Reply to Reviewer #2's comments

Goldberger et al. provides a detailed analysis of cloud phase classification using a suite of increasingly sophisticated machine learning (ML) models. The authors find that the traditional threshold-based approach (i.e., the THERMOCLDPHASE VAP) was limited in its ability to classify phase across varying regional climates due to the use of static thresholds, and lacked a general robustness when some inputs were missing (e.g. due to instrument outages). Shifting to an ML-based approach improved classification accuracy at two locations, and provided a framework for more easily understanding variable importance and uncertainty quantification in their predictions. The paper is well structured, and includes many of the minor, yet fundamental, points (e.g.., information about data standardization, splitting, model complexity) that are often missed in these types of manuscripts/projects. I also appreciated the additional work provided in the Supplement detailing the hyperparameter search space and dropout tests. I don't see any critical methodological issues here, the content aligns well with the journal's mission, and I believe this paper would be of great interest to the readers of Atmospheric Measurement Techniques (AMT). However, I do have a few minor points below that I'd like to see addressed on before I fully recommend the paper for publication.

Answer: We appreciate the reviewer's valuable suggestions and comments. We have carefully revised the manuscript in accordance with the feedback.

General Comments:

1. *While the performance of the ML models across both sites were mostly consistent, it was clear that the CNN really struggled with liquid/mixed-phase at ANX compared to NSA. Do the authors have more of a physical basis for this drop in accuracy? You discuss this issue briefly in 451-462, but I am curious if the regional cloud structures are different enough in Norway to cause the CNN to struggle for this very important classification category. On a similar note, while performance for other classes remained highest in the CNN, I found the MLP to often provide a consistent performance more generally across all classes (not only in Fig. 9, but also Fig. 6). For instance, in Fig. 6, the CNN really excels at ice-based classifications, but it is slightly worse in nearly every other class compared to the MLP/RF. Some may argue that a less complex model that is more consistent and quicker to train would be preferred compared to the CNN, even with worse performance in some categories.*

Answer: We thank the reviewer for these insightful comments. We agree that the regional differences in cloud structures likely play a key role in the CNN's reduced performance at ANX, particularly for the liquid and mixed-phase classifications. We attribute this degradation primarily to the more complex and vertically developed convective cloud structures observed during cold-air outbreaks (CAOs) at ANX, which pose greater challenges for classification. A sentence has been added between lines 488 and 489 to clarify this point.

As noted in lines 265–278, the CNN's strong performance in ice-phase classification is largely due to imbalanced training data, where ice-phase pixels dominate. We also discussed the challenges of applying balanced training data in CNNs. Additionally, in lines 371–377, we examined how imbalanced training data affect the performance of RF and MLP models.

2. *General comments about figure quality. As this is a visual project, I feel that the visual comparisons between methods are key for demonstrating performance differences to the reader. For instance, on Fig. 3, could you not zoom in more on the actual cloud structures? The plots extend to 12 km currently, but there is little-to-no activity above say 7 km (and similar for Fig. 9/10/11) which leaves a lot of blank space. Additionally, I like the point being made by Fig. 4, but it is quite challenging to read in its current state. I would make this a 4x2 plot instead of 2x4 and try to make the bars wider. I recognize that the versions of these images I am seeing is also compressed, but to highlight fine structural details, I'd recommend the authors make sure their resolution is as high as is allowed by the journal so readers can zoom in to see the interesting structural details better.*

Answer: We appreciate the reviewer's suggestions for improving the figure quality. In response, we limited the vertical axis to 8–9 km, just above the top of the highest cloud, in Figures 3, 9, 10, and 11. Additionally, we revised Figures 4 and S5 to a 4×2 layout as recommended. We will upload the highest-resolution versions of all figures permitted by the publication guidelines.

Specific Comments:

- *Lines 18-19: I like the evolution from decision-tree based methods to more advanced NNs, but was curious if you had considered vision transformers for this problem? They've seen quite a surge in popularity over the past few years especially in areas like remote sensing (e.g., Lenhardt et al., 2024 for clouds, or Thisanke et al., 2023 more generally), and while likely beyond the scope of this project, might be of interest down the road as they can pick up on global features often missed by the U-Net.*

Answer: We thank the reviewer for highlighting these recent studies that apply advanced neural networks. As noted in lines 648–650, we mentioned that more advanced segmentation algorithms, including vision transformers, could be explored in future work. We have now added the suggested recent studies to the manuscript as recommended.

- *Lines 177: I appreciate the authors providing a list of RF hyperparameters, and I was curious about using 20 as the maximum depth here. 20 is a fairly deep tree for 100k samples and I am wondering how sensitive the model is to changing this value to something smaller (10-15).*

Answer: 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. We have added a sentence in lines 190–192 to note this.

- *Line 185: Interesting, I've never seen someone use the scikit implementation to train an MLP (usually torch or tensorflow)! I'm guessing you used one of these for the CNN?*

Answer: We appreciate the reviewer's observation. We have updated the text to clarify that TensorFlow, through the Keras interface, was used for the CNN implementation.

