Atmos. Meas. Tech., 18, 5393–5414, 2025
https://doi.org/10.5194/amt-18-5393-2025

Atmospheric Measurement Techniques
Open Access
EGU

Classifying thermodynamic cloud phase using machine learning models

Lexie Goldberger[1,★], **Maxwell Levin**[1,★], **Carlandra Harris**[1,2], **Andrew Geiss**[1], **Matthew D. Shupe**[3,4], **and Damao Zhang**[1]

[1]Pacific Northwest National Laboratory, Richland, WA 99352, USA
[2]Alabama State University, Montgomery, 36104, USA
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA
[4]Physical Sciences Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA
★These authors contributed equally to this work.

Correspondence: Damao Zhang (damao.zhang@pnnl.gov)

Received: 28 March 2025 – Discussion started: 7 May 2025
Revised: 1 August 2025 – Accepted: 1 August 2025 – Published: 16 October 2025

Abstract. Vertically resolved thermodynamic cloud-phase classifications are essential for studies of atmospheric cloud and precipitation processes. The Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Thermodynamic Cloud Phase (THERMOCLDPHASE) value-added product (VAP) uses a multi-sensor approach to classify the thermodynamic cloud phase by combining lidar backscatter and depolarization, radar reflectivity, Doppler velocity, spectral width, microwave-radiometer-derived liquid water path, and radiosonde temperature measurements. The measured pixels are classified as ice, snow, mixed phase, liquid (cloud water), drizzle, rain, and liq_driz (liquid+drizzle). We use this product as the ground truth to train three machine learning (ML) models to predict the thermodynamic cloud phase from multi-sensor remote sensing measurements taken at the ARM North Slope of Alaska (NSA) observatory: a random forest (RF), a multi-layer perceptron (MLP), and a convolutional neural network (CNN) with a U-Net architecture. Evaluations against the outputs of the THERMOCLDPHASE VAP with 1 year of data show that the CNN outperforms the other two models, achieving the highest test accuracy, $F1$ score, and mean intersection over union (IOU). Analysis of ML confidence scores shows that ice, rain, and snow have higher confidence scores, followed by liquid, while mixed, drizzle, and liq_driz have lower scores. Feature importance analysis reveals that the mean Doppler velocity and vertically resolved temperature are the most influential data streams for ML thermodynamic cloud-phase predictions. Lidar measurements exhibit lower feature importance due to rapid signal attenuation caused by the frequent presence of persistent low-level clouds at the NSA site. The ML models' generalization capacity is further evaluated by applying them at another Arctic ARM site in Norway using data taken during the ARM Cold-Air Outbreaks in the Marine Boundary Layer Experiment (COMBLE) field campaign. The models demonstrated similar performance to that observed at the NSA site. Finally, we evaluate the ML models' response to simulated instrument outages and signal degradation and show that a CNN U-Net model trained with input channel dropouts performs better when input fields are missing.

1 Introduction

Arctic clouds are one of the least understood elements of the Arctic climate system, but they play a significant role in regulating radiative energy fluxes at the surface, through the atmosphere, and at the top of the atmosphere (Cesana and Chepfer, 2012; Curry et al., 1996; Kay et al., 2008; Kay and L'Ecuyer, 2013; Shupe and Intrieri, 2004). One major factor in this uncertainty is the thermodynamic phase of clouds, which is crucial for understanding many cloud processes, including ice particle production via the Wegener–Bergeron–Findeisen process (Storelvmo and Tan, 2015), precipitation

formation (Mülmenstädt et al., 2015), the evolution of the cloud life cycle (Pithan et al., 2014), and also the response of clouds to global warming (Tan et al., 2025). Ice particles and liquid droplets differ significantly in number, size, shape, fall velocity, and refractive index, leading to vastly different radiative properties for clouds with different thermodynamic structures (Shupe and Intrieri, 2004). Accurate thermodynamic cloud-phase representations in climate models enhance the reliability of climate projections (Cesana et al., 2024). In addition, thermodynamic cloud-phase classification is often a prerequisite for retrieving cloud properties from remote sensing data, as most retrieval algorithms are designed for specific thermodynamic cloud phases and types (Shupe et al., 2016, 2015; Platnick et al., 2017).

The thermodynamic cloud phase can be determined using either aircraft in situ measurements (McFarquhar et al., 2011; Verlinde et al., 2007; Wendisch et al., 2019) or remote sensing observations (Avery et al., 2020; Barker et al., 2008; Hogan et al., 2003; Shupe, 2007; Turner et al., 2003). Aircraft in situ measurements use captured particle images from onboard probes to identify the thermodynamic cloud phase based on the shape and size of cloud particles. While in situ measurements offer thermodynamic cloud-phase identification, it is challenging to gather large aircraft datasets under diverse environmental conditions, and these measurements cannot provide routine or continuous daily data. Remote sensing observations, however, offer long-term continuous thermodynamic cloud-phase identification. Spaceborne remote sensing, in particular, enables global-scale thermodynamic cloud-phase classification, which can effectively constrain global climate models (Cesana and Chepfer, 2013; Tan et al., 2016). High-resolution ground-based remote sensing observations allow for detailed thermodynamic cloud classification, supporting studies of cloud processes, and the validation of high-resolution cloud-resolving model simulations (Fan et al., 2011; Kalesse et al., 2016).

The Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) user facility deploys advanced remote sensing instruments in climate-critical locations to monitor atmospheric states and processes. To address the need for accurate thermodynamic cloud-phase identification, ARM developed the Thermodynamic Cloud Phase (THERMOCLDPHASE) value-added product (VAP) (Zhang and Levin, 2024). Using the multi-sensor approach developed by Shupe (2007), the THERMOCLDPHASE VAP combines data from active remote sensing lidars, radars, microwave radiometers, and radiosondes to determine the vertically resolved thermodynamic cloud phase at ARM sites. THERMOCLDPHASE data are available through ARM Data Discovery for ARM's North Slope of Alaska (NSA) atmospheric observatory at Utqiaġvik, Alaska (formerly Barrow), from 2018 to 2022, as well as six other ARM high-latitude observatories (2025). It is noted that multi-sensor thermodynamic cloud-phase classification has been specifically developed for observations of Arctic clouds (Shupe, 2007). Since the algorithm does not include the classification of hail and graupel, it has difficulties distinguishing these hydrometeor types in deep convective cloud regimes over tropical and mid-latitude regions.

Threshold-based algorithms for determining the thermodynamic cloud phase, such as those used in the THERMOCLDPHASE VAP, have two major limitations. First, standard algorithms are static and do not improve with additional data or generalize to new regions. For ARM to apply the Shupe (2007) algorithm to sites other than the Arctic, where it was originally developed, fine-tuning these thresholds and rigorous quality testing are necessary before the data product can be used. This limits how quickly the product can be made available to scientists. Second, the realities of instrument deployment to harsh, remote environments mean that instrumentation can go offline periodically, and most conventional algorithms are not able to adapt when data inputs are missing. For ARM's thermodynamic cloud-phase product, the thermodynamic cloud phase cannot be accurately classified when one or more input data streams are missing.

Machine learning methods, in combination with conventional methods, can improve thermodynamic cloud-phase classification. ML algorithms' performance generally improves as they are trained with more data, and they can be trained to adapt to data issues such as low quality or missing inputs. There are multiple years of ARM THERMOCLDPHASE VAP data from the NSA site, and the product contains both the VAP and the individual data streams used to derive it, making it an excellent source of training data for the ML algorithms.

In this work, we develop three machine learning models with increasing complexity: a random forest (RF), a multilayer perceptron (MLP) neural network, and a convolutional neural network (CNN) with a U-Net architecture for classifying the thermodynamic cloud phase. We use the ARM THERMOCLDPHASE VAP from the NSA site as ground truth for training. In addition to evaluation of model performance on NSA data, we evaluate the ML models' generalizability to another ARM site (ANX) and test each model's robustness against simulated instrument data loss.

2 Methods

2.1 Datasets and data pre-processing

This study leverages the THERMOCLDPHASE VAP, produced at the ARM NSA atmospheric observatory (https://www.arm.gov/capabilities/science-data-products/vaps/thermocloudphase, last access: 2 September 2025), as the training data. The ARM NSA site (71°19′ N, 156°36′ W) is located on the northern Alaskan coastline (Verlinde et al., 2016). It experiences a variety of cloud types throughout the year, with predominantly ice clouds in winter, mixed-phase clouds in spring and fall, and liquid clouds in summer

(Shupe, 2011). The observatory is equipped with advanced atmospheric observing instruments, including cloud radars, depolarization lidars, radiometers, and radiosondes. These instruments provide comprehensive data for describing cloud and radiative processes at high latitudes. These data have been used to improve the representation of high-latitude cloud and radiation processes in Earth system models (Shupe et al., 2015; Zheng et al., 2023; Balmes et al., 2023).

The classification algorithm used to create the THERMOCLDPHASE VAP exploits the complementary strengths of cloud radar, depolarization lidar, microwave radiometer, and temperature soundings to classify cloud hydrometeors observed in the vertical column as ice, snow, mixed phase, liquid, drizzle, rain, and liq_driz (liquid+drizzle). The liq_driz class represents cases with liquid cloud and drizzle in the same volume, whereas the drizzle class indicates drizzle that has fallen below the cloud. In short, lidar backscatter is sensitive to small cloud droplets with high concentrations, while the lidar depolarization ratio can distinguish between spherical (i.e., liquid) and irregularly shaped particles such as ice crystals and snow. Radar reflectivity is dominated by large particles such as ice particles, snow, or raindrops, while higher-order radar moments provide more detailed information on, for example, particle fall speed. Supplemental data, such as the liquid water path from the microwave radiometer and temperature profiles from radiosondes, can be used to further refine thermodynamic cloud-phase identification. Combining these complementary observations provides a reliable approach to identifying thermodynamic cloud phases. An "unknown" label is assigned in cases when the thermodynamic phase of the hydrometeor cannot be identified due to missing input datasets or when the determined thermodynamic cloud phase is inconsistent with our understanding of cloud structures and physics based on past studies. The VAP also includes a "clear" classification when no hydrometeors are present. A full description of the method can be found in Shupe (2007). While Shupe (2007) used lidar and radar measurements to distinguish between clear and cloudy pixels, the THERMOCLDPHASE VAP applies the phase classification algorithm to cloudy pixels identified by the ARM Active Remote Sensing of CLouds (ARSCL) VAP (https://www.arm.gov/data/science-data-products/vaps/arscl, last access: 2 September 2025) (Clothiaux et al., 2001). The ARSCL VAP provides cloud boundaries for up to 10 cloud layers by combining radar, lidar, and radiometer measurements.

The THERMOCLDPHASE VAP reads in micropulse lidar (MPL) or high-spectral-resolution lidar (HSRL) backscatter and depolarization ratio data from the Micropulse Lidar Cloud Mask (MPLCMASK) VAP (https://www.arm.gov/data/science-data-products/vaps/mplcmask, last access: 2 September 2025) (Flynn et al., 2023) or HSRL data (Goldsmith 2016), respectively; radar reflectivity, mean Doppler velocity, and Doppler spectral width data from the ARSCL VAP; liquid water path data from the Microwave Radiometer Retrievals (MWRRET) VAP (https://www.arm.gov/data/science-data-products/vaps/mwrret, last access: 2 September 2025) (Gaustad et al., 2011); and temperature data from the Interpolated Sonde (INTERPSONDE) VAP (https://www.arm.gov/data/science-data-products/vaps/interpsonde, last access: 2 September 2025) (Fairless et al., 2021). The HSRL system is deployed at only a few ARM observatories and ARM mobile facility (AMF) field campaigns. When HSRL data are available, the THERMOCLDPHASE VAP uses the HSRL backscatter coefficients and LDR thresholds, as outlined in Shupe (2007), to distinguish between liquid and ice. The MPL system, on the other hand, is deployed at all ARM fixed atmospheric observatories and in nearly all AMF field campaigns. The THERMOCLDPHASE VAP uses the gradient of MPL backscatter (MPLGR), following Wang and Sassen, (2001), to distinguish between liquid and ice. We employ the thermodynamic cloud-phase classification data that utilize the MPLGR method to train ML models so that the trained models can be readily extended to other ARM observatories. Ultimately, the THERMOCLDPHASE VAP then outputs seven hydrometeor-phase classifications at 30 m vertical and 30 s temporal resolutions. The VAP and the datasets used to produce it are publicly available through ARM's Data Discovery tool (https://adc.arm.gov/discovery/, last access: 2 September 2025).

