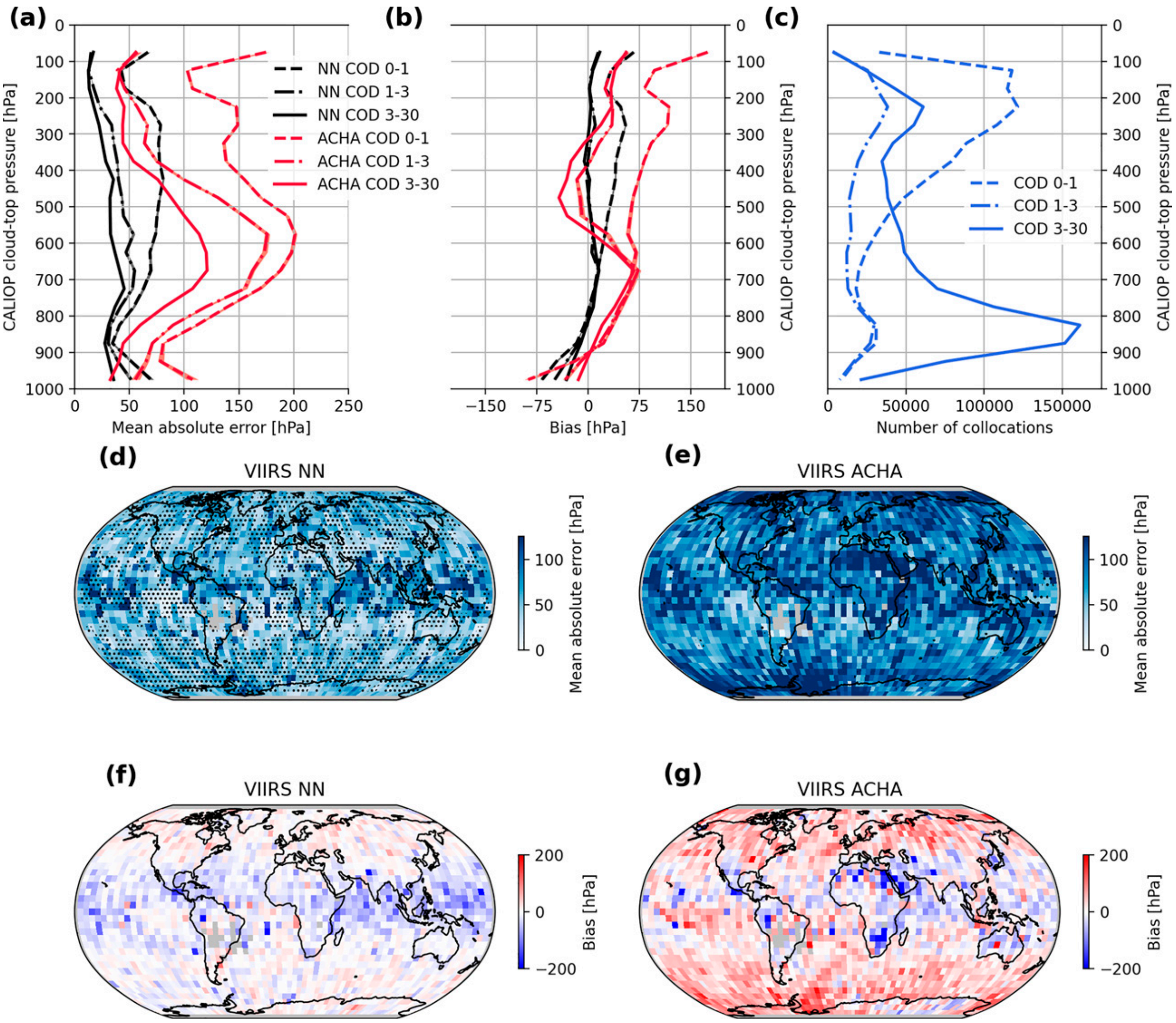

FIG. 4. (a) The MAE of the NN (black) and ACHA (red) for several values of COD. (b) The bias of the NN and ACHA relative to CALIOP over the same values of COD. In (a) and (b), 99% confidence intervals are in lighter shading but are often obscured by the mean values because of the narrow intervals. (c) The number of collocations occurring between CALIOP and VIIRS. (d),(e) The MAE on a 5° × 5° latitude/longitude grid. Stippling in (d) and (e) indicates that the respective approach has an improvement over the other that is statistically significant with a p value less than 0.001 at the grid point. (f),(g) The mean bias on the same grid.

set of predictions. We attempt to give different perspectives on feature importance using several methods for explaining NN CTP models.

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

a. *Sequential backward selection*

Sequential backward selection (SBS) is commonly used to find reduced feature sets with minimal reduction in model