- *Lines 202-203: How are missing pixels handled before being ingested by the CNN? Like what if you have to QC a small patch of bad radar data and mark it as missing, is that whole scene dropped?*

Answer: Missing instrument data values are filled with a value of –9999 prior to dataset normalization, as specified in Table 1, which results in clipping to the bottom of their valid input range. 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. We added a sentence in line 217-220 to clarify this.

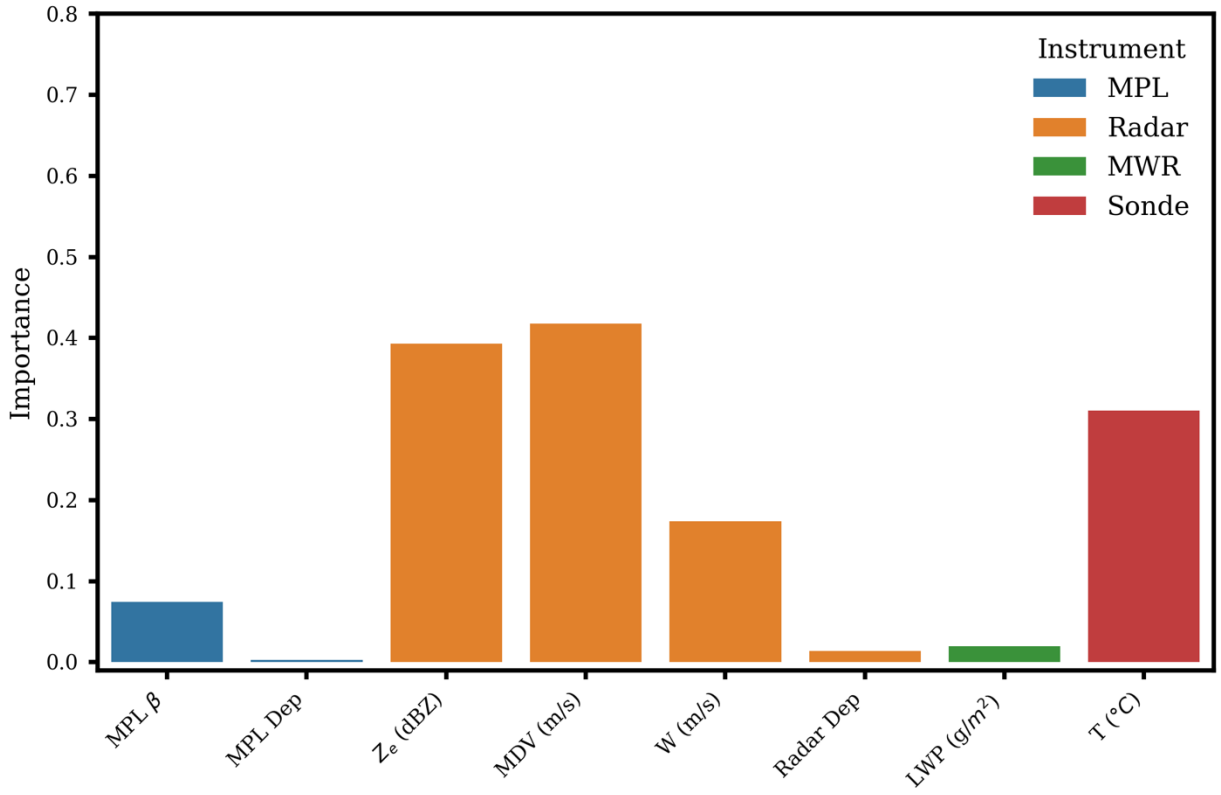
- *Figure 3: Zooming in on e-g of Fig. 3, we can see the melting layer is one of the most challenging areas to accurately predict for the CNN, with low confidence throughout. As this is an area where change is quickly happening to particle phase, shape and fallspeed, the CNN appears to struggle in resolving these fine scale details that are quite important (the presence, location, and width of the ML are often some of the most important features we would like to accurately classify for looking at regime shifts). Do you have any insights on how to improve this detection here? There have been some projects using ML to better predict this region (e.g., Brast and Markmann, 2020), but it remains an active area of research.*

Answer: We appreciate the reviewer's thoughtful observation regarding the CNN's difficulty in accurately classifying the melting layer. We agree that this region—where rapid transitions in particle phase, shape, and fall speed occur—is particularly important and challenging to resolve. In our case, the limitations of the THERMOCLDPHASE VAP, including its coarse vertical resolution and reliance on interpolated sounding data, contribute to the reduced model confidence in this layer. As a result, all ML models show diminished performance in this region. We concur that improving detection in the melting layer will likely require a refined training dataset that specifically targets this transitional zone. As the reviewer noted, this remains an active area of research. We have added a sentence between lines 303-307 to reflect this discussion.