An example of multi-sensor remote sensing measurements and thermodynamic cloud-phase classification from the THERMOCLDPHASE VAP on 15 August 2021 at the ARM NSA observatory is shown in Fig. 1. The day started with a deep precipitating system with some embedded convection before 09:00 UTC, with cloud tops reaching up to 8 km and temperatures near the cloud top close to −40 °C. KAZR radar signals can penetrate through the cloud and provide measurements of the cloud structure. Increased radar reflectivity (Z_e), downward motion (indicated by negative mean Doppler velocity, MDV), and Doppler spectral width (W) around 1 km suggest a transition from cold to warm precipitation (Fig. 1c, d, and e). Furthermore, the radar bright band is observable at ∼ 1 km when large falling ice crystals start to become coated with melted liquid water (Fig. 1c). Lidar signals, however, were quickly attenuated by warm raindrops near the surface, as shown in Fig. 1a and b. The large liquid water path (LWP) retrieved from the microwave radiometer and warm temperatures near the surface provide support for this identification (Fig. 1f and g). As shown in Fig. 1g, the THERMOCLDPHASE VAP identifies ice and mixed-phase regions in the middle and upper portions of the cloud system, with snow pixels occasionally present in the middle layers. Below approximately 1 km, warm cloud phases and precipitation, including liquid, drizzle, and rain, are observed. Two additional, relatively shallower cloud systems with similar cloud-phase structures were observed between 10:00 and 13:00 and 15:00 and 19:00 UTC. Interestingly, two mid-level thin liquid-layer clouds were observed after each of the first two systems. However, due to warmer

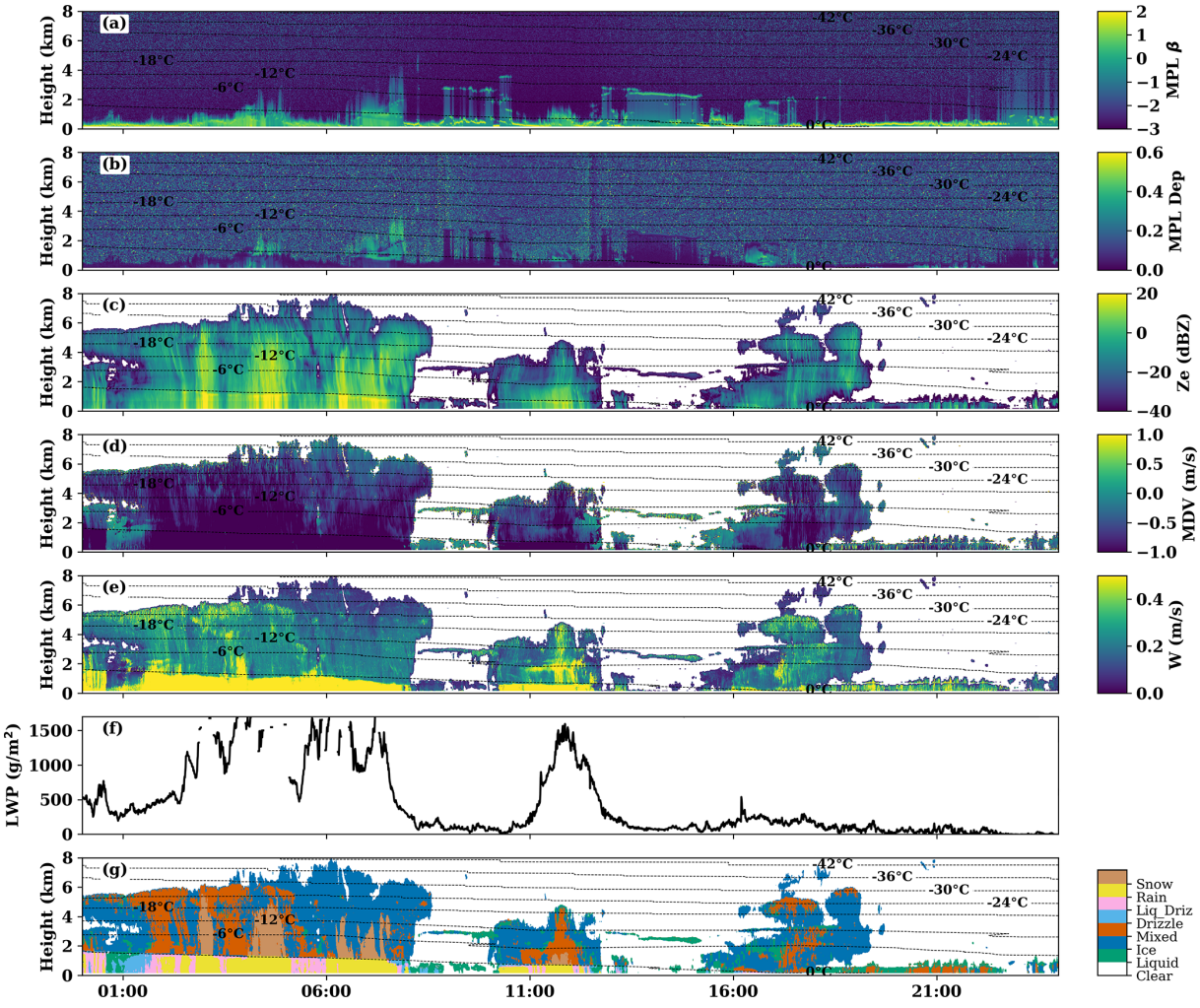

Figure 1. An example of multi-sensor remote sensing measurements of clouds and the thermodynamic cloud-phase classification from the THERMOCLDPHASE VAP on 15 August 2021 at the ARM NSA site. Panels from top to bottom are the **(a)** MPL attenuated backscatter (MPL β), **(b)** MPL linear depolarization ratio (MPL LDR), **(c)** Ka-band ARM zenith radar (KAZR) equivalent reflectivity factor (Z_e), **(d)** KAZR mean Doppler velocity (MDV), **(e)** KAZR Doppler spectral width (W), **(f)** liquid water path (LWP) from the MWRRET VAP, and **(g)** thermodynamic cloud-phase classification from the THERMOCLDPHASE VAP. Negative MDV values represent downward motions toward the surface. The dashed lines in panel **(g)** are isothermal lines based on the ARM Interpolated Sonde (INTERPSONDE) VAP.

cloud top temperatures, a lack of ice-nucleating particles (INPs), or other processes, these liquid cloud layers did not produce ice or produced ice that was immediately sublimated right below the cloud base. Further cloud model simulations could provide insights into these processes (Solomon et al., 2018, 2011). After 18:00 UTC, a typical polar low-level stratocumulus cloud with liquid droplets at the top and mixed-phase pixels below is observed. Note that the low radar reflectivity that appears to be detached from the cloud below between 20:00 and 22:00 UTC could be artifacts caused by side-lobe impacts of the KAZR moderate sensitivity mode (MD) (Silber et al., 2018). Accurately detecting and removing these radar artifacts are being investigated by the ARM radar data team (Ya-Chein Feng, personal communication, January 2025).

The various input fields for the VAP are listed in Table 1. These inputs have different units, can differ in scale by orders of magnitude, and may include extreme outlier values. To facilitate training of the neural networks, which can be sensitive to input scaling, each input was range limited and then re-scaled to an approximate range between -2 and 2. The range limiting was chosen to only cut off erroneous or missing data points (the ARM datasets assign missing data a value of -9999) and restricts the inputs to a physically plausible range. The scaling values were determined manually based on histograms of the training dataset and are detailed in Table 1. Additionally, the MPL backscatter and MWRRET LWP variables were log-scaled because these observations span several orders of magnitude. Training data were limited to periods where all instruments were operational, and in-

Table 1. The formulas used to normalize the input data for the MLP and CNN models. The clip function is used to constrain the values in an array within a specified range. For example, $\text{clip}(x, x_l, x_u) = \{x_l, (x < x_l); x, (x_l \le x \le x_u); x_u, (x > x_u)\}$.

Variable	Units	Lower bound	Upper bound	Full normalization formula
MPL backscatter (MPL β)	Counts/µs	1×10^{-8}	1×10^{4}	$(\log(\text{clip}(x, 1 \times 10^{-8}, 1 \times 10^{4})) + 6)/8$
MPL linear depolarization ratio (MPL dep)	NA	0	1	$\text{clip}(x, 0, 1) \times 2 - 1$
Radar reflectivity (Z_e)	dBz	−70	70	$(\text{clip}(x, -70, 70) + 20)/30$
Radar mean Doppler velocity (MDV)	m s^{-1}	−8	8	$\text{clip}(x + 0.5, -8, 8)/2$
Radar spectral width (W)	m s^{-1}	−1	4	$\text{clip}(x \times 5, -1, 4) - 0.5$
Radar linear depolarization ratio (radar dep)	NA	−20	20	$\text{clip}(x + 20, -20, 20)/6$
MWRRET liquid water path (LWP)	g m^{-2}	0.1	2000	$(\log(\text{clip}(x, 0.1, 2000)) - 3)/2$
Temperature profile (T)	°C	−100	50	$(\text{clip}(x, -100, 50) + 30)/30$

stances where the VAP output was labeled "unknown" were excluded from the training process. Based on 1 year of data from 2021 at the NSA site, 5.9 % of cloud hydrometeors were classified as unknown.

2.2 Machine learning models

2.2.1 Random forest

A random forest (RF) is a meta-estimator that fits multiple decision tree classifiers using a best-split strategy on various sub-samples of the dataset. The individual tree's predictions are then averaged to improve predictive accuracy and control over-fitting (Breiman, 2001). The RF model uses an ensemble of 100 decision trees and operates on individual pixels (1.6 million samples), unlike the CNN, which processes time–height images. We used the Scikit-learn library (Pedregosa et al., 2011) to train a random forest classifier, which took less than 2 h to train using CPUs. The RF was trained using a standard scaler to re-scale the input variables and excluding any pixels marked as "unknown" in the VAP. A total of 90 RF configurations were tested, with the best model determined by considering training accuracy, validation accuracy, and validation $F1$ score (precision) (Eq. 1). Categorical accuracy evaluates the overall percentage of correct predictions but can be biased in imbalanced datasets. Precision measures the proportion of correct positive predictions out of all positive predictions made. A higher precision indicates fewer false positive predictions. Recall evaluates the proportion of actual positive instances correctly identified. A higher recall indicates fewer false negative predictions. The $F1$ score is the harmonic mean of precision and recall, which is defined as

$$F1 = 2\text{TP}/(2\text{TP} + \text{FN} + \text{FP}). \quad (1)$$

The $F1$ score provides a balanced measure of both precision and recall. The best model used 40 trees, with 10^5 samples used to train each tree, and was trained with a maximum of 2 features used for each split, a maximum depth of 20, and no restriction on the maximum number of leaf nodes. Our initial experiments with the RF model showed that its performance did not change substantially with hyperparameter adjustments, and the best validation performance was achieved using the default Scikit-learn hyperparameters.

2.2.2 Neural network

We also trained a conventional multi-layer perceptron (MLP)-style neural network. The MLP is a supervised learning algorithm that can learn a non-linear, continuous, and differentiable mapping between the input data and the target classifications (Bishop, 2006). The MLP takes the 8 normalized input values (Table 1), has 5 hidden layers with rectified linear unit (ReLU) activation functions and 100 neurons each, and has a 7-neuron output layer that applies a softmax activation function. Like the RF model, the MLP operates pixel by pixel to generate phase classifications. The MLP was also trained using the Scikit-learn library (Pedregosa et al., 2011) with the same dataset used to train the RF, which was standardized as well. A total of 41 variants of the MLP were tested with a robust scalar, quantile transformer, or standard scalar applied directly to the data. The best was trained using the Adam optimizer with an adaptive learning rate initialized at 0.001, a batch size of 200, and a categorical cross-entropy loss function. Training was terminated after 134 epochs due to early stopping. The validation fraction was 0.2.

2.2.3 Convolutional neural network

Deep convolutional neural networks (CNNs) are powerful machine learning models originally developed for computer vision and image processing tasks (Krizhevsky et al., 2017; LeCun et al., 1998; Heaton, 2018). CNNs learn convolutional kernels that can efficiently represent information about spatial structures in their input fields and are translationally equivariant models, making them optimal for image-recognition and segmentation tasks. Both RFs and CNNs have demonstrated effectiveness in labeling radar and lidar data for the classification of radial velocity and precipitating hydrometeors (Lu and Kumar, 2019; Veillette et al., 2023).

Ronneberger et al. (2015) introduced the "U-Net", a CNN architecture designed for image segmentation that maps an input image to pixel-level class labels, and several improved although more complex variants have been developed since its introduction (Huang et al., 2020; Zhou et al., 2018). U-Net and its variants are broadly applicable to both classification- and regression-style image-to-image mapping problems and have now been adapted for a wide range of use cases in the atmospheric sciences (Galea et al., 2024; Geiss and Hardin, 2021; Lagerquist et al., 2023; Sha et al., 2020; Weyn et al., 2021; Wieland et al., 2019). Here, we use a CNN similar to the original U-Net that has been modified for the thermodynamic cloud- and precipitation-phase retrieval task as shown in Fig. 2. The U-Net was implemented using the TensorFlow Keras Python library (Chollet, 2015) and trained to ingest inputs of size $128 \times 384 \times 8$ and produce a 128×384 pixel-level phase classification mask. Missing instrument data values are filled with a value of -9999 prior to dataset normalization, as specified in Table 1. This means that, after normalization, they will be mapped to the lowest allowed value for the corresponding input field by the clip function. On the other hand, if the ground truth labels for any batch of data are missing or classified as unknown, the entire batch is discarded and not used to train the model. The $128 \times 384 \times 8$ input shape corresponds to samples that are 1 h in duration, 12 km in height, and have eight input fields, respectively. The U-Net was trained using the Adam optimizer with an initial learning rate of 0.001, a batch size of 16, and categorical cross-entropy loss. Training was terminated when the mean intersection over union (IOU) reached a maximum value after epoch 10. IOU is defined per class as

$$\mathrm{IOU} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN} + \mathrm{FP}), \tag{2}$$

where TP, FN, and FP represent true positives, false negatives, and false positives, respectively. The mean IOU is calculated by averaging the IOU of each class and is not biased by the class imbalance.

An ablation study was used to determine the optimal U-Net design. This involved altering one design choice at a time, retraining the model, and evaluating the model's performance on the validation data (evaluating on cloudy pixels only, no clear sky). We tested cases with and without dropout layers, channel-wise dropout layers (applied only to the input tensor to simulate instrument dropouts), and batch normalization layers. We also ran experiments varying the number of convolutional layers in each block, the number of channels in the convolutional layers, the type of activation functions, and the class weights used during training (Ioffe and Szegedy, 2015; Srivastava et al., 2014). During the ablation study, the U-Nets were evaluated using several metrics, including categorical cross-entropy computed only on cloudy regions, training loss (categorical cross-entropy computed on all regions), mean IOU, and categorical accuracy. The categorical accuracy is averaged over all pixel classifications and, because of class imbalances, is more representative of model skill on the most common classes.

Ultimately, the best U-Net configuration performed the best across all four metrics (lowest cloudy cross-entropy and all-sky cross-entropy and highest mean IOU score and categorical accuracy). The best results with the CNN were achieved with no channel-wise dropout layers; 2 convolutions in each block with the first followed by a dropout layer and the second followed by a batch normalization layer; 64, 64, 64, 128, 128, and 256 channels in the convolutional layers (where the ordering represents depth in the U-Net); leaky ReLU activation functions following the dropout and batch normalization layers; and no class weighting. These design choices resulted in a mean IOU score of 0.810 on the testing dataset, about 0.1 larger than the results of other model configurations we tested. This also resulted in a training loss of 0.018, which was 0.025 less than the other configurations. Notably, the U-Net configuration with channel-wise dropout layers was the second-best model, with an IOU score of 0.528 and training loss of 0.054. We note that these results are based on testing with complete inputs; however, when the U-Net is evaluated with simulated instrument outages, the versions that were trained with channel-wise dropout applied to the inputs performed better (details in Sect. 4). The results for all the ablation tests are documented in the Supplement.