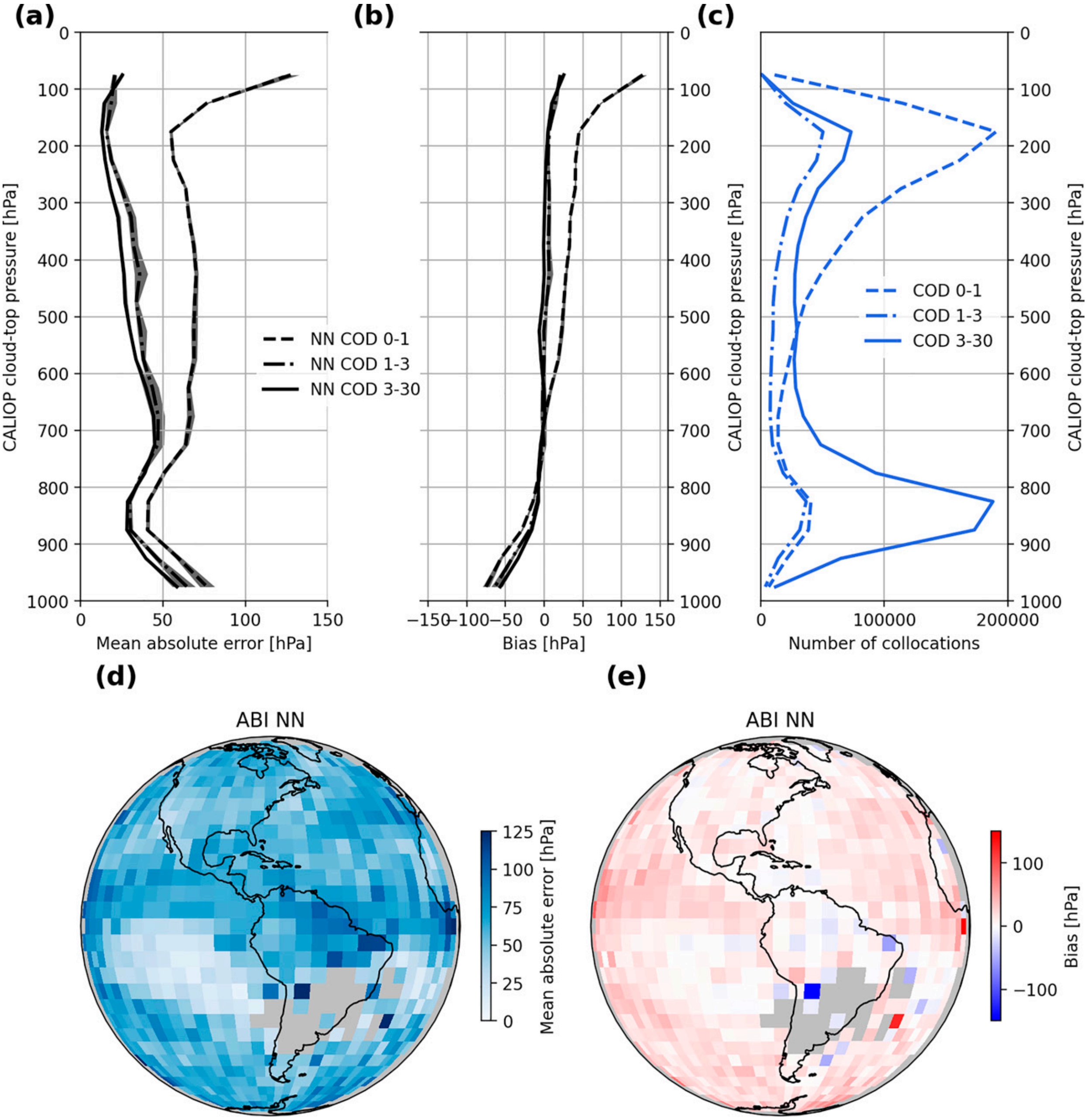

FIG. 5. (a) The MAE for the NN over several ranges of COD. (b) The bias over the same ranges of COD. In (a) and (b), 99% confidence intervals are in lighter shading but are often obscured by the mean values because of the narrow intervals. (c) The number of collocations between ABI and CALIOP. Also shown are the (d) MAE and (e) bias of the neural network on a 5° × 5° latitude/longitude grid.

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

To isolate the value of a given channel's information, we perform a full SBS, iterating over conceptually linked groups of features associated with each channel (Figs. 7 and 8). Removing

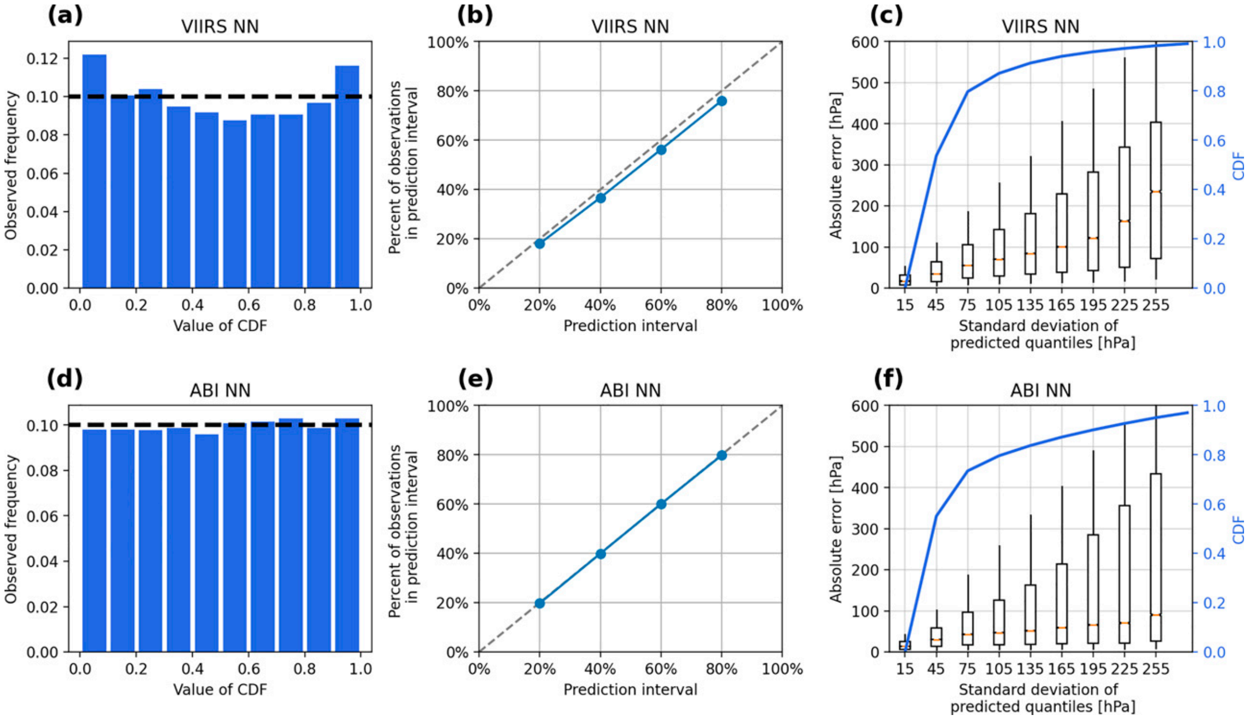

FIG. 6. For (top) VIIRS and (bottom) ABI: (a),(d) The observed frequency of CALIOP observations as a function of the value of the predicted CDF from the predicted quantiles. (b),(e) The fraction of CALIOP observations that fall within the prediction intervals derived from the predicted quantiles. Dashed lines in (a), (b), (d), and (e) indicate a well-calibrated model. (c),(f) The distribution of absolute errors between the neural networks and CALIOP for several ranges of the standard deviations of the predicted quantiles that fall within 30-hPa bins specified on the *x* axis. The middle orange line represents the 50th percentile, box edges represent the 30th and 70th percentile, and the whiskers represent the 10th and 90th percentile of absolute error (left *y* axis) with respect to CALIOP. The cumulative distribution function is shown in blue and represented on the right *y* axis.

groups of features allows us to determine feature importance as a function of spectral band. Otherwise, a feature's relevant information for the estimation of CTP could also be found in other features from the same channel. This also allows us to quantify the contribution of NWP model output to the NN's performance. Satellite estimates are often useful because they are based on observations (as compared with an NWP model forecast). Thus, quantifying the contribution of NWP and comparing it with that of the actual observations could be one way of determining how useful or reliable a given estimate of CTP is, in addition to the uncertainty estimates provided by the neural network. However, it is worth noting that the NWP group contains a larger number of features than groups associated with each channel.

The feature group SBS analysis shows many intuitive characteristics of CTP estimation. For VIIRS (Fig. 7), these results imply that the 8.6-μm channel is the most important channel followed by 10.8 and the 12.0. The 8.6-μm channel, in conjunction with window channels such as 10.8 μm, could be used to identify cloud phase (Strabala et al. 1994) and place a cloud in the upper or lower portion of the troposphere. The most useful fusion channels appear to be the 6.7 and 7.3 μm, which contain information about water vapor absorption and might be useful for placing a cloud above or below the bulk of the water vapor in a given scene. The low increases in MAE of CO_2 fusion channels (13.3–14.2 μm) are surprising given that CO_2 slicing has proven a useful approach for CTP estimation.

In most cases, a model's reliance on an individual channel increases when the number of channels decrease. However, there are a few exceptions to this generalization for VIIRS, including the impact of removing information from 8.6–10.8-μm channels once the 13.9- and 14.2-μm channels are removed (Fig. 7, rounds 3–5). NWP information ranks highly in the first few rounds, and as channels are removed, we see an increasing reliance on NWP information. This indicates that the usage of NWP information changes as a function of the channels included in an NN.

The same analysis indicates a few similarities for ABI (Fig. 8). The 8.4-, 10.3-, and 12.3-μm channels all have relatively large increase in MAE when tested in the first several rounds. Unlike VIIRS, the 13.3-μm channel ranks fairly high. On ABI, the 11.2-μm channel, ozone channel (9.6 μm), and strongly absorbing water vapor channels (6.2–7.3 μm) do not benefit the model strictly in terms of MAE.