- *Figure 8: While I think this is a useful figure, it is a bit busy with how it is currently set up and takes a moment to digest (variable by phase by model). Do you have an overall feature importance chart anywhere to give a general overview of each input's importance across all phase types?*

Answer: We appreciate the reviewer's helpful observations regarding Figure 8. We agree that the figure is somewhat busy and may take a moment to interpret. We have updated the figure to improve visual separation of the features and to make the feature labels clearer.

Because of the cold temperatures at NSA, warm cloud phase categories represent extreme minority classes. As a result, computing overall feature importance based on the change in multi-class recall from input permutations would largely reflect the importance for the dominant ice category. Given this class imbalance, it is important to analyze feature importance separately for each phase. For this reason, we believe it remains valuable to present phase-specific results in Figure 8. For the reviewer's reference, we have also provided a figure below showing the average permutation importance for each feature in the CNN model, calculated as the mean across phase-specific importances.



- *Table 4: I see in Table 4 that you experiment with leaving out different variables, but did you perform retraining on these models without the combination of MPL/LWP variables which display little relevance in Fig. 8? It would appear as though they don't provide much useful context beyond what is given in the other variables.*

Answer: As noted in lines 424–426, "input features from lidar measurements (panels a, e, and i) and the MWRRET LWP (panels c, g, and k) are less influential, likely because lidar signals are often attenuated by persistent low-level clouds at the NSA site (Shupe et al., 2011; Zhang et al., 2017), and LWP provides only column-integrated information rather than detailed vertical profiles." While lidar backscatter and depolarization ratio provide direct and reliable indicators of liquid-phase presence, radar-based variables—such as reflectivity, mean Doppler velocity, and spectral width—also contain useful signatures of liquid-phase clouds. However, in situations where radar data are missing, MPL and LWP can become critical for

cloud phase classification. Moreover, as shown in Figure 8, including MPL and LWP does contribute to improving model predictions in certain scenarios. Therefore, we did not excluded these variables.

- *Section 3.3: While not critical to the current manuscript, did you consider looking at feature/saliency maps in the CNN (e.g., Haar et al., 2023)? As a vision model, this can sometimes be useful for sanity checks that the model is looking at regions in the data that we would expect from a physical standpoint as being relevant (e.g. cloud edges, hydrometeor gradients, gaps etc.).*

Answer: We appreciate the reviewer's valuable suggestion. We agree that incorporating feature or saliency map analysis is a promising avenue for future work to enhance the interpretability of model predictions. We have added a sentence and the suggested reference in lines 648–650, which reads: "*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 (Haar et al., 2023).*" In addition to verifying that the model is looking at appropriate regions in the input data, this type of analysis could also help determine how much spatial information the model is using and how large the perceptive field needs to be, which could aid in future model development.

Section 3.4: It is great to see the model also tested at an independent site. I touched on this in my general comment above, but how well do you believe this model could generalize to other regional climates? For instance, other DOE ARM sites like SBS/SGP/ENA or even AWR with similar instrument setups?

Answer: As noted by Shupe (2007), the multi-sensor cloud thermodynamic phase classification method "has been specifically developed for observations of Arctic clouds." Accordingly, the THERMOCLDPHASE VAP is currently applied only at the seven ARM high-latitude observatories including AWR. We have revised the sentence on line 57 to read: "as well as six other ARM high-latitude observatories." Additionally, 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. To improve clarity, we added the following sentence between lines 57 and 60: "It is noted that the multi-sensor cloud thermodynamic phase classification was specifically developed for observations of Arctic clouds (Shupe 2007). Since the algorithm does not include the classification of hail and graupel, it has difficulty distinguishing these hydrometeor types in deep convective cloud regimes over tropical and mid-latitude regions.".

- *Line 459: This should be CNN not CCN I believe.*

Answer: We changed 'CCN' to 'CNN' in the manuscript.

- *Line 602: Advanced U-Net models like these have already started to be used in recent literature at other DOE sites looking at clouds using radar data (e.g., King et al., 2025).*

Answer: We appreciate the reviewer for pointing out the recent literature. We added the reference in line 647.

- *Line 612: The GitHub repo link appears to be broken here. It would be great to also have a Google Colab notebook available to reproduce some of the model results if possible.*

Answer: We appreciate the reviewer's suggestions for reproducibility of our results. We have fixed the link and made the GitHub repo public. The data are too large for Google Colab, but we have included code to download the data from ARM and all of the code used to generate results and figures for the manuscript.

References

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.

Haar, L. V., Elvira, T., & Ochoa, O. (2023). 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

King, F., C. Pettersen, C. G. Fletcher, and A. Geiss, 2024: 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.

Lenhardt, J., Quaas, J., Sejdinovic, D., and Klocke, D.: CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-2724, 2024.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., & Herath, D. (2023). Semantic segmentation using Vision Transformers: A survey. Engineering Applications of Artificial Intelligence, 126, 106669. https://doi.org/10.1016/j.engappai.2023.106669