2.3 Training dataset

A total of 3 years of data at the ARM NSA site, from 2018–2020, were used for training and validation, and 1 year of data, from 2021, were used for testing. For the MLP and RF models, a subset of 40 000 pixels from the 3 years of training data selected randomly were used for each cloud phase and 10 000 pixels for each cloud phase for validation. For the CNN model, the first 80 % of data from 2018–2020 were used for training and the remaining 20 % for validation. The input fields were organized as three-dimensional arrays time $\times$ height $\times$ channel samples. The seven unique cloud-phase classifications produced by the THERMOCLDPHASE VAP were used as targets (the eighth was clear sky and was not used). The training time for each model is reported in Table 2. The RF and MLP ran on CPUs, while the CNN was trained using GPU. For this reason, the MLP and CNN train in comparable time, around $\sim$ 110 min, though the CNN requires more computation. Meanwhile, the RF trains an order of magnitude faster, around 12 min. Inference time was inconsequential for all three models, which can each classify a day of data within a few seconds.

The different methods of training set construction and input format used by each of the models create different class imbalances and inherently complicate a direct comparison between models. For the RF and MLP models, an equal number of samples of each of the cloud-phase types was used to train and validate the models because they operate at a pixel level. Meanwhile, the CNN processes full time–

Table 2. Model performance metrics for the three machine learning models on the test dataset.

Model	Accuracy (%)	Precision*	Recall*	$F1$ score*	IOU*
CNN	95.7	0.890	0.894	0.891	0.811
MLP	85.7	0.760	0.905	0.815	0.704
RF	87.2	0.789	0.913	0.837	0.735

* Using a macro-average across classes.

height images, and its performance will be biased towards the most common pixel types (ice is the most common class observed at NSA). These challenges are inherent to the different ML models; for example, the RF cannot be trained on the huge dataset the CNN uses due to computational constraints. In the future, the CNN could potentially be trained with a class-weighted loss function to ensure that the model can identify the minority classes with greater accuracy, but class weighting does not have the exact same effect as rebalancing the class frequency, particularly when the class imbalance is large. Balancing the class distribution ensures that the model receives gradients of a similar scale from each class at approximately the same frequency throughout training. In contrast, altering the class weights results in predominantly small gradients from the majority classes, with occasional large gradients from minority classes. Therefore, achieving optimal performance is likely not as straightforward as selecting class weights that are inversely proportional to class frequency and will likely require fine-tuning of hyperparameters. Recent research has reported better results with combo loss (Taghanaki et al., 2019) rather than weighting schemes in similar applications (Xie et al., 2025).

3 Results

Once the ML models were trained and validated, they were applied to 1 year of multi-sensor remote sensing measurements from 2021 to predict the thermodynamic cloud phase (THERMOCLDPHASE-ML). The predicted phase classifications were compared to the VAP to evaluate the performance of the three ML models.

3.1 Applying trained ML models to remote sensing measurements

Figure 3 provides an example of thermodynamic cloud-phase classifications from the three ML models compared with the THERMOCLDPHASE VAP on 15 August 2021 at the ARM NSA site. Among the predictions from the three ML models, the CNN demonstrates the best agreement with the THERMOCLDPHASE VAP, capturing nearly identical thermodynamic cloud-phase structures. The MLP and RF models also show good agreement with the VAP but tend to

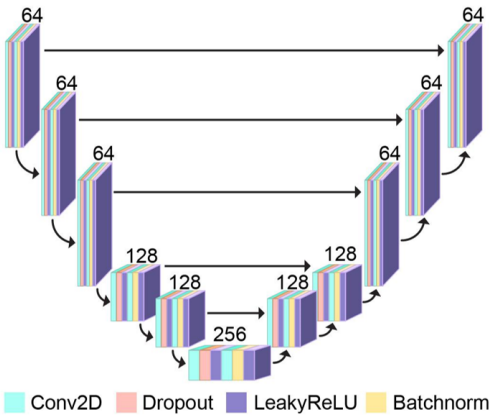

Figure 2. An illustration of the most effective U-Net architecture tested, showcasing both its encoding and its decoding paths along with their channel dimensions. Given two-dimensional eight-channel inputs $128 \times 384 \times 8$, where the eight channels are the variables in Table 1 and a cloud mask, the model produces a $128 \times 384 \times 8$ output, where each channel represents the softmax probability of a pixel belonging to one of the eight cloud-phase classes.

overestimate mixed-phase pixels in the ice-dominated high clouds between 00:00–09:00 and 15:00–18:00 UTC and underestimate ice-phase pixels in the low-level clouds between 15:00–23:00 UTC. Notably, the ML models provide confidence scores for their predictions, where higher scores indicate greater certainty. For the CNN and MLP models, the raw model output is a softmax probability score for each phase class. For the RF, confidence is calculated using the mean of the predictions of trees in the RF. As shown in Fig. 3e–g, the CNN consistently generates higher confidence scores compared to the MLP and RF models. Regions with low confidence scores from the MLP and RF models often correspond to areas where these models misclassify thermodynamic cloud phases. As shown in Fig. 3e–g, all ML models exhibit significantly lower confidence scores within the melting layer – a region characterized by rapid transitions in particle phase, shape, and fall speed. While this zone is critical for understanding cloud regime shifts, it remains difficult to resolve. Improving detection in this region will require a refined training dataset specifically focused on the melting layer, which remains an active area of research (Brast and Markmann, 2020; Xie et al., 2025). In Fig. S1 in the Supplement, we plot confidence score bins versus accurate classifications for the 2021 data. The MLP and RF models' accuracy linearly increases with higher confidence. For the CNN, accuracy increases linearly for confidence scores above 40 %. There is a local maximum in accuracy for low confidence scores between 20 %–30 %. For these cases, there are several orders of magnitude fewer data points, and the majority of correctly classified cases are ice. Because NSA is dominated by ice, classifying a non-clear-sky pixel as ice, even with low confidence, has a high chance of being correct for

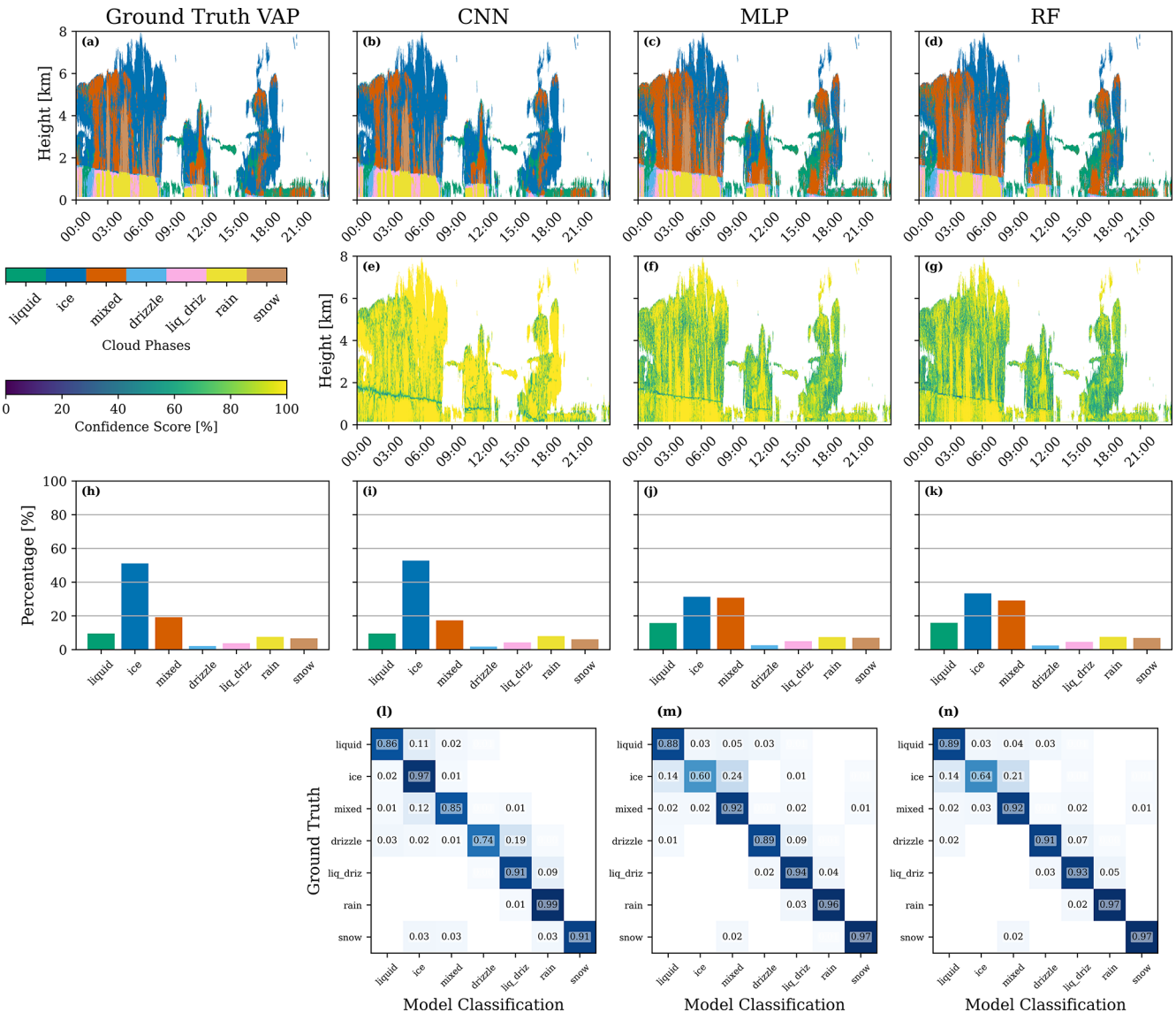

Figure 3. Thermodynamic cloud-phase classifications from the three ML models and their comparisons against the THERMOCLDPHASE VAP on 15 August 2021 at the NSA site. **(a–d)** Time–height plots of thermodynamic cloud-phase classifications from the VAP, as well as from CNN, MLP, and RF model predictions, respectively; **(e–g)** confidence scores of thermodynamic cloud-phase classification predictions from the three ML models; **(h–k)** histograms of thermodynamic cloud-phase distributions; and **(l–n)** normalized confusion matrices for each model. Panel **(a)** is identical to Fig. 1g.

this dataset. At the pixel level of the thermodynamic cloud-phase classification, Fig. 3h indicates that the day was dominated (volume-wise) by ice-phase pixels, followed by liquid and mixed-phase pixels. Small numbers of snow, drizzle, and liq_driz pixels were also identified. The histogram plots of ML-predicted thermodynamic cloud phases in Fig. 3i–k show that the CNN produces a histogram closely matching the VAP. In contrast, the MLP and RF models tend to underestimate ice-phase pixels while overestimating liquid, mixed-phase, and liq_driz pixels.

Figure 3l–n provide a more detailed evaluation of thermodynamic cloud-phase classifications from the three models through confusion matrices. The multi-class confusion matrix is a 7×7 grid with a row and column for each of the cloud phases. Each row represents the class reported by the VAP, and columns show the class predicted by ML. Correct predictions (true positives) are found along the diagonal, while misclassifications are in the off-diagonal elements. The sum of the columns ideally would equal 1; a sum greater than 1 indicates an over-classification of pixel type. On this day, liquid, mixed-phase, drizzle, and snow pixels were ac-

curately identified by all three ML models, with accuracies exceeding 0.8. While the CNN model also accurately classified ice-phase pixels, the MLP and RF models frequently misclassified them as liquid or mixed-phase pixels. This case has pixel percentages above 5 % for all cloud-phase types and has high accuracy for all types, including liq_driz and rain pixels. In cases consisting predominately of ice clouds, relatively low accuracy for liq_driz and rain pixels is reported compared to other categories, with the CNN performing the worst, likely due to the extremely low occurrence of these pixel types and an overzealousness in predicting ice.

3.2 Analyses of ML model performance

Given that the confidence score reflects the uncertainty of ML predictions, it is essential to analyze confidence scores and their relationship to accuracy for different thermodynamic cloud phases. Figure 4 presents a comprehensive statistical analysis of ML model confidence scores based on 1 year of data from 2021 at the NSA site. Overall, the confidence scores for thermodynamic cloud-phase predictions peak near 100 %, which is promising. Among the phases, predictions for ice, rain, and snow generally exhibit higher confidence scores across all three ML models. The ice phase, in particular, is reliably predicted – especially by the CNN model – due to the availability of key information such as the lidar backscatter and depolarization ratio, radar reflectivity, mean Doppler velocity and spectral width, and temperature (Shupe, 2007). The rain and snow phases, representing large particles in warm and cold conditions, respectively, can be identified using key information such as radar reflectivity, mean Doppler velocity, and temperature. In contrast, the confidence scores for the liquid-phase predictions are lower than those for the ice, rain, and snow phases. While the liquid phase can theoretically be reliably determined using lidar backscatter and depolarization ratio measurements, lidar signals are often quickly attenuated by low-level clouds, as illustrated in Fig. 1a and b. Under such conditions, identifying liquid-phase pixels becomes challenging when relying solely on radar reflectivity and spectral width data (Silber et al., 2020). The mixed, drizzle, and liq_driz phases have even lower confidence scores, likely due to the inherent difficulties in extracting their distinguishing characteristics from available measurements. Among the three ML models, the CNN achieves the highest confidence scores across all thermodynamic cloud phases. The MLP model exhibits confidence scores comparable to those of the RF model for the liquid, ice, mixed, drizzle, and liq_driz phases but shows significantly lower confidence scores for the rain and snow phases.

Figure 5 shows the frequency of thermodynamic cloud phases at the NSA site, as derived from the THERMOCLDPHASE VAP (labeled "VAP"), and predictions from the three ML models using 1 year of testing data. Due primarily to the low polar temperatures, the NSA site is primarily dominated by the ice phase, followed by mixed, snow, and liquid phases. Warm phases, including drizzle, liq_driz, and rain, occur much less frequently and are mostly confined to the summer season (Shupe, 2011). Comparing the ML predictions with the THERMOCLDPHASE VAP, the CNN closely matches the VAP's percentage distribution of thermodynamic cloud phases. In contrast, both the MLP and the RF models predict lower percentages for the ice phase but higher percentages for the liquid, mixed, drizzle, and liq_driz phases, consistent with the case observed in Fig. 3h–k.