In the first round, NWP information appears to be more essential for accurate predictions from ABI relative to VIIRS. This impact becomes more similar after the CO_2 channels (with wavelengths 13.3 μm and above) are removed from the VIIRS models. VIIRS appears to rely more heavily on

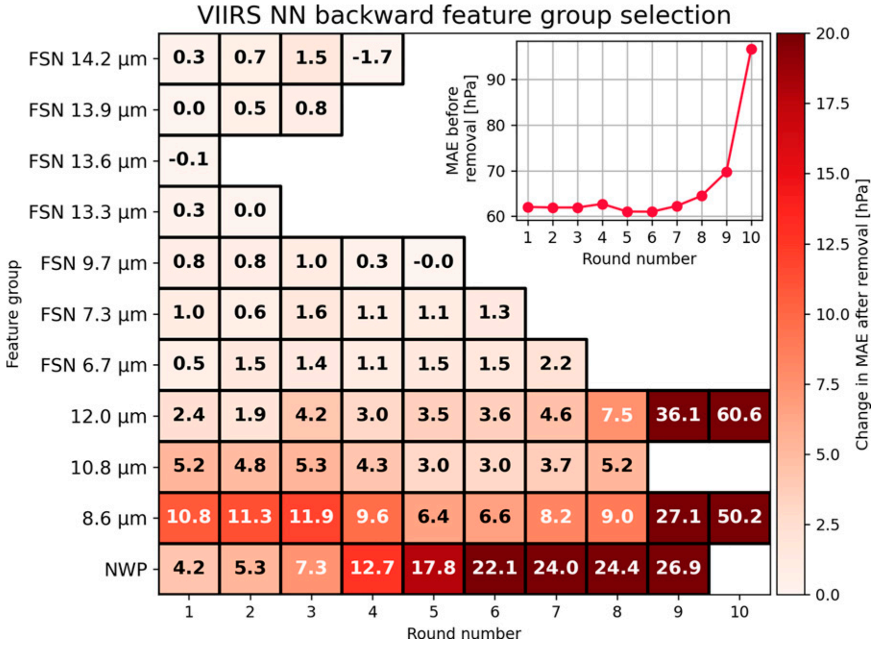

FIG. 7. The results of a backward selection performed on features linked to each channel used in the VIIRS CTP models. This figure is most easily interpreted by considering each column from left to right. Each column represents a single round of backward selection. The inset plot shows the MAE of a model that includes all remaining features present in a given column. In each round, a feature's impact is tested by training five identical but randomly initialized models without that feature and recording the MAE. The value in each box represents the mean increase in MAE (of the three best-performing models) relative to a model that includes all features present in a column. Note that the feature group that increases MAE the least in a given round is permanently removed from the model and is no longer tested in the following rounds.

information from the 8.6-μm channel. The impact of removing the lone CO_2 channel on ABI (13.3 μm) is different than the impact of removing the four fusion CO_2 channels for VIIRS. For instance, in round 7 after the 13.3-μm channel is removed, the MAE from removing the 12.3- or 8.4-μm channel is nearly tripled on ABI. After removing the fusion CO_2 channels from VIIRS, it is mostly the impact of NWP information, which is increased.

Figures 7 and 8 make it clear that similar performance can be achieved for these CTP models with reduced feature sets. Ignoring differences in reliance on NWP information, similar models could be created using the feature sets after round 5 for both instruments. We continue to use the full feature set to keep the latter experiments consistent.

b. *Neural network explanation methods*

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

The first method used is layerwise relevance propagation (LRP; Bach et al. 2015). LRP is a popular method for model attribution and has been used to explain models in applications such as radar reflectivity estimation from satellite imagers (Hilburn et al. 2021) and for detecting common change patterns among climate models (Barnes et al. 2020). LRP can be generally described as computing a backward pass through an NN, starting with the activations at the last layer. A prediction score is propagated backward through each layer of the model and projected onto the dimensions of the original input at the first layer. There are several different propagation rules that dictate how the prediction score is distributed to the units of each layer. In our application, we use the epsilon rule for all layers, which adds a small positive value to the denominator of the relevance propagation rule to improve numerical stability.

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

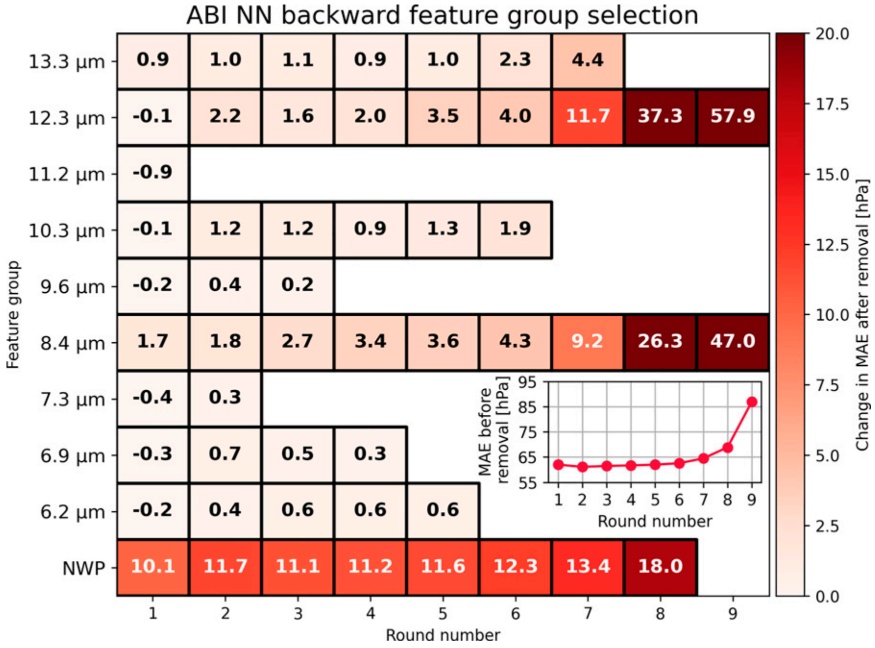

FIG. 8. As in Fig. 7, but for the ABI CTP models.

Deep SHAP takes advantage of the per-node attribution rules used in DeepLIFT.

These methods were selected on the basis of their recent popularity in the atmospheric sciences and broader machine learning communities and the relative ease-of-use of open-source implementations. SHAP has several advantages over LRP including producing featurewise additive explanations that sum to the value of the model's prediction. SHAP also uses a dataset-dependent sample as reference values for producing explanations. Both LRP and SHAP can produce signed attributions, according to whether an input feature acted to increase or decrease the prediction. We note that SHAP explanations took several orders of magnitude longer than LRP to produce feature attributions, which likely prevents the possibility of using SHAP to explain all predictions if these neural networks were implemented in a near-real-time application for VIIRS or ABI. LRP, on the other hand, could produce explanations in roughly the same amount of time it takes to make a prediction. Mamalakis et al. (2022) provides a more thorough comparison of LRP, SHAP, and many other NN explanation approaches,

In itself, interpreting the output from local explanation methods can be a difficult task. We attempt to simplify this by standardizing the local explanations. We take the absolute value of the LRP relevance and SHAP values and express them relative to the feature with the greatest value for each input example. For each prediction, each feature has relative importance ranging from 0, which implies it was not important, to 1, which implies the feature was the most important or tied with the most important.

Figures 9 and 10 show the global relative feature importance calculated over conceptually linked groups of features. These values are calculated by summing the absolute value of the LRP and SHAP attributions for each group of features and dividing by the value from the largest group. The 8.4-μm (ABI) and 8.6-μm (VIIRS) channels are suggested to be the most important channels for CTP estimation. LRP and SHAP assign low relative feature importance to the ozone channels around 9.7 μm on both instruments and the 6.7- and 7.3-μm channels on VIIRS. Both methods rank spatial information lower than spectral information for both imagers. Both methods also rank spatial metrics from fusion channels lower than those from native features, despite there being more than double the number of features from the fusion channels.