Figure 6 presents the confusion matrices for the three models computed on the testing set. All models achieved over 80 % accuracy for each cloud phase. The correct prediction percentages are close for the three ML models, except that CNN has a dramatically higher correct prediction for ice than the other two ML models. The CNN correctly identified ice 99 % of the time. However, it occasionally misclassified liquid (8 %), mixed (12 %), and drizzle (1 %) as ice. Because there are so few total instances of these phases (Fig. 5), these misidentifications did not contribute much to reducing the overall accuracy of the model. However, to do a true comparison of the models to the best of our ability, we retrained the RF and MLP models on a random sample of 1.6 million pixels from the training dataset (using the same number of samples as the class-balanced training and the same inputs and normalizations used by the CNN), where the distributions of phases match closely with the overall phase distribution in the VAP. We examined how the "imbalanced" RF and MLP compared to the CNN (Fig. S2). Focusing on the prediction of ice, the "balanced" RF and MLP models only misclassify liquid and mixed phases as ice 4 % and 5 % of the time, respectively (Fig. 6), while the "imbalanced" RF misclassifies liquid and mixed phases 25 % and 22 % of the time, and the "imbalanced" MLP misclassifies them 22 % and 21 % of the time (Fig. S2). Regarding the performance of the "imbalanced" models on the warm cloud phases, for drizzle, the CNN correctly identifies it 83 % of the time, the imbalanced RF 86 %, and the imbalanced MLP 81 %. Compared to the balanced RF (90 %) and MLP (88 %), the imbalanced datasets perform worse on this metric.

The performance of the three ML models was statistically evaluated using performance metrics listed in Table 2. These metrics include categorical accuracy, precision, recall, $F1$ score, and mean IOU (Eq. 1). Here, we calculated the test accuracy as the percentage of pixels that match the VAP. Precision, recall, $F1$ score, and IOU are calculated for each phase class and reported as an average across the classes to reduce bias due to class imbalance. These metrics provide us with information to evaluate the performance of the three ML models in classifying thermodynamic cloud phases on a pixel-by-pixel level.

Table 2 shows that each model agreed with the THERMOCLDPHASE VAP in more than 85 % of the utilized samples. The CNN achieved the highest test accuracy, $F1$ score, and mean IOU. The RF model performed slightly better than the MLP across these metrics but was significantly outper-

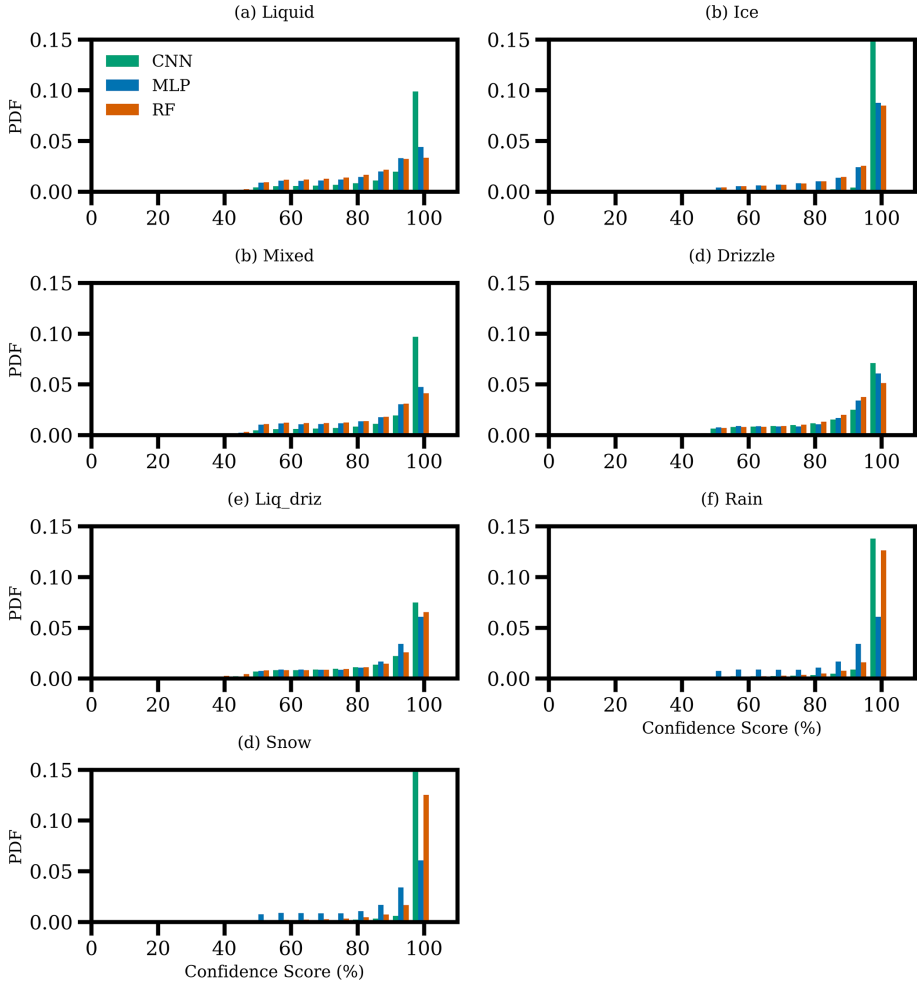

Figure 4. Probability density functions (PDFs) of confidence scores for thermodynamic cloud-phase predictions from the three ML models using 1 year of data in 2021 at the NSA site.

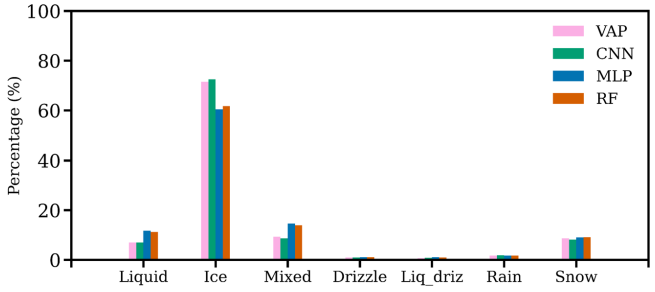

Figure 5. Percentage distributions of thermodynamic cloud phases from the THERMOCLDPHASE VAP (labeled "VAP") and predictions from the three ML models, based on 1 year of data from 2021 at the NSA site.

formed by the CNN. We hypothesize that the CNN's superior performance is due to its ability to evaluate the input time–height arrays (sections of data covering 11 km in height by 1 h) holistically rather than on a pixel-by-pixel basis. This approach allows the CNN to leverage information from neighboring pixels and potentially assess larger-scale features, such as cloud shape, to improve classification accuracy.

Another aspect of evaluation is the performance of the models with respect to altitude. Figure 7 presents vertically resolved $F1$ scores and mean IOU scores for the ML models, overlaid on a stacked histogram of thermodynamic cloud-phase category occurrences based on the VAP. Vertically resolved cloud phases converge toward ice-only clouds due to the extremely cold environment at higher altitudes. A peak in liquid-phase occurrence is observed around ~ 1 km, which may be due to the prevalence of low-level polar stratiform mixed-phase clouds with a thin liquid layer at the top in the VAP (Zhang et al., 2010; Silber et al., 2021; Zhang et al., 2017) or due to the artifacts caused by the KAZR MD mode (MD) side lobe, as discussed in Sect. 2.1. The $F1$ scores and mean IOU are consistent with altitude until 8 km, after which they start to increase across the three ML models, primarily

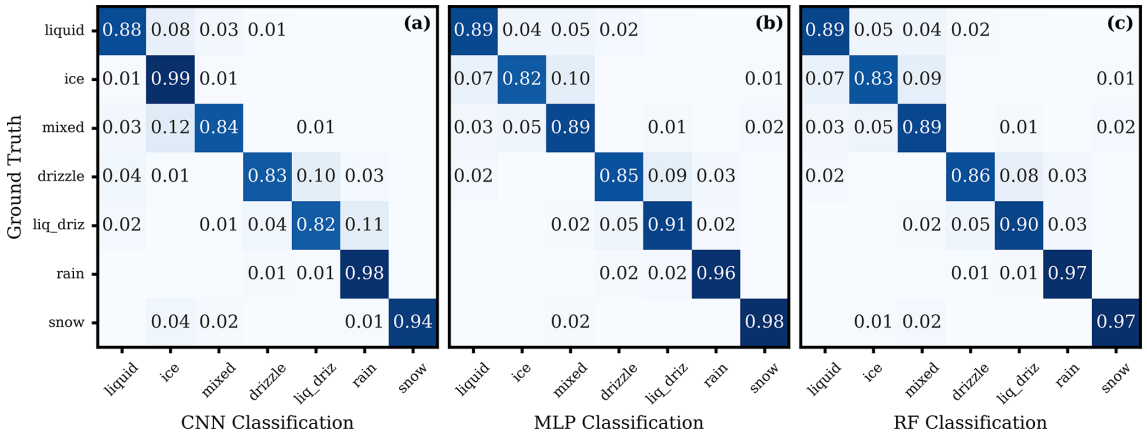

Figure 6. Confusion matrices computed on the 2021 NSA test dataset for **(a)** the CNN U-Net model, **(b)** the MLP model, and **(c)** the RF model. The values are normalized by row, with the main diagonal showing true positive predictions and values off the main diagonal representing incorrect predictions.

due to the higher frequency of the ice phase at greater altitudes and the fact that the ice phase is more reliably predicted by all three ML models, as shown in Fig. 4b. The CNN consistently achieves significantly higher $F1$ scores than the MLP and RF models at altitudes below ~ 6 km. This is attributed to the greater diversity of thermodynamic cloud phases at lower altitudes and the CNN's strong performance across all phases, as shown in Fig. 4.

3.3 Input feature importance

To identify which input features are most influential in determining cloud phase and to provide additional context for model performance, we calculate permutation feature importance (Breiman, 2001) for the three ML models by cloud-phase class. We assess the permutation importance of an input feature, defined as the model's recall score for a specific phase category on the test set minus its recall score resulting from shuffling the values of the input feature (randomly reordering their positions within the column), which effectively removes its relationship with a specific phase category. A significant difference between recall scores indicates that the feature is important, while little or no change suggests the feature has minimal importance. This is done for each phase class, and the recall score is used specifically because it shows the reduction in the models' ability to positively identify specific thermodynamic phases. This process is repeated for the CNN, MLP, and RF models and is reported in Fig. 8.

Overall, input features from radar measurements (Fig. 8b, f, and j), including Z_e, MDV, and W, and radiosonde temperature measurements (Fig. 8d, h, and l) are the most significant for classifying thermodynamic cloud phases across all three models. In contrast, input features from lidar measurements (Fig. 8a, e, and i) and the MWRRET LWP (Fig. 8c, g, and k) are less influential, probably because lidar signals are quickly attenuated by persistent low-level clouds at the

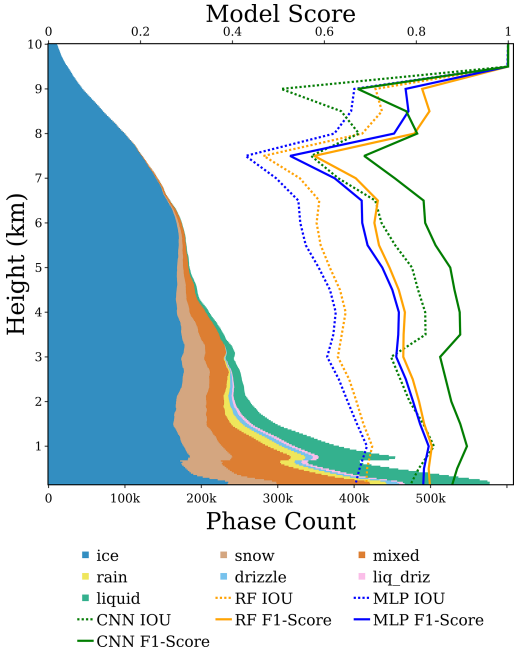

Figure 7. Vertically resolved ML model $F1$ scores and mean IOU scores, overlaid on a stacked histogram of the frequency of the thermodynamic cloud-phase categories. A height bin size of 0.5 km is used to calculate the vertical profiles of mean IOU and $F1$ scores. Noise around 7.5–10 km is likely due to phase extinction and low pixel count.

NSA site (Shupe et al., 2011; Zhang et al., 2017), and LWP provides only column-integrated information rather than detailed vertical profiles. Future work may want to explore the feature importance restricted to pixels that were observed by both radar and lidar to reevaluate the lidar's importance.

The colors in Fig. 8 represent different phase categories and enable feature importance to be assessed for each cate-

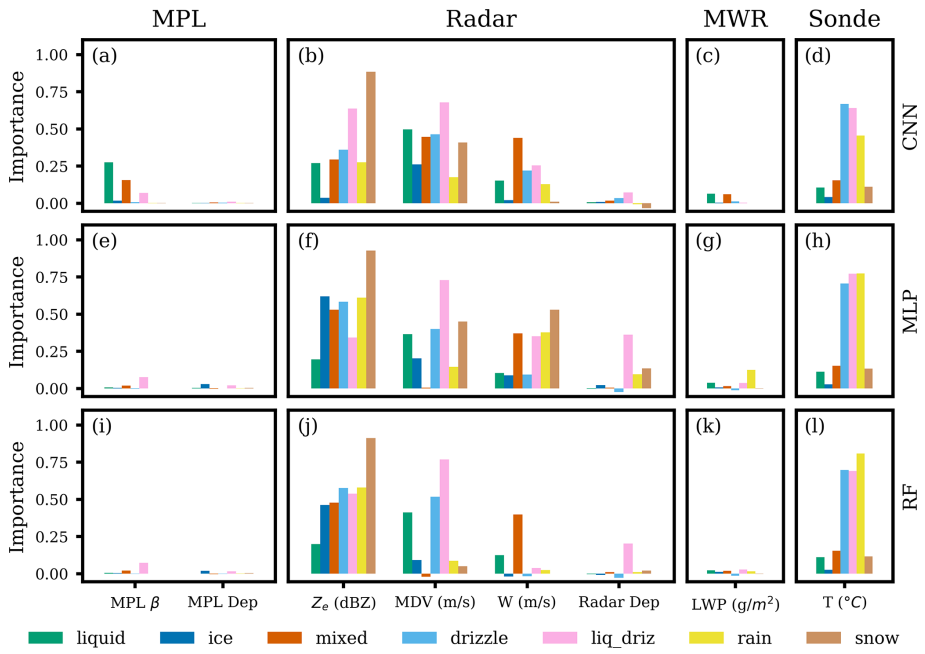

Figure 8. Permutation feature importance of predicting thermodynamic cloud phases from **(a–d)** the CNN U-Net model, **(e–h)** the MLP model, and **(i–l)** the RF model. Features are grouped by the instruments they are derived from. The abbreviations are defined in Fig. 1. Radar dep represents radar linear depolarization.

gory. The main focus of permutation feature importance is the relative importance of the features instead of their absolute values. This is because the sum of the importance is not necessarily meaningful, given that feature interactions and the non-additive nature of the method can affect the results. For the CNN model, radar Z_e, MDV, and MPL β are identified as the three most important input features for determining the liquid phase. This aligns with the logic used in threshold-based algorithms by Shupe (2007) for liquid-phase identification. As shown in Fig. 8a, lidar backscatter shows notable importance in the CNN model. While the lidar backscatter and depolarization ratio offer direct and reliable indicators of liquid-phase presence, radar-based variables – such as reflectivity, mean Doppler velocity, and spectral width – can also contain useful signatures of liquid-phase clouds (Luke et al., 2010; Yu et al., 2014; Kalogeras et al., 2021; Schimmel et al., 2022), as evidenced in Fig. 8b, f, and j. The lidar measurement's lower feature importance relative to radar measurements was also observed on days with single-layer, low-level liquid clouds (Fig. S3). For the ice phase, input feature importance is generally lower, likely because the ice phase can be independently identified using multiple input features. As a result, even when one input feature is missing, the ice phase can still be accurately classified using the remaining features. The key features for identifying the mixed phase are Z_e, MDV, and W. For drizzle, liq_driz, and rain, Z_e, MDV, and temperature are the most important, likely due to the complexity of distinguishing these phases, requiring multiple measurements. Z_e is the primary feature for snow identification, followed by MDV and temperature, consistent with Shupe et al. (2016), where snow identification relied on Z_e and temperature. The importance of MDV for snow may result from its covariance with Z_e. The input feature importance for the other two models (Fig. 8b and c) is generally similar to that of the CNN model. Broadly, the feature importance in Fig. 8 aligns with physical intuition and with the logic used by Shupe (2007), indicating that ML models successfully captured the relationships between remote sensing measurements and the thermodynamic cloud phases.

3.4 Application to another ARM site: COMBLE

To assess the generalization capability of the ML models, we applied them at a different ARM mobile facility (AMF) observatory. The ARM Cold-Air Outbreaks in the Marine Boundary Layer Experiment (COMBLE) field campaign deployed an AMF at a coastal site in Andenes, Norway (69.141° N, 15.684° E; referred to as the "ANX" site), from December 2019 to May 2020 (Geerts et al., 2022). The campaign aimed to investigate the relationships between surface fluxes, boundary layer structure, aerosol properties, cloud and precipitation characteristics, and mesoscale circulations during cold-air outbreaks (CAOs) over open Arctic waters (Geerts et al., 2022). A key focus was to enhance the understanding of thermodynamic cloud phases and their evolution during CAOs. The deployment at the main site included all remote sensing measurements required to run the THER-

MOCLDPHASE VAP, as well as the input features needed for the ML models. However, MPL data were missing until 11 February 2020. Consequently, the THERMOCLDPHASE VAP between 11 February and 31 May 2020 was produced for this site shortly after the field campaign and has since been utilized in recent studies to analyze cloud-phase structures over the polar regions (Lackner et al., 2024; Van Weverberg et al., 2023; Xia and McFarquhar, 2024).

We evaluated the models' ability to classify thermodynamic cloud phases for a CAO event identified on 25 February 2020. Figure 9 presents thermodynamic cloud-phase classifications from the THERMOCLDPHASE VAP and the three ML model predictions and evaluations of ML model performance against the THERMOCLDPHASE VAP. Convective cloud structures and production of heavy snowfall during the CAO can be clearly observed from the time–height plot of thermodynamic cloud phases in Fig. 9a. ML model predictions compare well with those of the THERMOCLDPHASE VAP (Fig. 9b–d). A figure with the data streams used to create the VAP (similar to Fig. 1) is available in the Supplement (Fig. S4). All three models captured the time period accurately, with ice and snow dominating the ML-classified thermodynamic cloud phases. Interestingly, there are some "unknown" phase pixels at the beginning of the day from the THERMOCLDPHASE VAP, where the static algorithm was unable to resolve the cloud phase because the phase identification is inconsistent with our understanding of cloud physics based on past studies. Large Z_e and cold temperatures suggest that these pixels are snow, yet they exhibit falling velocities exceeding $2.5\,m\,s^{-1}$. Snow typically has low terminal velocities due to its small mass density and large surface area. However, during the CAO event's strong convective conditions, snow velocities may increase significantly in intense downdraft regions. The three ML models consistently predicted "snow" in this region, which is consistent with surrounding pixels, demonstrating an advantage of using ML models for cloud-phase classifications.

Both the CNN and the MLP have high confidence scores that are generally greater than 90 % for ice and snow pixels but significantly lower confidence scores for liquid- and mixed-phase pixels. Indeed, it is challenging to reliably distinguish liquid- and mixed-phase pixels from ice-phase pixels when they are embedded in ice-dominated clouds (Shupe, 2007; Silber et al., 2021). The lower model performance at ANX compared to at NSA is likely due to the more complex convective cloud structures associated with cold-air outbreaks (CAOs) at ANX (Geerts et al., 2022). The RF has lower confidence scores, except for ice-phase pixels at high altitudes after 12:00 UTC. The histogram plots in Fig. 9i–k show that all three ML models produce histograms that closely match the VAP, with the MLP and RF models slightly over-predicting the liquid category and under-predicting the ice category. The confusion matrices in Fig. 9l–n confirm that all three ML models predict the dominant ice and snow phases reasonably well, with accuracies exceeding 0.85. The three models all showed lower accuracy for the liquid phase (< 0.7), which is a minority category in this sample. In addition, both the MLP and the RF showed good predictions of the mixed-phase pixels, while the CNN showed a much lower accuracy in predicting mixed-phase pixels for this day. Overall, the CNN outperformed the MLP and RF models in terms of accuracy when predicting the dominant categories but performed worse than the other two models when predicting the minority categories.

Table 3. Model performance metrics for the three ML models on the dataset from COMBLE at ANX.

Model	Accuracy (%)	Precision*	Recall*	$F1$ score*	IOU*
CNN	92.5	0.841	0.777	0.805	0.69
MLP	80.4	0.684	0.827	0.725	0.594
RF	81.1	0.703	0.806	0.726	0.597

* Using a macro-average for each output class.

Model performance metrics for the entire study period in which the THERMOCLDPHASE VAP was produced at ANX are reported in Table 3. ANX plots, in the same format as those produced for NSA (Figs. 4, 5, and 6), are presented in Figs. S5, S6, and S7. Every performance metric using ANX as a test dataset (accuracy, precision, recall, $F1$ score, and IOU) is reduced in comparison to NSA (Table 3). The NSA test dataset comprised 12 months of data, and the ANX dataset comprised 4 months of data (February–May). Differences emerge when comparing the PDFs of confidence scores for the cloud-phase predictions for the three models. The CNN model behaved similarly at both sites, likely because the CNN incorporates information from neighboring pixels and because of the prevalence of ice at both locations, and for all phases predictions peaked at 100 % confidence (Fig. S5). The RF model also peaks at 100 % for all phases, except for liquid and liq_driz, which peak at 90 % and display a secondary local maximum at 40 %. The MLP diverges the most, with only the PDF of ice classification confidence peaking at 100 %. The PDFs for all other phases for the MLP model are more symmetrical and peak between 50 %–60 %. In addition, all three models reported higher false negatives for drizzle, liq_driz, and rain (Fig. S7). Comparing frequency distributions of cloud phases, ANX and NSA are similar as they are both high-latitude locations. Ice accounts for ~ 60 % of all cloud phases detected, followed by mixed, snow, and liquid (Fig. S6).

4 Data dropout experiment (improving threshold algorithm)

One advantage of using machine learning models for thermodynamic phase classification is that, unlike the VAP, they can still provide classifications in missing data scenarios. To

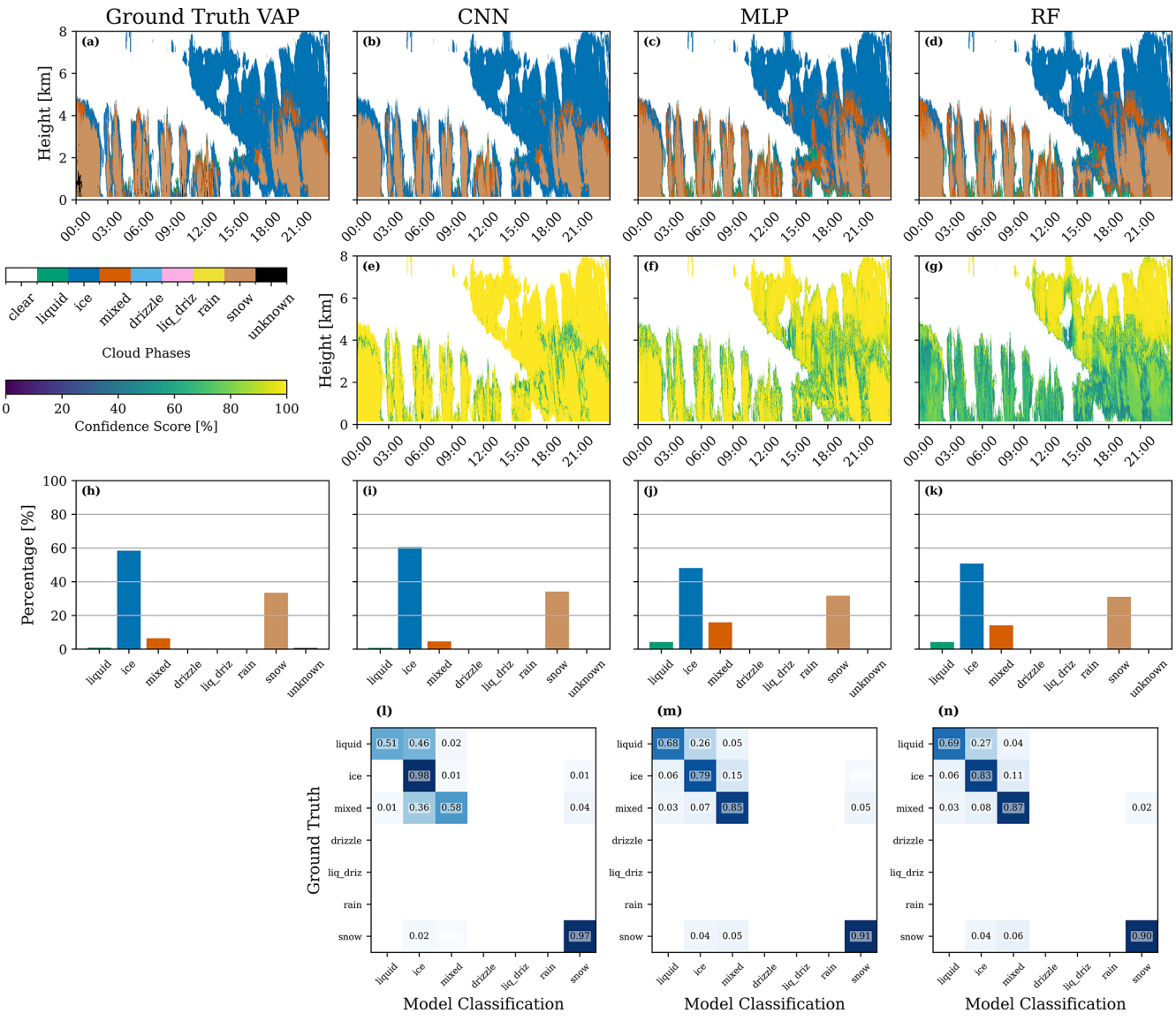

Figure 9. The results of applying the ML models to data collected at the ANX site in Norway on 25 February 2020 during the COMBLE campaign, with the same subplot structure as Fig. 3. The case day chosen is experiencing a CAO event. Note that at 00:00 UTC in panel **(a)**, the VAP has unknown pixels, which the ML models are able to resolve **(b, c, d)**.

assess model robustness against missing inputs, we tested our models by systematically removing either a single variable or all variables from a specific instrument to simulate scenarios where the instrument was offline. We also trained a variant of the U-Net designed to be resilient to missing data by including a layer to drop out random input channels with a likelihood of $p = 0.125$ during training, referred to as "CNN-ICD" (input channel dropouts). The CNN-ICD model was the second-best-performing CNN in the ablation study in Sect. 2.2.3, when all input channels were used, but the addition of the input channel dropout during training makes it far more robust in missing data scenarios.

We tested our models on a year's worth of data in 2021 at the NSA site. For each test, we evaluated the IOU score for each cloud-phase type over the year, the overall mean (with respect to phases) IOU score, and the total accuracy. Table 4 shows the results for the CNN-ICD model. Results for the other models are in the Supplement (Tables S2–S4, Fig. S8). The two instruments that had the greatest effect on accuracy were the radiosonde temperature and the radar data streams. For instance, for 2021, the accuracy of the CNN dropped from 95 % to 88 % without temperature data (mean IOU dropped from 0.81 to 0.37), and the accuracy of the RF dropped from 86 % to 74 % (IOU 0.72 to 0.28) (Tables S2 and S4). The CNN-ICD model, in comparison with temper-

Table 4. Performance of the CNN-ICD model in the data dropout study.