Despite the agreement on some broad points, there are a few differences between LRP and SHAP. In general, the relative feature importance values from LRP are more distributed across features relative to SHAP, which gives sparser explanations that emphasize the most important features. The high rankings of the 8.4- and 8.6-μm channels from SHAP imply that most explanations are dominated by these channels. Both methods agree on the relative ranking of most features, with the significant exception of NWP data for both sensors, where LRP reports values that are more than 2 times that of SHAP. SHAP's attribution here also contrasts with the backward selection results that imply that NWP information is very useful for CTP estimation.

There are several other differences between what information these methods suggest that the NNs use in comparison with backward selection. LRP reports that the fusion channels, ignoring spatial metrics, have a roughly similar value of relative feature importance as the VIIRS native channels. However, when we remove all fusion channels from the VIIRS models, MAE only increases by less than 5 hPa, which

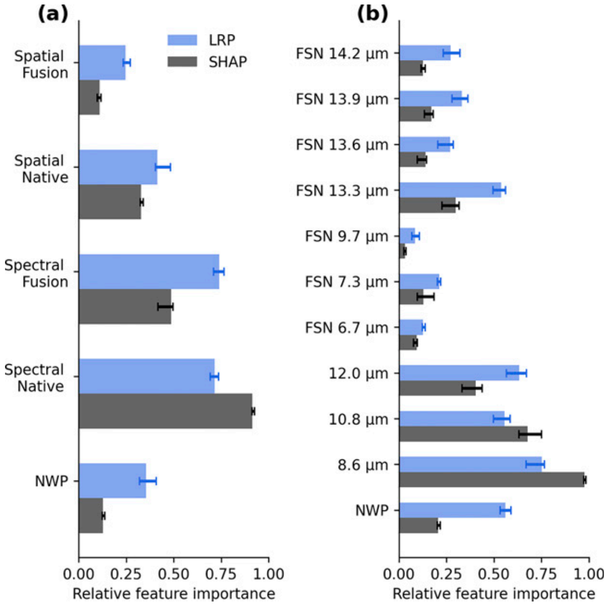

FIG. 9. Relative feature importance for different groups of features calculated over five VIIRS NN models. (a) Channel brightness temperatures are separated from their associated spatial metrics and fusion channels from native VIIRS channels. (b) Features are separated based on their associated channel. Error bars are computed from the maximum and minimum relative feature importance of LRP and SHAP values computed for five models with different weight initializations.

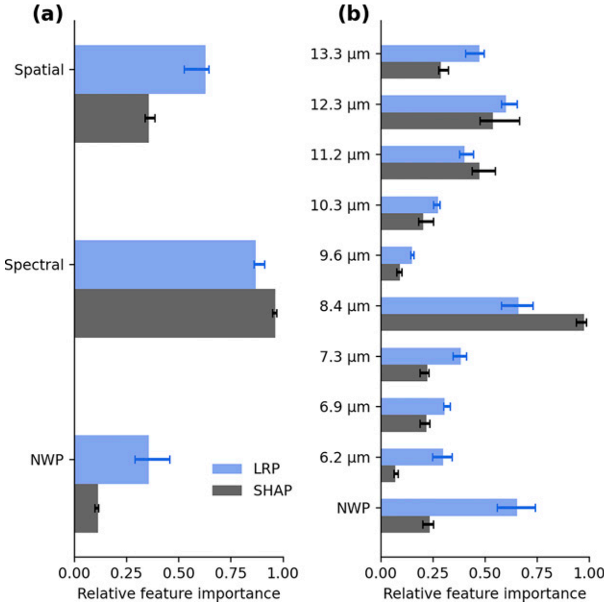

FIG. 10. As in Fig. 9, but for ABI NN models.

is less than removing information from the 8.6- or 10.8-μm channels. Of all fusion channels, the 6.7- and 7.3-μm channels appear to the have the largest impact when removed during backward selection but are ranked fairly low when compared with 13.3 μm, which both LRP and SHAP rank as the most important fusion channel.

c. *Local explanation clustering*

Next, we explore the local explanations for these models. We attempt to find conceptually similar explanations among the local attributions. We then analyze these explanations as a function of their dominant features, CTP, cloud-top phase, opacity, and location. We find these explanations by using a k-means clustering (from scikit-learn, version 0.24; Pedregosa et al. 2011) on the local attributions. Thus, each imager–CALIOP collocation belongs to a specific cluster. For concision, we only perform the following analysis using LRP. We specify four clusters, but do not conclude that it is the optimal number, nor that there are discrete clusters at all. We use the clustering to partition the local explanations into more homogeneous groups and visualize differences among them. The motivation for this analysis is to help understand the kinds of relationships that might be used in the neural networks in predicting CTP for different types of clouds. Figure 11 shows the clustering analysis performed for VIIRS and describes each cluster in terms of the relative feature importance for predictions that belong to each cluster. Figure 11 also illustrates the fraction of all collocations belonging to each cluster in terms of CTP, cloud-top phase, opacity, and location.

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

Figure 12 illustrates a similar analysis for ABI. Overall, the clusters appear to be less sensitive to opacity and more sensitive to cloud-top pressure. Some similar patterns exist in the spatial distribution of the clusters when comparing ABI with VIIRS.

ABI cluster 1 shows importance in channels with water vapor absorption (6.2, 6.9, and 7.2 μm), many spatial metrics, and NWP