CNN-ICD model results		Intersection over union (IOU) score								
Model	Missing data stream/ instrument	Liquid	Ice	Mixed	Drizzle	Liquid drizzle	Rain	Snow	Mean IOU	Total accuracy (%)
CNN-ICD	Control	0.441	0.875	0.530	0.426	0.429	0.849	0.808	0.622	88.4
CNN-ICD	Micropulse lidar, all data streams	0.535	0.894	0.555	0.412	0.546	0.844	0.890	0.668	90.2
CNN-ICD	Micropulse lidar, backscatter	0.467	0.877	0.553	0.362	0.463	0.850	0.860	0.633	88.7
CNN-ICD	Micropulse lidar, linear depolarization ratio	0.469	0.877	0.508	0.407	0.448	0.841	0.819	0.624	88.6
CNN-ICD	Microwave, radiometer	0.436	0.876	0.533	0.438	0.440	0.850	0.802	0.625	88.5
CNN-ICD	Radar, all data streams	0.180	0.800	0.001	0.103	0.244	0.003	0.204	0.219	76.8
CNN-ICD	Radar, linear depolarization ratio	0.432	0.869	0.525	0.388	0.411	0.849	0.799	0.611	87.9
CNN-ICD	Radar, mean Doppler velocity	0.347	0.891	0.374	0.488	0.467	0.694	0.836	0.585	89.2
CNN-ICD	Radar, reflectivity	0.374	0.870	0.445	0.450	0.500	0.770	0.109	0.502	84.3
CNN-ICD	Radar, spectral width	0.459	0.879	0.470	0.600	0.316	0.802	0.873	0.629	88.9
CNN-ICD	Radiosonde, temperature	0.143	0.883	0.367	0.450	0.456	0.788	0.809	0.557	88.5

ature, dropped from 88 % to 85 % accuracy and 0.62 to 0.55 IOU, so while its control case performs worse, it is the least affected by data outages. It is also worthwhile to note how and where the cloud-phase classification failed without certain instruments. Dropping the MWR data had a minimal effect on model performance for all four models. However, without the radar mean Doppler velocity, the CNN, for example, had trouble distinguishing between rain and drizzle in liquid clouds. This is because Doppler velocity is key for determining whether a liquid particle is falling (Shupe, 2007). Another example is temperature, without which the model has trouble distinguishing solid from liquid water phases.

Table 4 shows that the CNN-ICD model performs well even with missing data, generally achieving a mean IOU > 0.5 and accuracy $> 75\,\%$. We hypothesize that with the addition of the two-dimensional dropout layers, which mimic instrument dropouts, it had greater elasticity to adapt to missing data and thus will be more robust to these events. When all input fields are available, we achieved the best results without the addition of these layers. Interestingly, in some cases the CNN-ICD model has greater accuracy and a better IOU score if some of the instrument data streams are missing, such as the linear depolarization ratio for the lidar and radar. This could indicate that some of the data streams give conflicting phase information or add input noise, in which case their inclusion actually makes the model less robust. We do not see this with the other models however.

Figure 10 demonstrates how each model responds to the absence of temperature data from the interpolated sonde on 15 August 2021 at the NSA site. These temperature data were identified as one of the most important input features for all the ML models in Fig. 8. On this day, deep clouds were observed at the beginning and end of the day and low-level clouds during the middle of the day. Due to elevated surface temperatures, the low-altitude clouds were predominantly composed of warm classes. This case serves as an excellent example for the data dropout experiment, as it includes all thermodynamic cloud phases. When all input features are available, the four ML models demonstrate strong performance compared to the THERMOCLDPHASE VAP (Fig. 10a–e). When temperature data are removed, all models show reduced performance (Fig. 10f–i), with the CNN-ICD model exhibiting the smallest reduction in performance. It accurately identifies mid- and high-level cloud phases but misclassifies liquid, drizzle, and rain as ice, mixed phase, and snow, particularly for low-altitude cloud pixels at the beginning and end of the day when temperature data are missing (Fig. 10f). Interestingly, the CNN-ICD model still correctly identifies low-altitude warm cloud classes between 03:00 and 20:00 UTC. The CNN, MLP, and RF models also correctly classify thermodynamic cloud phases for mid- and high-level

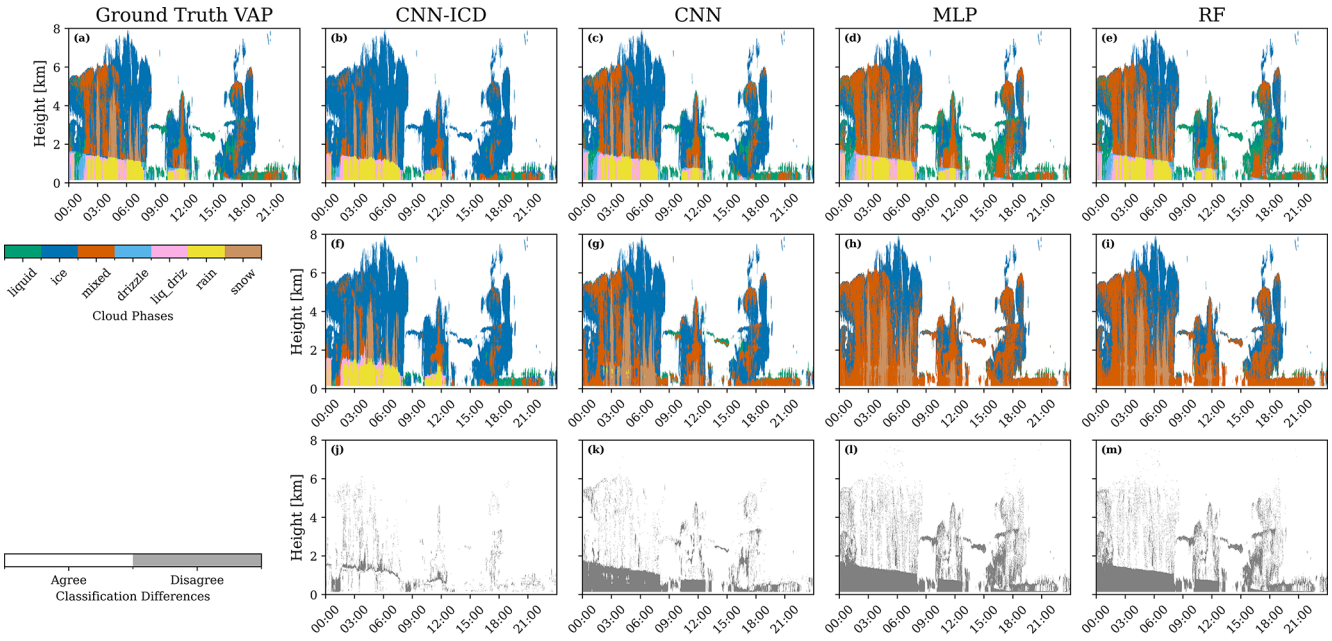

Figure 10. Example of how each model responds to missing temperature data from the interpolated sonde on 15 August 2021 at the NSA site. **(a–e)** Time–height plots of thermodynamic cloud phases from the THERMOCLDPHASE VAP (ground truth VAP), CNN-ICD, CNN, MLP, and RF, respectively; **(f–i)** thermodynamic cloud-phase classifications from the four ML models when temperature data are dropped out from the input features; and **(j–m)** the differences in thermodynamic cloud-phase classifications between model predictions with and without temperature data for the four ML models.

cloud pixels but frequently misclassify liquid, drizzle, and rain as ice, mixed phase, and snow for low-altitude cloud pixels throughout the day (Fig. 10g and h). The responses of each ML model to the removal of other input features are detailed in different rows in Figs. S9 and S10. Overall, the CNN-ICD model performs the best in the absence of data, followed by the CNN model and the MLP and RF models, which perform roughly equally.

Figure 11 shows how the CNN-ICD model responds to the removal of different variables for predicting thermodynamic cloud phases for the same case shown in Fig. 10. Consistent with the input feature analysis shown in Fig. 8, removing the MPL β, MPL dep, radar dep, LWP, and all MPL variables has a minimal impact on the performance of the CNN-ICD model. When Z_e is missing, the model sometimes fails to distinguish between liquid and drizzle for low-altitude cloudy pixels throughout the day and between ice and snow for mid- and high-level cloud pixels at the end of the day (Fig. 11c). Without radar W, the model sometimes fails to identify mixed-phase pixels for mid-level clouds, although they are only present for short periods in this example (Fig. 11e). Dropping out radar MDV causes the model to sometimes fail to distinguish between rain and drizzle between 03:00 and 06:00 UTC (Fig. 11f). Dropping out T causes the model to sometimes fail to distinguish between ice and liquid at the beginning of the day and between ice and drizzle at the end of the day (Fig. 11g). Overall, dropping out individual radar variables (including Z_e, MDV, W), all radar variables simultaneously, or temperature data had the largest effect on predicting thermodynamic cloud phases. This general result is also true for the other ML models for this case study, which are detailed in Figs. S9 and S10. This result shows general agreement with the feature importance results presented in Sect. 3.3.

5 Summary and conclusions

The ARM THERMOCLDPHASE VAP offers vertically resolved thermodynamic cloud-phase classifications using the multi-sensor approach developed by Shupe (2007), which combines lidar backscatter and depolarization ratio, radar reflectivity, Doppler velocity and spectra width, liquid water path, and temperature measurements. This study leveraged multiple years of the VAP product as the ground truth to train and evaluate three ML models for identifying thermodynamic cloud phases based on multi-sensor remote sensing data collected at the ARM NSA observatory. The models are an RF, an MLP, and a CNN with a U-Net architecture. Input features for the three ML models include MPL β and MPL dep, radar Z_e, MDV, W, and radar dep, MWR-derived LWP, and radiosonde T. An ablation study was conducted to find the optimal configuration of the CNN model. A total of 3 years of data at the ARM NSA site, from 2018–2020, were used for training and validation, while 1 year of data, from

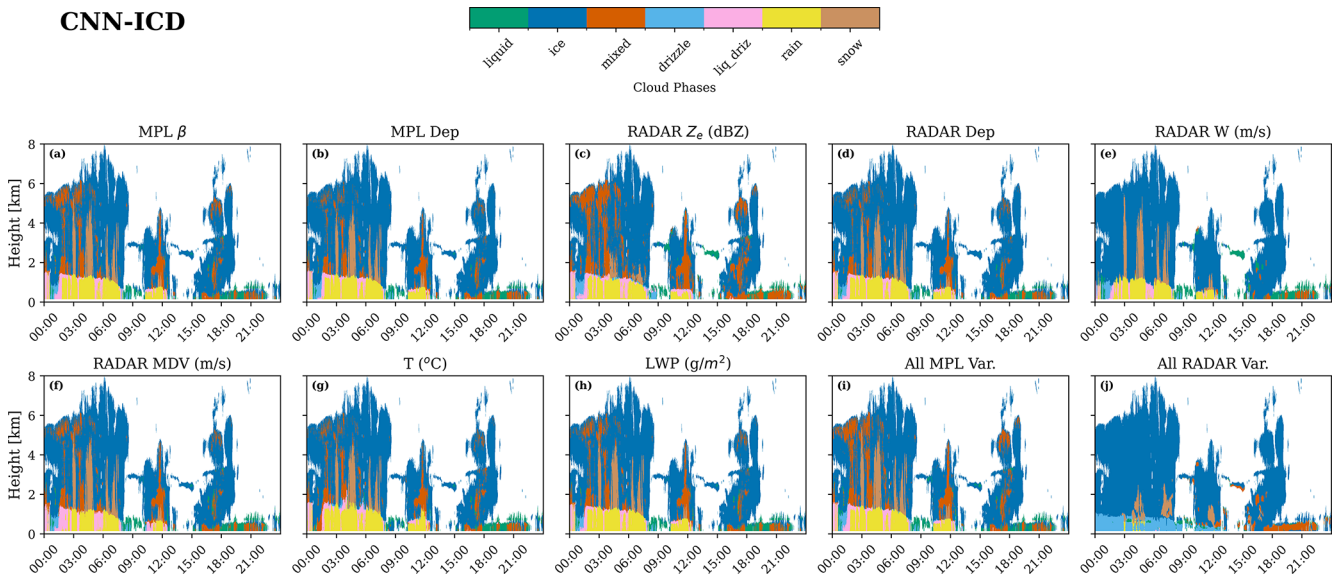

Figure 11. An example of how the CNN trained with input channel dropouts (CNN-ICD) responds to different missing input variables, mimicking data loss in the field for the same case shown in Fig. 10. The title of each panel shows the variable or all variables from a specific instrument that was dropped out. "All MPL Var" and "All Radar Var" represent all lidar variables and all radar variables that were dropped out, respectively.

2021, were used for testing. The input fields were organized as three-dimensional arrays (time × height × channel), with the channel dimension containing the nine individual ARM data stream inputs. The seven unique cloud-phase classifications produced by the THERMOCLDPHASE VAP were used as target variables.

The three trained ML models were applied to 1 year of multi-sensor remote sensing measurements from 2021 to predict the thermodynamic cloud phase (THERMOCLDPHASE-ML). The accuracy of these predictions was evaluated against the outputs of the THERMOCLDPHASE VAP. Evaluations included a detailed 1 d case study and year-long statistical assessment using performance metrics such as categorical accuracy, precision, recall, $F1$ score, and mean IOU. Among the ML models, the CNN demonstrated superior performance, achieving the highest categorical accuracy, $F1$ score, and mean IOU. This success is likely attributed to its holistic evaluation of input time–height arrays rather than the pixel-by-pixel approach used by the MLP and RF models. The CNN's success may also be due to site dependency, as NSA is ice dominated, and this model best predicts ice. The evaluations were further extended to data from an ARM AMF observatory during the ARM Cold-Air Outbreaks in the Marine Boundary Layer Experiment (COMBLE) field campaign at a coastal site in Andenes, Norway.

We also demonstrated three possible advantages of using ML models for thermodynamic cloud-phase classification, including the following:

1. ML models provide confidence scores for their predictions, with higher scores indicating greater certainty. Statistical analysis of 1 year of ML classification data reveals that predictions for ice, rain, and snow generally exhibit higher confidence scores, followed by the liquid phase. The mixed, drizzle, and liq_driz phases show lower confidence scores. Among the three ML models, the CNN produced the highest confidence scores across all thermodynamic cloud phases.

2. ML models enable feature importance analysis to identify the input features most influential in determining thermodynamic cloud phases. Analyzing the calculated permutation feature importance for the three ML models reveals that radar moments – specifically Z_e, MDV, and W – as well as temperature, are the most significant features for classifying thermodynamic cloud phases. In contrast, input features from lidar measurements and MWRRET LWP were found to be less influential.

3. ML models can predict thermodynamic cloud phases even when one or more input datasets are missing. To evaluate this capability, we conducted data dropout experiments by systematically removing either a single input variable or all variables from a specific instrument to simulate scenarios where the instrument was offline. We also trained a CNN U-Net model with input channel dropouts during training (referred to as CNN-ICD), hypothesizing that the inclusion of channel-wise dropout layers would mimic real instrument dropouts and enhance the model's ability to adapt to missing data, thus

making the model more robust. Overall, the CNN-ICD model performs better than the others when input fields are missing, followed by the standard CNN and MLP models, with the RF model performing the worst. Dropping out radar variables, including radar Z_e, MDV, and W and all of them together, as well as dropping out temperature data, had the largest negative impacts on predicting thermodynamic cloud phases for all models.

We utilized thermodynamic cloud-phase classifications from the THERMOCLDPHASE VAP as the ground truth. However, the VAP, which employs empirical threshold-based algorithms, can misclassify thermodynamic cloud phases (Shupe, 2007). Therefore, we do not expect the trained ML models to produce better thermodynamic cloud-phase classifications than the THERMOCLDPHASE VAP in most cases. Instead, we demonstrated the feasibility of using ML models to predict thermodynamic cloud-phase classifications with accuracy close to the VAP while adding additional information, such as confidence scores and feature importance. Furthermore, ML models can extend classification to scenarios where some instruments are offline, which are typically problematic for the VAP, and can produce reasonable classifications in specific cases when the VAP algorithm cannot. The ML models demonstrate elasticity in their ability to classify cloud phase, such as when the VAP was unable to classify snow in the COMBLE case study. Even so, we note that CNNs have limited interpretability and are less physics-informed than a hand-crafted retrieval. There are other more advanced segmentation algorithms than U-Nets that could be tested in future studies, e.g., U-Net++ (Zhou et al., 2018; King et al., 2024) and vision transformers (Springenberg et al., 2023). Furthermore, feature or saliency map analysis could offer valuable insights into whether the CNN focuses on physically meaningful regions of the data and represents a promising direction for future work to enhance the interpretability of model predictions and aid future model development (Haar et al., 2023). Our next step will involve creating a multiple-year, expert-labeled dataset of thermodynamic cloud phases to train ML models. The goal is to have an ML model that ultimately predicts better thermodynamic cloud phases than models derived from empirical threshold-based algorithms. It is important to note that the definition of thermodynamic phases depends on instrument sample volume and detection limit (Korolev and Milbrandt, 2022). The seven thermodynamic cloud-phase categories used in this study are empirical and might not precisely represent true thermodynamic cloud phases in nature. Therefore, we also plan to explore using unsupervised machine learning schemes for classifying thermodynamic cloud phases, using the THERMOCLDPHASE data only as a reference of comparison in future work.

Code availability. Code is currently hosted on GitHub: https://code.arm.gov/machine_learning/thermocldphase_amt, last access: 3 October 2025.

Data availability. The THERMOCLDPHASE VAP and all input data to the VAP and the ML models used in this study can be directly downloaded from the ARM Data Discovery website: https://doi.org/10.5439/2568095 (Goldberger et al., 2025).

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/amt-18-5393-2025-supplement.

Author contributions. Conceptualization, LG and DZ; methodology, LG, ML, AG, and DZ; software, LG, ML, and AG; validation, LG, MS, and DZ; formal analysis, LG, ML, AG, and DZ; investigation, LG and ML; resources, LG ML, AG, and DZ; data curation, LG and ML; writing – original draft preparation, LG and ML; writing – review and editing, all co-authors; visualization, LG and ML; supervision, DZ; project administration, DZ; funding acquisition, DZ. All authors have read and agreed to the published version of the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. Data were obtained from the ARM user facility, a US DOE Office of Science user facility managed by the Biological and Environmental Research (BER) program. This research was supported by the DOE ARM program. Matthew D. Shupe was supported by the DOE (DE-SC0021341), the NOAA Cooperative Agreement (NA22OAR4320151), and the NOAA Global Ocean Monitoring and Observing Program (fund ref. https://doi.org/10.13039/100018302).

Financial support. This research has been supported by the ARM user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research (BER) program (grant no. KP1704014/77783), the DOE Atmospheric System Research grant (grant no. DOE DE-SC0021341), the NOAA Cooperative Agreement (grant no. NA22OAR4320151), and the NOAA Global Ocean Monitoring and Observing Program (fund ref. https://doi.org/10.13039/100018302).

Review statement. This paper was edited by Gianfranco Vulpiani and reviewed by two anonymous referees.

References

Avery, M. A., Ryan, R. A., Getzewich, B. J., Vaughan, M. A., Winker, D. M., Hu, Y., Garnier, A., Pelon, J., and Verhappen, C. A.: CALIOP V4 cloud thermodynamic phase assignment and the impact of near-nadir viewing angles, Atmos. Meas. Tech., 13, 4539–4563, https://doi.org/10.5194/amt-13-4539-2020, 2020.

Balmes, K. A., Sedlar, J., Riihimaki, L. D., Olson, J. B., Turner, D. D., and Lantz, K.: Regime-Specific Cloud Vertical Overlap Characteristics From Radar and Lidar Observations at the ARM Sites, Journal of Geophysical Research: Atmospheres, 128, https://doi.org/10.1029/2022JD037772, 2023.

Barker, H. W., Korolev, A. V., Hudak, D. R., Strapp, J. W., Strawbridge, K. B., and Wolde, M.: A comparison between CloudSat and aircraft data for a multilayer, mixed phase cloud system during the Canadian CloudSat-CALIPSO Validation Project, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008JD009971, 2008.

Bishop, C. M.: Pattern recognition and machine learning, 4, 738 pp., New York, springer, 2006.

Brast, M., and Markmann, P.: Detecting the melting layer with a micro rain radar using a neural network approach, Atmos. Meas. Tech., 13, 6645–6656, https://doi.org/10.5194/amt-13-6645-2020, 2020.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cesana, G. and Chepfer, H.: How well do climate models simulate cloud vertical structure? A comparison between CALIPSO-GOCCP satellite observations and CMIP5 models, Geophysical Research Letters, 39, https://doi.org/10.1029/2012GL053153, 2012.

Cesana, G. and Chepfer, H.: Evaluation of the cloud thermodynamic phase in a climate model using CALIPSO-GOCCP, Journal of Geophysical Research: Atmospheres, 118, 7922–7937, https://doi.org/10.1002/jgrd.50376, 2013.

Cesana, G. V., Ackerman, A. S., Fridlind, A. M., Silber, I., Del Genio, A. D., Zelinka, M. D., Chepfer, H., Khadir, T., and Roehrig, R.: Observational constraint on a feedback from supercooled clouds reduces projected warming uncertainty, Communications Earth & Environment, 5, 181–181, https://doi.org/10.1038/s43247-024-01339-1, 2024.

Clothiaux, E., Miller, M., Perez, R., Turner, D., Moran, K., Martner, B., Ackerman, T., Mace, G., Marchand, R., Widener, K., Rodriguez, D., Uttal, T., Mather, J., Flynn, C., Gaustad, K., and Ermold, B.: The ARM Millimeter Wave Cloud Radars (MMCRs) and the Active Remote Sensing of Clouds (ARSCL) Value Added Product (VAP), https://doi.org/10.2172/1808567, 2001.

Chollet, F.: Keras, GitHub, https://github.com/fchollet/keras (last access: 15 October 2025), 2015.

Curry, J. A., Schramm, J. L., Rossow, W. B., and Randall, D.: Overview of Arctic Cloud and Radiation Characteristics, Journal of Climate, 9, 1731–1764, https://doi.org/10.1175/1520-0442(1996)009<1731:OOACAR>2.0.CO;2, 1996.

Fairless, T., Jensen, M., Zhou, A., and Giangrande, S.: Interpolated Sounding and Gridded Sounding Value-Added Products, https://doi.org/10.2172/1248938, 2021.

Fan, J., Ghan, S., Ovchinnikov, M., Liu, X., Rasch, P. J., and Korolev, A.: Representation of Arctic mixed-phase clouds and the Wegener-Bergeron-Findeisen process in climate models: Perspectives from a cloud-resolving study, Journal of Geophysical Research, 116, D00T07, https://doi.org/10.1029/2010JD015375, 2011.

Flynn, D., Cromwell, E., and Zhang, D.: Micropulse Lidar Cloud Mask Machine-Learning Value-Added Product Report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (US), https://doi.org/10.2172/1824785, 2023.

Galea, D., Ma, H.-Y., Wu, W.-Y., and Kobayashi, D.: Deep Learning Image Segmentation for Atmospheric Rivers, Artificial Intelligence for the Earth Systems, 3, https://doi.org/10.1175/AIES-D-23-0048.1, 2024.

Gaustad, K. L., Turner, D. D., and McFarlane, S. A.: MWRRET Value-Added Product: The Retrieval of Liquid Water Path and Precipitable Water Vapor from Microwave Radiometer (MWR) Data Sets (Revision 2), https://doi.org/10.2172/1019284, 2011.

Geerts, B., Giangrande, S. E., McFarquhar, G. M., Xue, L., Abel, S. J., Comstock, J. M., Crewell, S., DeMott, P. J., Ebell, K., Field, P., Hill, T. C. J., Hunzinger, A., Jensen, M. P., Johnson, K. L., Juliano, T. W., Kollias, P., Kosovic, B., Lackner, C., Luke, E., Lüpkes, C., Matthews, A. A., Neggers, R., Ovchinnikov, M., Powers, H., Shupe, M. D., Spengler, T., Swanson, B. E., Tjernström, M., Theisen, A. K., Wales, N. A., Wang, Y., Wendisch, M., and Wu, P.: The COMBLE Campaign: A Study of Marine Boundary Layer Clouds in Arctic Cold-Air Outbreaks, Bulletin of the American Meteorological Society, 103, E1371–E1389, https://doi.org/10.1175/BAMS-D-21-0044.1, 2022.

Geiss, A. and Hardin, J. C.: Inpainting radar missing data regions with deep learning, Atmos. Meas. Tech., 14, 7729–7747, https://doi.org/10.5194/amt-14-7729-2021, 2021.

Goldberger, L., Levin, M., Zhang, D., and Geiss, A.: Thermodynamic Cloud Phase Determination using Machine Learning, ARM Data Discovery [data set], https://doi.org/10.5439/2568095, 2025.

Haar, L. V., Elvira, T., and Ochoa, O.: An analysis of explainability methods for convolutional neural networks. Engineering Applications of Artificial Intelligence, 117, 105606, https://doi.org/10.1016/j.engappai.2022.105606, 2023.

Heaton, J.: Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning, Genet Program Evolvable Mach, 19, 305–307, https://doi.org/10.1007/s10710-017-9314-z, 2018.

Hogan, R. J., Illingworth, A. J., O'Connor, E. J., and Baptista, J. P. V. P.: Characteristics of mixed-phase clouds. II: A climatology from ground-based lidar, Quarterly Journal of the Royal Meteorological Society, 129, 2117–2134, https://doi.org/10.1256/qj.01.209, 2003.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W., and Wu, J.: UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, https://doi.org/10.1109/ICASSP40776.2020.9053405, 2020.

Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Pro-

ceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015, 448–456, 2015.

Kalesse, H., de Boer, G., Solomon, A., Oue, M., Ahlgrimm, M., Zhang, D., Shupe, M. D., Luke, E., and Protat, A.: Understanding Rapid Changes in Phase Partitioning between Cloud Liquid and Ice in Stratiform Mixed-Phase Clouds: An Arctic Case Study, Monthly Weather Review, 144, 4805–4826, https://doi.org/10.1175/MWR-D-16-0155.1, 2016.

Kalogeras, P., Battaglia, A., and Kollias, P.: Supercooled Liquid Water Detection Capabilities from Ka-Band Doppler Profiling Radars: Moment-Based Algorithm Formulation and Assessment, Remote Sens., 13, 2891, https://doi.org/10.3390/rs13152891, 2021.

Kay, J. E. and L'Ecuyer, T.: Observational constraints on Arctic Ocean clouds and radiative fluxes during the early 21st century, Journal of Geophysical Research: Atmospheres, 118, 7219–7236, https://doi.org/10.1002/jgrd.50489, 2013.

Kay, J. E., L'Ecuyer, T., Gettelman, A., Stephens, G., and O'Dell, C.: The contribution of cloud and radiation anomalies to the 2007 Arctic sea ice extent minimum, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL033451, 2008.

King, F., Pettersen, C., Fletcher, C. G., and Geiss, A.: Development of a Full-Scale Connected U-Net for Reflectivity Inpainting in Spaceborne Radar Blind Zones, Artif. Intell. Earth Syst., 3, e230063, https://doi.org/10.1175/AIES-D-23-0063.1, 2024.

Korolev, A. and Milbrandt, J.: How Are Mixed-Phase Clouds Mixed?, Geophysical Research Letters, 49, https://doi.org/10.1029/2022GL099578, 2022.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Communications of the ACM, 60, https://doi.org/10.1145/3065386, 2017.

Lackner, C. P., Geerts, B., Juliano, T. W., Kosovic, B., and Xue, L.: Characterizing Mesoscale Cellular Convection in Marine Cold Air Outbreaks With a Machine Learning Approach, Journal of Geophysical Research: Atmospheres, 129, https://doi.org/10.1029/2024JD041651, 2024.

Lagerquist, R., Turner, D. D., Ebert-Uphoff, I., and Stewart, J. Q.: Estimating Full Longwave and Shortwave Radiative Transfer with Neural Networks of Varying Complexity, Journal of Atmospheric and Oceanic Technology, 40, 1407–1432, https://doi.org/10.1175/JTECH-D-23-0012.1, 2023.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86, https://doi.org/10.1109/5.726791, 1998.