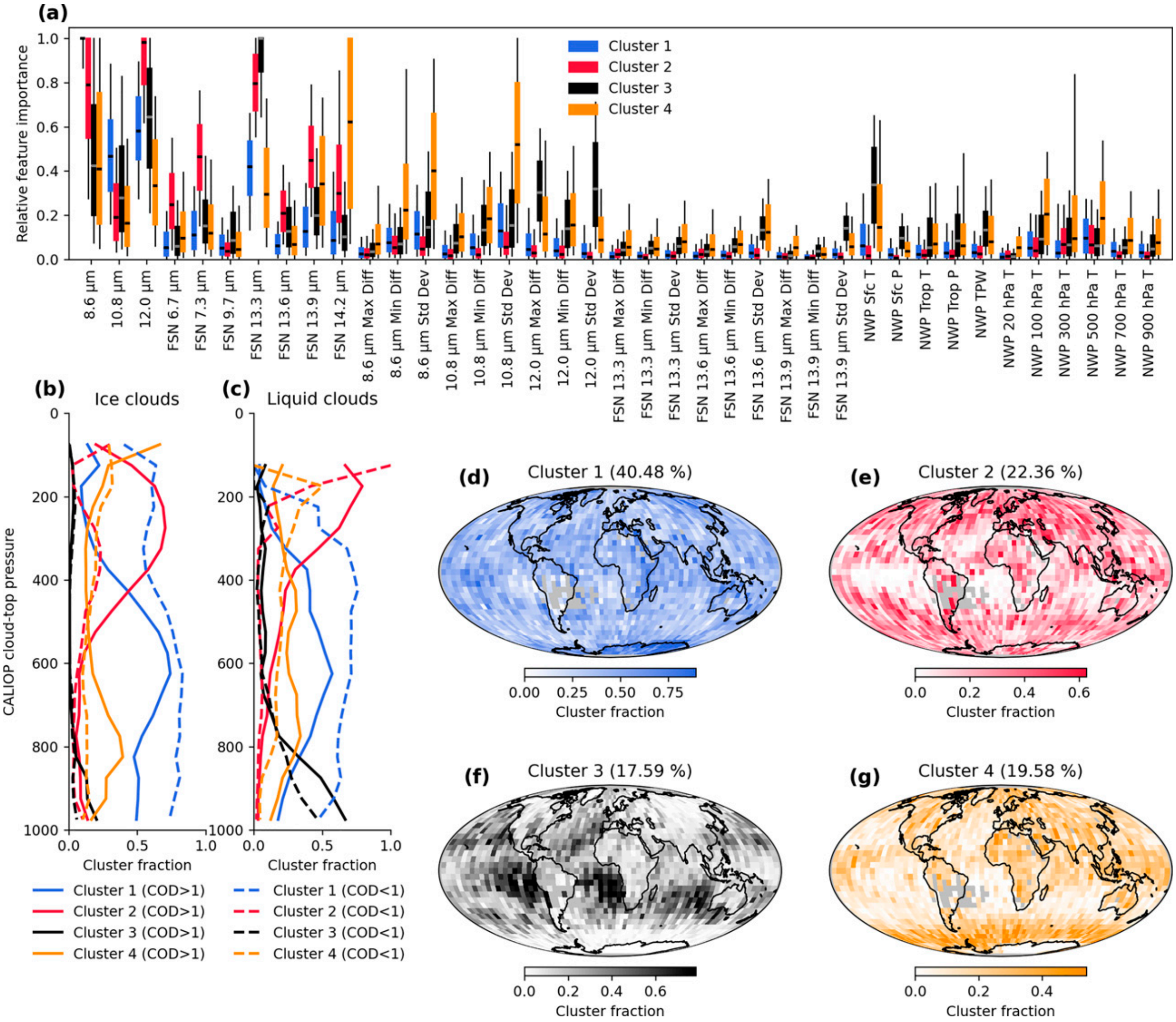

FIG. 11. Cluster analysis of relative feature importance values calculated from LRP for the VIIRS CTP models. (a) The distribution of feature importance values for each cluster where the black middle line, box edges, and whiskers represent the 50th, 30th/70th, and 10th/90th quantiles of each feature. Also shown are the distributions of each cluster with respect to CTP and optical depth of the uppermost cloud for (b) ice clouds and (c) liquid clouds. (d),(e),(f),(g) The spatial distribution of each cluster on a regular 5° grid and the proportion of collocations falling within each cluster, listed above the maps. Note that the color bars represent slightly different ranges and are chosen to emphasize spatial variability within each cluster. The analysis in this figure is subject to the requirement described in Fig. 3 because of the use of optical depth. Some fusion-channel spatial metrics are not shown in (a) because of very low values and to ease visualization.

information. Cluster 1 primarily represents clouds at all levels but slightly favors low-level opaque ice clouds. ABI cluster 2 explains a large fraction of low-level liquid clouds and relies heavily on the 12.3-μm channel where it is the leading feature for over 90% of examples. ABI cluster 2 is frequent in areas with low-level cloud cover. ABI cluster 3 has the largest feature importance in the 8.4-μm channel, where it is frequently the leading feature. It is present at various levels, slightly favoring optically thin liquid clouds and mostly occurs at the tropics. ABI cluster 4 has high feature importance in the 6.2-, 6.9-, 7.3-, and 8.4-μm channels and an otherwise low importance in spatial metrics and NWP data aside from 300 hPa temperatures. It describes the vast majority of predictions for ice clouds between 600 and 200 hPa and is primarily located at the high latitudes.

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

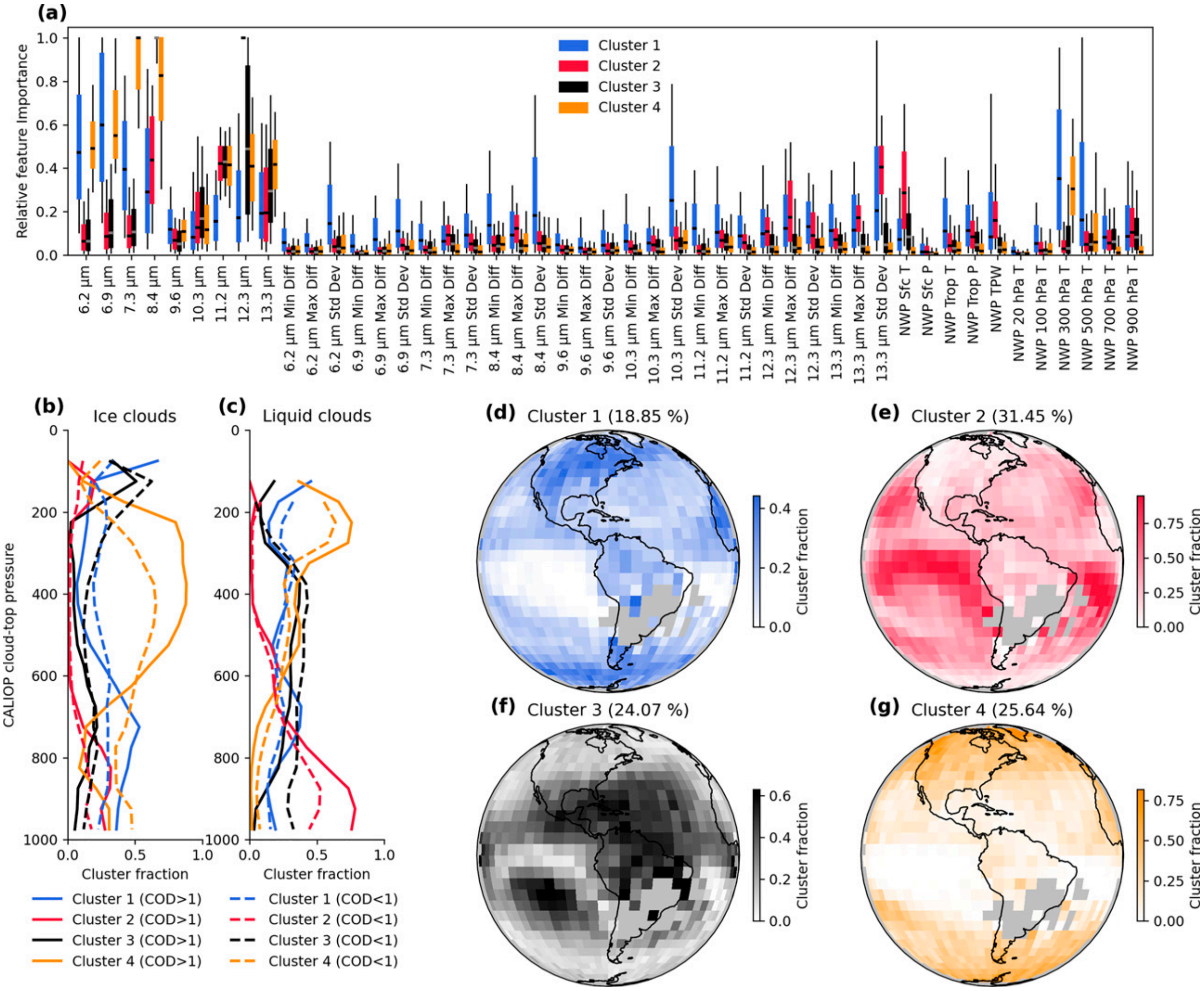

FIG. 12. As in Fig. 11, but for the ABI CTP models.

similarity can be found between VIIRS cluster 1 and ABI cluster 3, which have high importance in the 8.4- and 8.6-μm channels but have very different spatial distributions.

d. *Local explanation example*

In an effort to contextualize the LRP attributions and illustrate potential relationships between VIIRS cluster 4 and cloud edges, we calculate relative feature importance from LRP for an example VIIRS scene centered over 55°S, 100°E (Fig. 13). The LRP values are standardized in the same way and are reported as a function of the same conceptual groups in Fig. 9a. The neural network is not capable of cloud detection, so predictions are provided in all pixels regardless of whether there is a cloud present.

Shown in Figs. 13a and 13b is 10.8-μm channel from VIIRS and the predictions of CTP made by the NN. The width of the predicted distribution can be large near edges of clouds that have high contrast with the surface (Fig. 13c). The predicted distributions are also wider where upper-level clouds overlap with midlevel clouds (Fig. 13c, lower left). In this scene, NWP information is most important for mid- and lower-level clouds. Native spectral observations (Fig. 13e) are most important for upper and lower-level clouds, but there is a strong decrease in the importance of native spectral observations near clouds edges (Fig. 13e, right). These low values of the relative feature importance of native spectral observations near cloud edges correspond to large importance of the spatial metrics from native VIIRS observations (Fig. 13g). The relative feature importance for spectral fusion feature group is largest for lower- and midlevel clouds and has more moderate values for upper-level clouds. The feature importance of the spatial metrics from fusion channels (Fig. 13h) appear to have the lowest values in this scene overall and only have moderate impact for more spatially uniform low-level clouds and very low importance for upper-level clouds.

We see a few potential explanations for the importance of spatial metrics around cloud edges. At cloud edges, spectral features can be difficult to interpret because of the possibility of a pixel being only partially cloudy and the resulting

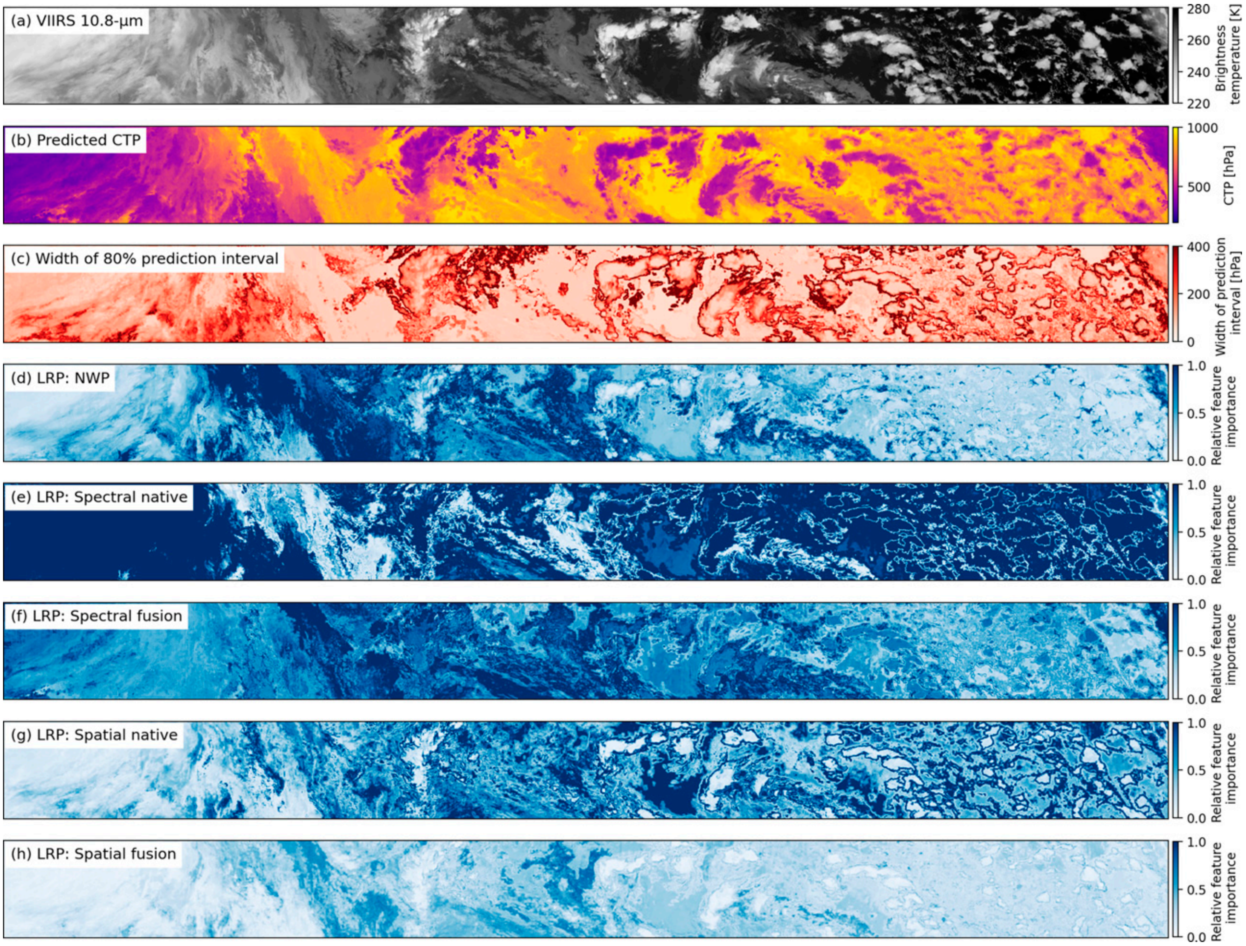

FIG. 13. Example of CTP predictions for the VIIRS scene centered over 55°S, 100°E. (a) The 10.8-μm infrared channel. (b) The estimates of CTP from the 50th quantile. (c) The width of the 80% prediction interval constructed from the 10th and 90th quantiles. Also shown is the LRP relative feature importance for the (d) NWP, (e) spectral native, (f) spectral fusion, (g) spatial native, and (h) spatial fusion groups discussed in the text.

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

5. Discussion

The efforts to explain the models in this work give slightly different perspectives on feature importance for NN CTP estimation. In many cases, this is expected when comparing SBS with LRP and SHAP. Observing a small increase in error when removing a set of features does not necessarily imply that they are not useful for CTP estimation. It instead might be an indicator that the information is not unique to a feature. This may be the case for the fusion CO_2 channels in the VIIRS model. When removed through backward selection (Fig. 7), they yield only small increases in model error; however, the LRP and SHAP attribute a moderate amount of feature importance to them. This indicates that while fusion CO_2 channels may be useful for CTP estimation, similar performance, in terms of MAE, can be attained without them. However, when comparing SBS with the LRP and SHAP, it is important to note that SBS ultimately describes a different set of models. The backward selection performed involves fitting models with access to fewer and fewer features. Thus, the rankings of feature importance in the latter rounds may become less consistent with those from LRP and SHAP due to this.

Other differences are less easily explained, such as the differences in importance of NWP information between LRP and SHAP. LRP assigns a large relative FI to NWP information and is in agreement with the first round of the SBS feature group analysis for both instruments, indicating that this might be a failure of SHAP's attribution. A few potential sources of the differences between LRP and SHAP could be rooted in our choice of model architecture, the nature of our prediction task, and choice of LRP rules. LRP was initially developed to explain the output of convolutional neural networks trained for image classification (Bach et al. 2015). It is unclear how well these attributions generalize to regression problems or nonconvolutional neural networks like ours. Similarly, during development we noticed slight differences in attributions depending on the exact LRP propagation rules used (Montavon et al. 2019), but qualitatively similar takeaways overall (not shown).