Lu, Y. and Kumar, J.: Convolutional Neural Networks for Hydrometeor Classification using Dual Polarization Doppler Radars, 2019 International Conference on Data Mining Workshops (ICDMW), 288–295, https://doi.org/10.1109/ICDMW.2019.00050, 2019.

Luke, E. P., Kollias, P., and Shupe, M. D.: Detection of supercooled liquid in mixed-phase clouds using radar Doppler spectra, J. Geophys. Res. Atmos., 115, D19201, https://doi.org/10.1029/2009JD012884, 2010.

McFarquhar, G. M., Ghan, S., Verlinde, J., Korolev, A., Strapp, J. W., Schmid, B., Tomlinson, J. M., Wolde, M., Brooks, S. D., Cziczo, D., Dubey, M. K., Fan, J., Flynn, C., Gultepe, I., Hubbe, J., Gilles, M. K., Laskin, A., Lawson, P., Leaitch, W. R., Liu, P., Liu, X., Lubin, D., Mazzoleni, C., Macdonald, A.-M., Moffet, R. C., Morrison, H., Ovchinnikov, M., Shupe, M. D., Turner, D. D., Xie, S., Zelenyuk, A., Bae, K., Freer, M., and Glen, A.: Indirect and Semi-direct Aerosol Campaign, Bulletin of the American Meteorological Society, 92, 183–201, https://doi.org/10.1175/2010BAMS2935.1, 2011.

Mülmenstädt, J., Sourdeval, O., Delanoë, J., and Quaas, J.: Frequency of Occurrence of Rain from Liquid-, Mixed-, and Ice-Phase Clouds Derived from A-Train Satellite Retrievals, Geophys. Res. Lett., 42, 6502–6509, https://doi.org/10.1002/2015GL064604, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830, https://doi.org/10.48550/arXiv.1201.0490, 2011.

Pithan, F., Medeiros, B., and Mauritsen, T.: Mixed-phase clouds cause climate model biases in Arctic wintertime temperature inversions, Clim. Dynam., 43, 289–303, https://doi.org/10.1007/s00382-013-1964-9, 2014.

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., and Riedi, J.: The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua, IEEE T. Geosci. Remote, 55, 502–525, https://doi.org/10.1109/TGRS.2016.2610522, 2017.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Sha, Y., Gagne, D. J., West, G., and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part i: Daily maximum and minimum 2-m temperature, Journal of Applied Meteorology and Climatology, 59, https://doi.org/10.1175/JAMC-D-20-0057.1, 2020.

Schimmel, W., Kalesse-Los, H., Maahn, M., Vogl, T., Foth, A., Garfias, P. S., and Seifert, P.: Identifying cloud droplets beyond lidar attenuation from vertically pointing cloud radar observations using artificial neural networks, Atmos. Meas. Tech., 15, 5343–5366, https://doi.org/10.5194/amt-15-5343-2022, 2022.

Shupe, M. D.: A ground-based multisensor cloud phase classifier, Geophysical Research Letters, 34, https://doi.org/10.1029/2007GL031008, 2007.

Shupe, M. D.: Clouds at Arctic Atmospheric Observatories. Part II: Thermodynamic Phase Characteristics, Journal of Applied Meteorology and Climatology, 50, 645–661, https://doi.org/10.1175/2010JAMC2468.1, 2011.

Shupe, M. D. and Intrieri, J. M.: Cloud Radiative Forcing of the Arctic Surface: The Influence of Cloud Properties, Surface Albedo, and Solar Zenith Angle, Journal of Climate, 17, 616–628, https://doi.org/10.1175/1520-0442(2004)017<0616:CRFOTA>2.0.CO;2, 2004.

Shupe, M. D., Turner, D. D., Zwink, A., Thieman, M. M., Mlawer, E. J., and Shippert, T.: Deriving Arctic Cloud Microphysics at Barrow, Alaska: Algorithms, Results, and Radiative Closure, Journal of Applied Meteorology and Climatology, 54, 1675–1689, https://doi.org/10.1175/JAMC-D-15-0054.1, 2015.

Shupe, M. D., Comstock, J. M., Turner, D. D., and Mace, G. G.: Cloud Property Retrievals in the ARM Program, Meteorological Monographs, 57, 19.11–19.20,

https://doi.org/10.1175/AMSMONOGRAPHS-D-15-0030.1, 2016.

Shupe, M. D., Walden, V. P., Eloranta, E., Uttal, T., Campbell, J. R., Starkweather, S. M., and Shiobara, M.: Clouds at Arctic Atmospheric Observatories. Part I: Occurrence and Macrophysical Properties, J. Appl. Meteorol. Clim., 50, 626–644,https://doi.org/10.1175/2010JAMC2467.1, 2011.

Silber, I., Fridlind, A. M., Verlinde, J., Ackerman, A. S., Cesana, G. V., and Knopf, D. A.: The prevalence of precipitation from polar supercooled clouds, Atmos. Chem. Phys., 21, 3949–3971, https://doi.org/10.5194/acp-21-3949-2021, 2021.

Silber, I., Verlinde, J., Eloranta, E. W., and Cadeddu, M.: Antarctic cloud macrophysical, thermodynamic phase, and atmospheric inversion coupling properties at McMurdo Station: I, Principal data processing and climatology, Journal of Geophysical Research: Atmospheres, 123, 6099–6121. https://doi.org/10.1029/2018JD028279, 2018.

Silber, I., Verlinde, J., Wen, G., and Eloranta, E. W.: Can Embedded Liquid Cloud Layer Volumes Be Classified in Polar Clouds Using a Single- Frequency Zenith-Pointing Radar? IEEE Geoscience and Remote Sensing Letters, 17, 222–226, https://doi.org/10.1109/LGRS.2019.2918727, 2020.

Solomon, A., Shupe, M. D., Persson, P. O. G., and Morrison, H.: Moisture and dynamical interactions maintaining decoupled Arctic mixed-phase stratocumulus in the presence of a humidity inversion, Atmos. Chem. Phys., 11, 10127–10148, https://doi.org/10.5194/acp-11-10127-2011, 2011.

Solomon, A., de Boer, G., Creamean, J. M., McComiskey, A., Shupe, M. D., Maahn, M., and Cox, C.: The relative impact of cloud condensation nuclei and ice nucleating particle concentrations on phase partitioning in Arctic mixed-phase stratocumulus clouds, Atmos. Chem. Phys., 18, 17047–17059, https://doi.org/10.5194/acp-18-17047-2018, 2018.

Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., and Strodthoff, N.: From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology, Medical Image Analysis, 87, https://doi.org/10.1016/j.media.2023.102809, 2023.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research, 15, 1929–1958, https://doi.org/10.5555/2627435.2670313, 2014.

Storelvmo, T., and Tan, I.: The Wegener-Bergeron-Findeisen Process – Its Discovery and Vital Importance for Weather and Climate, Meteorol. Z., 24, 455–461, https://doi.org/10.1127/metz/2015/0626, 2015.

Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., and Hamarneh, G.: Combo Loss: Handling Input and Output Imbalance in Multi-Organ Segmentation. Computerized Medical Imaging and Graphics, 75, 24–33, https://doi.org/10.1016/j.compmedimag.2019.04.005, 2019.

Tan, I., Storelvmo, T., and Zelinka, M. D.: Observational constraints on mixed-phase clouds imply higher climate sensitivity, Science, 352, 224–227, https://doi.org/10.1126/science.aad5300, 2016.

Tan, I., Zhou, C., Lamy, A., and Stauffer, C. L.: Moderate climate sensitivity due to opposing mixed-phase cloud feedbacks, Clim. Atmos. Sci., 8, 86, https://doi.org/10.1038/s41612-025-00948-7, 2025.

Turner, D. D., Ackerman, S. A., Baum, B. A., Revercomb, H. E., and Yang, P.: Cloud Phase Determination Using Ground-Based AERI Observations at SHEBA, Journal of Applied Meteorology, 42, 701–715, https://doi.org/10.1175/1520-0450(2003)042<0701:CPDUGA>2.0.CO;2, 2003.

Van Weverberg, K., Giangrande, S., Zhang, D., Morcrette, C. J., and Field, P. R.: On the Role of Macrophysics and Microphysics in Km-Scale Simulations of Mixed-Phase Clouds During Cold Air Outbreaks, Journal of Geophysical Research: Atmospheres, 128, https://doi.org/10.1029/2022JD037854, 2023.

Veillette, M. S., Kurdzo, J. M., Stepanian, P. M., McDonald, J., Samsi, S., and Cho, J. Y. N.: A Deep Learning–Based Velocity Dealiasing Algorithm Derived from the WSR-88D Open Radar Product Generator, Artificial Intelligence for the Earth Systems, 2, https://doi.org/10.1175/AIES-D-22-0084.1, 2023.

Verlinde, J., Harrington, J. Y., McFarquhar, G. M., Yannuzzi, V. T., Avramov, A., Greenberg, S., Johnson, N., Zhang, G., Poellot, M. R., Mather, J. H., Turner, D. D., Eloranta, E. W., Zak, B. D., Prenni, A. J., Daniel, J. S., Kok, G. L., Tobin, D. C., Holz, R., Sassen, K., Spangenberg, D., Minnis, P., Tooman, T. P., Ivey, M. D., Richardson, S. J., Bahrmann, C. P., Shupe, M., DeMott, P. J., Heymsfield, A. J., and Schofield, R.: The Mixed-Phase Arctic Cloud Experiment, Bulletin of the American Meteorological Society, 88, 205–222, https://doi.org/10.1175/BAMS-88-2-205, 2007.

Verlinde, J., Zak, B. D., Shupe, M. D., Ivey, M. D., and Stamnes, K.: The ARM North Slope of Alaska (NSA) Sites, Meteorological Monographs, 57, 8.1–8.13, https://doi.org/10.1175/AMSMONOGRAPHS-D-15-0023.1, 2016.

Wang, Z., and Sassen, K.: Cloud Type and Macrophysical Property Retrieval Using Multiple Remote Sensors. Journal of Applied Meteorology, 40, 1665–1682, https://doi.org/10.1175/1520-0450(2001)040<1665:CTAMPR>2.0.CO;2, 2001.

Wendisch, M., Macke, A., Ehrlich, A., Lüpkes, C., Mech, M., Chechin, D., Dethloff, K., Velasco, C. B., Bozem, H., Brückner, M., Clemen, H.-C., Crewell, S., Donth, T., Dupuy, R., Ebell, K., Egerer, U., Engelmann, R., Engler, C., Eppers, O., Gehrmann, M., Gong, X., Gottschalk, M., Gourbeyre, C., Griesche, H., Hartmann, J., Hartmann, M., Heinold, B., Herber, A., Herrmann, H., Heygster, G., Hoor, P., Jafariserajehlou, S., Jäkel, E., Järvinen, E., Jourdan, O., Kästner, U., Kecorius, S., Knudsen, E. M., Köllner, F., Kretzschmar, J., Lelli, L., Leroy, D., Maturilli, M., Mei, L., Mertes, S., Mioche, G., Neuber, R., Nicolaus, M., Nomokonova, T., Notholt, J., Palm, M., van Pinxteren, M., Quaas, J., Richter, P., Ruiz-Donoso, E., Schäfer, M., Schmieder, K., Schnaiter, M., Schneider, J., Schwarzenböck, A., Seifert, P., Shupe, M. D., Siebert, H., Spreen, G., Stapf, J., Stratmann, F., Vogl, T., Welti, A., Wex, H., Wiedensohler, A., Zanatta, M., and Zeppenfeld, S.: The Arctic Cloud Puzzle: Using ACLOUD/PASCAL Multiplatform Observations to Unravel the Role of Clouds and Aerosol Particles in Arctic Amplification, Bulletin of the American Meteorological Society, 100, 841–871, https://doi.org/10.1175/BAMS-D-18-0072.1, 2019.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2021MS002502, 2021.

Wieland, M., Li, Y., and Martinis, S.: Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network, Remote Sensing of Environment, 230, https://doi.org/10.1016/j.rse.2019.05.022, 2019.

Xia, Z. and McFarquhar, G. M.: Dependence of Cloud Macrophysical Properties and Phase Distributions on Environmental Conditions Over the North Atlantic and Southern Ocean: Results From COMBLE and MARCUS, Journal of Geophysical Research: Atmospheres, 129, https://doi.org/10.1029/2023JD039869, 2024.

Xie, Y., King, F., Pettersen, C., and Flanner, M.: Machine learning detection of melting layers from radar observations, Journal of Geophysical Research: Machine Learning and Computation, 2, e2024JH000521, https://doi.org/10.1029/2024JH000521, 2025.

Yu, G., Verlinde, J., Clothiaux, E. E., and Chen, Y.-S.: Mixed-phase cloud phase partitioning using millimeter wavelength cloud radar Doppler velocity spectra, J. Geophys. Res.-Atmos., 119, 7556–7576, https://doi.org/10.1002/2013JD021182, 2014.

Zhang, D., and Levin, M.: Thermodynamic cloud phase (THERMOCLDPHASE), 2017-03-01 to 2024-07-01, North Slope Alaska (NSA), Central Facility, Barrow AK (C1), Atmospheric Radiation Measurement (ARM) User Facility [data set], https://doi.org/10.5439/1871014, 2024.

Zhang, D., Wang, Z., and Liu, D.: A global view of midlevel liquid-layer topped stratiform cloud distribution and phase partition from CALIPSO and CloudSat measurements, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012143, 2010.

Zhang, D., Wang, Z., Luo, T., Yin, Y., and Flynn, C.: The occurrence of ice production in slightly supercooled Arctic stratiform clouds as observed by ground-based remote sensors at the ARM NSA site, Journal of Geophysical Research: Atmospheres, 122, 2867–2877, https://doi.org/10.1002/2016JD026226, 2017.

Zheng, X., Tao, C., Zhang, C., Xie, S., Zhang, Y., Xi, B., and Dong, X.: Assessment of CMIP5 and CMIP6 AMIP Simulated Clouds and Surface Shortwave Radiation Using ARM Observations over Different Climate Regions, Journal of Climate, 36, 8475–8495, https://doi.org/10.1175/JCLI-D-23-0247.1, 2023.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham, 2018, 3–11, https://doi.org/10.1007/978-3-030-00889-5_1, 2018.