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

Several methods also agree on the relative unimportance of particular features. These include the ozone channels from both instruments, the 6.2-μm ABI band, which is sensitive to upper tropospheric water vapor, and the spatial metrics calculated from the VIIRS–CrIS fusion channels. In this case it is helpful to remember that the fusion channels are derived from the relatively coarse CrIS observations and interpolated using infrared channels from VIIRS. It is plausible that fine-scale spatial variability on the scale of 3.75 km (the edge length of 5 VIIRS pixels at nadir) is not well represented. Regardless of disagreement between LRP and SHAP, we can conclude that the VIIRS–CrIS fusion channels only have small benefit when included in the VIIRS NN since they increased MAE only by roughly 5 hPa (Fig. 7) when all are removed. However, several fusion channels indicated no benefit after removal (Fig. 7, rounds 1–5).

One point made earlier in this paper is that removing these channels had the effect of increasing the reliance on NWP information (Figs. 7 and 8). The SBS analysis shows that the fusion CO_2 channels do not substantially reduce model error when included. However, their inclusion is suggested to reduce the reliance on NWP information. This is an important point since it can change how much a given CTP prediction depends on observations, as compared with ancillary information from an NWP model forecast. When included in climate records, changing the source of the ancillary NWP data can yield small but meaningful changes in the variability of cloud properties estimated from imagers (Foster et al. 2016). Thus, the physical interpretation CTP estimates can change depending on the reliance on NWP information.

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

It is interesting to compare results of this analysis with the physical information exploited in approaches like CO_2 slicing. CO_2 channels do not seem to add much value to a model that already has access to infrared window channels including a channel around 8.6 μm. There is more value in including channels with lower- and midlevel water vapor absorption, such as the 6.7 and 7.3 μm. However, it is not clear if this observation holds for imagers, such as MODIS, which have native channels in these spectral regions, rather than inferred channels from sounder observations in the case of our VIIRS neural network. Despite this caveat, most of the feature importance metrics used in this analysis imply that not exploiting variability of the 8.6-μm or infrared window channels between 10 and 12 μm will yield a suboptimal result.

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

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

As previously mentioned in section 4, much of the difficulty in interpreting the results of this analysis comes from use of correlated features. One approach we experimented with in order to ease the interpretation of this analysis was to add L1 and L2 regularization to the parameters of each layer in our models. L1 regularization acts to penalize the magnitude of the parameters of the model and L2 regularization penalizes the squared magnitude of the parameters. Our hypothesis was that by adding this kind of regularization, the models would be incentivized to rely on a smaller subset of features and ease the interpretation of the results. However, we found that regularization with small L1 and L2 penalties had almost no impact on the relative feature importance from LRP and

SHAP. Larger L1 and L2 penalties decreased the performance of these models to unacceptable levels and increased the impact of different random weight initializations. Thus, we have not included these models as a part of the analysis.

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

6. Conclusions

We characterize the use of individual channels of LEO and GEO imagers for NN CTP estimation. We first perform a short comparison between our NNs and an operational approach that demonstrates large improvement in CTP estimation with respect to CALIOP. We then use backward selection, LRP, and SHAP to infer the relative importance of features. We find many instances of disagreement between these different perspectives on feature importance, but broad agreement on the importance of a few channels, including the VIIRS 8.4-μm channel (8.6 μm for ABI) and other infrared window channels around 10–12 μm. We also observe a small benefit in including absorption channels that are sensitive to midlevel and lower-level water vapor. VIIRS–CrIS fusion CO_2 channels and spatial metrics derived from them appear add little to no improvement to CTP models where native infrared channels are already present but have impact on the reliance of a given model on NWP model output. Clustering local explanations from LRP illustrates how NN models can exploit variability related to CTP, phase, and opacity from infrared measurements. The LRP clustering also suggests, intuitively, that the NNs use different infrared channel combinations for estimating CTP of the different cloud types. While this analysis reveals several interesting aspects of the relative importance of infrared channels for CTP estimation, this work also illustrates the immense challenge of attributing physical significance to both global and local explanations for neural networks.

Acknowledgments. This research has been supported by the National Oceanic and Atmospheric Administration (NA15NES4320001). The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. government position, policy, or decision.

Data availability statement. All VIIRS data used in this study are freely available from the NASA VIIRS atmosphere Science Investigator-Led Processing System (SIPS; https://sips.ssec.wisc.edu/). ABI data can be obtained from the NOAA Comprehensive Large Array-Data Stewardship System (CLASS; https://class.noaa.gov). CFS data can be obtained from the National Centers for Environmental Information (NCEI; https://www.ncei.noaa.gov/products/weather-climate-models/climate-forecast-system).

REFERENCES

Abadi, M., and Coauthors, 2016: TensorFlow: A system for large-scale machine learning. *Proc. 12th USENIX Symp. on Operating Systems Design and Implementation*, Savannah, GA, USENIX, 265–283, https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Alin, A., 2010: Multicollinearity. *Wiley Interdiscip. Rev.: Comput. Stat.*, **2**, 370–374, https://doi.org/10.1002/wics.84.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 2015: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10**, e0130140, https://doi.org/10.1371/journal.pone.0130140.

Barnes, E. A., B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2020: Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002195, https://doi.org/10.1029/2020MS002195.

Braun, B. M., T. H. Sweetser, C. Graham, and J. Bartsch, 2019: CloudSat's A-Train exit and the formation of the C-Train: An orbital dynamics perspective. *2019 IEEE Aerospace Conf.*, Big Sky, MT, IEEE, 1–10, https://doi.org/10.1109/AERO.2019.8741958.

Cannon, A. J., 2018: Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environ. Res. Risk Assess.*, **32**, 3207–3225, https://doi.org/10.1007/s00477-018-1573-6.

Chahine, M. T., 1974: Remote sounding of cloudy atmospheres. I. The single cloud layer. *J. Atmos. Sci.*, **31**, 233–243, https://doi.org/10.1175/1520-0469(1974)031<0233:RSOCAI>2.0.CO;2.

Chase, R. J., S. W. Nesbitt, and G. M. McFarquhar, 2021: A dual-frequency radar retrieval of two parameters of the snowfall particle size distribution using a neural network. *J. Appl. Meteor. Climatol.*, **60**, 341–359, https://doi.org/10.1175/JAMC-D-20-0177.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020: A deep-learning model for automated detection of intense midlatitude convection using geostationary satellite images. *Wea. Forecasting*, **35**, 2567–2588, https://doi.org/10.1175/WAF-D-20-0028.1.

Daniels, J., W. Bresky, S. Wanzong, C. Velden, and H. Berger, 2012: GOES-R Advanced Baseline Imager (ABI) algorithm theoretical basis document for derived motion winds—Version 2.5. NOAA/NESDIS/STAR Tech. Rep., 98 pp., https://www.star.nesdis.noaa.gov/goesr/docs/ATBD/DMW.pdf.

Daoud, J. I., 2017: Multicollinearity and regression analysis. *J. Phys.*, **949**, 012009, https://doi.org/10.1088/1742-6596/949/1/012009.

Dawid, A. P., 1984: Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Roy. Stat. Soc.*, **147A**, 278–292, https://doi.org/10.2307/2981683.

Dormann, C. F., and Coauthors, 2013: Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46, https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Farrar, D. E., and R. R. Glauber, 1967: Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.*, **49**, 92–107, https://doi.org/10.2307/1937887.

Foster, M. J., and A. Heidinger, 2014: Entering the era of +30-year satellite cloud climatologies: A North American case study. *J. Climate*, **27**, 6687–6697, https://doi.org/10.1175/JCLI-D-14-00068.1.

——, ——, M. Hiley, S. Wanzong, A. Walther, and D. Botambekov, 2016: PATMOS-x cloud climate record trend sensitivity to reanalysis products. *Remote Sens.*, **8**, 424, https://doi.org/10.3390/rs8050424.

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

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

Heidinger, A. K., and M. J. Pavolonis, 2009: Gazing at cirrus clouds for 25 years through a split window. Part I: Methodology. *J. Appl. Meteor. Climatol.*, **48**, 1100–1116, https://doi.org/10.1175/2008JAMC1882.1.

——, and Y. Li, 2017: AWG cloud height algorithm (ACHA)—Version 3.1. NOAA/NESDIS/STAR Tech. Rep., 59 pp., https://www.star.nesdis.noaa.gov/goesr/documents/ATBDs/Enterprise/ATBD_Enterprise_Cloud_Height_v3.1_Mar2017.pdf.

——, and Coauthors, 2019: Using sounder data to improve cirrus cloud height estimation from satellite imagers. *J. Atmos. Oceanic Technol.*, **36**, 1331–1342, https://doi.org/10.1175/JTECH-D-18-0079.1.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2021: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using GOES-R satellite observations. *J. Appl. Meteor. Climatol.*, **60**, 3–21, https://doi.org/10.1175/JAMC-D-20-0084.1.

Holz, R., S. Ackerman, F. Nagle, R. Frey, S. Dutcher, R. Kuehn, M. Vaughan, and B. Baum, 2008: Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP. *J. Geophys. Res.*, **113**, D00A19, https://doi.org/10.1029/2008JD009837.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Inoue, T., 1985: On the temperature and effective emissivity determination of semi-transparent cirrus clouds by bi-spectral measurements in the 10μm window region. *J. Meteor. Soc. Japan*, **63**, 88–99, https://doi.org/10.2151/jmsj1965.63.1_88.

Jakob, C., and G. Tselioudis, 2003: Objective identification of cloud regimes in the tropical western pacific. *Geophys. Res. Lett.*, **30**, 2082, https://doi.org/10.1029/2003GL018367.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://arxiv.org/abs/1412.6980v9.

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

Lagerquist, R., A. McGovern, and D. J. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, **34**, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1.

Li, Y., B. A. Baum, A. K. Heidinger, W. P. Menzel, and E. Weisz, 2020: Improvement in cloud retrievals from VIIRS through the use of infrared absorption channels constructed from VIIRS+CrIS data fusion. *Atmos. Meas. Tech.*, **13**, 4035–4049, https://doi.org/10.5194/amt-13-4035-2020.

Lundberg, S., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. arXiv, 1705.07874v2, https://arxiv.org/abs/1705.07874.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2021: Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. arXiv, 2103.10005v1, https://arxiv.org/abs/2103.10005v1.

——, E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. arXiv, 2202.03407v1, https://arxiv.org/abs/2202.03407.

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

Menzel, W. P., and Coauthors, 2008: MODIS global cloud-top pressure and amount estimation: Algorithm description and results. *J. Appl. Meteor. Climatol.*, **47**, 1175–1198, https://doi.org/10.1175/2007JAMC1705.1.

Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, 2019: Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer International Publishing, 193–209, https://doi.org/10.1007/978-3-030-28954-6_10.

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

Oreopoulos, L., N. Cho, and D. Lee, 2017: Using MODIS cloud regimes to sort diagnostic signals of aerosol-cloud-precipitation interactions. *J. Geophys. Res. Atmos.*, **122**, 5416–5440, https://doi.org/10.1002/2016JD026120.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pfreundschuh, S., P. Eriksson, D. Duncan, B. Rydberg, N. Håkansson, and A. Thoss, 2018: A neural network approach to estimating a posteriori distributions of Bayesian retrieval problems. *Atmos. Meas. Tech.*, **11**, 4627–4643, https://doi.org/10.5194/amt-11-4627-2018.

Poulsen, C. A., and Coauthors, 2012: Cloud retrievals from satellite data using optimal estimation: Evaluation and application to ATSR. *Atmos. Meas. Tech.*, **5**, 1889–1910, https://doi.org/10.5194/amt-5-1889-2012.

Rodgers, C. D., 1976: Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys.*, **14**, 609–624, https://doi.org/10.1029/RG014i004p00609.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

Seaman, C. J., Y.-J. Noh, S. D. Miller, A. K. Heidinger, and D. T. Lindsey, 2017: Cloud-base height estimation from VIIRS. Part I: Operational algorithm validation against *CloudSat*. *J. Atmos. Oceanic Technol.*, **34**, 567–583, https://doi.org/10.1175/JTECH-D-16-0109.1.

Shapley, L. S., 1953: A value for *n*-person games. *Contributions to the Theory of Games*, Vol. II, Princeton University Press, 307–317.

Shrikumar, A., P. Greenside, A. Shcherbina, and A. Kundaje, 2017: Not just a black box: Learning important features through propagating activation differences. arXiv, 1605.01713v3, https://arxiv.org/abs/1605.01713v3.

Smith, L. N., 2017: Cyclical learning rates for training neural networks. *2017 IEEE Winter Conf. on Applications of Computer Vision*, Santa Rosa, CA, IEEE, 464–472, https://doi.org/10.1109/WACV.2017.58.

Smith, W., and C. Platt, 1978: Comparison of satellite-deduced cloud heights with indications from radiosonde and ground-based laser measurements. *J. Appl. Meteor.*, **17**, 1796–1802, https://doi.org/10.1175/1520-0450(1978)017<1796:COSDCH>2.0.CO;2.

Strabala, K. I., S. A. Ackerman, and W. P. Menzel, 1994: Cloud properties inferred from 8–12-μm data. *J. Appl. Meteor. Climatol.*, **33**, 212–229, https://doi.org/10.1175/1520-0450(1994)033<0212:CPIFD>2.0.CO;2.

Vaughan, M. A., and Coauthors, 2009: Fully automated detection of cloud and aerosol layers in the CALIPSO lidar measurements. *J. Atmos. Oceanic Technol.*, **26**, 2034–2050, https://doi.org/10.1175/2009JTECHA1228.1.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Watts, P. D., R. Bennartz, and F. Fell, 2011: Retrieval of two-layer cloud properties from multispectral observations using optimal estimation. *J. Geophys. Res.*, **116**, D16203, https://doi.org/10.1029/2011JD015883.

Weisz, E., B. A. Baum, and W. P. Menzel, 2017: Fusion of satellite-based imager and sounder data to construct supplementary high spatial resolution narrowband IR radiances. *J. Appl. Remote Sens.*, **11**, 034506, https://doi.org/10.1117/1.JRS.13.034506.

Wheeler, D., and M. Tiefelsdorf, 2005: Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J. Geogr. Syst.*, **7**, 161–187, https://doi.org/10.1007/s10109-005-0155-6.

White, C. H., 2022: Application of machine learning methods to imager cloud property estimation and the feasibility of their use in operations and climate data records. Ph.D. thesis, University of Wisconsin–Madison, 167 pp., https://www.aos.wisc.edu/aosjournal/Volume40/White_PhD.pdf.

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

Winker, D. M., M. A. Vaughan, A. Omar, Y. Hu, K. A. Powell, Z. Liu, W. H. Hunt, and S. A. Young, 2009: Overview of the CALIPSO mission and CALIOP data processing algorithms. *J. Atmos. Oceanic Technol.*, **26**, 2310–2323, https://doi.org/10.1175/2009JTECHA1281.1.