

Thank you to the reviewers for comments and suggestions, your input has improved our paper. Our responses are in blue italics.

Response to reviewer 1: pages 1-6

Response to reviewer 2: pages 7-21

There have been several major changes I would like to summarise here:

- 1. The paper has been majorly restructured, where the results and discussion sections have been merged. Please see pages 10 and 11 for details.*
- 2. The scope of paper has been narrowed. Evaluation and conclusions are focused on habits that are dominant in the DCMEX campaign.*
- 3. One person evaluation of the CNNs. Images used to test CNN have now only been labelled by one person. In the latest version of the manuscript figures 2, 3, 5 and 6 have been updated to reflect this (new figures are on pages 22-27). All references to values from these figures have also been updated. Overarching results and conclusions have not changed.*

Review 1

Review: Evaluation and application of a convolutional neural network for graupel identification in DCMEX deep convective cloud

Ezri E. Alkilani-Brown, Declan L. Finney, Alan M. Blyth, Paul R. Field, Jonathan Crosier, and Chetan R. Deva

A) Manuscript Summary

This study by Alkilani-Brown et al. presents an independent evaluation and application of open-source Convolutional Neural Network (CNN) models developed by Jaffeux et al. (2022, 2025) for hydrometeor classification from Optical Array Probe (OAP) images. The evaluation uses two-dimensional stereo (2D-S) and High Volume Particle Spectrometer (HVPS) images collected during the Deep Convective Microphysics EXperiment (DCMEX) campaign over New Mexico in summer 2022. A subset of images was manually labelled by multiple human annotators to construct a consensus ground-truth, against which three CNN models (global, 2D-S-specific, and HVPS-specific) were evaluated. Performance was assessed using accuracy, precision, recall, and F1 score. The authors then scale up CNN application to the full DCMEX dataset to characterise ice particle habits and infer graupel concentrations, introducing a correction factor to address systematic CP-to-HPC misclassification.

The study aims to address a useful community need: an independent evaluation of publicly available classification tools on a dataset from a different institution without re-tuning. The resulting manual labelling disagreement analysis (Figure 1) provides valuable insight into the inherent ambiguity of OAP image interpretation and the porosity of class boundaries. Combined with a random-labelling control experiment, these results constitute a genuine characterisation of the morphological composition of the DCMEX dataset and of human perception of ice particle habits within it. The graupel approximation is practically valuable, and the resulting campaign-wide habit distributions are physically interpretable and represent a first characterisation of DCMEX ice microphysics.

The main weakness of the paper is a mismatch between the methodology used and what the paper claims to deliver. The random extraction of images for evaluation produces a test set that reflects the heavily skewed class distribution of the DCMEX dataset, which is largely dominated by rimed particles. This choice is well-suited for the characterisation of the dataset and specific human labelling uncertainty. However, it is insufficient to rigorously evaluate classification models themselves: classes being represented by very few images make per-class metrics statistically weak, and aggregate scores are dominated almost entirely by CP performance. The existing results should consequently be reframed as a dataset and human perception characterisation, and complemented by a class-balanced evaluation, ideally following the methodology of Jaffeux et al. (2025), who used 200 images per class per instrument, to substantiate conclusions about the skill of CNN models. Although the study demonstrates the discrepancies between the class definitions and the content of the dataset used to train the models, it falls short of translating its own findings into constructive community-level recommendations, such as the formalisation of a consensus training dataset through existing international expert gatherings, such as the Cloud Probe Workshops (latest publication McFarquhar et al. 2017).

B) Recommendation

This paper makes a welcome contribution to the ice microphysics in-situ community by providing an independent assessment of the Jaffeux et al. (2025) CNN tools on a new campaign dataset, and by delivering the first characterisation of ice habit composition for the DCMEX campaign. The human labelling effort and disagreement analysis are valuable. However, the central claim that the CNN models have been evaluated is not adequately supported by the methodology as currently presented. The random sampling approach is valuable for dataset characterisation, but on its own it is not sufficient for a fully rigorous model-performance assessment across all classes. Furthermore, the conclusions could be substantially improved to benefit the OAP user community. These issues can be addressed by complementing the existing results with a class-focused evaluation for sufficiently populated classes. If this is not feasible, the paper should still be publishable after major revision provided the claims are correspondingly narrowed and the scope is clearly reframed.

Recommended decision: Major revisions

C) Major Issues

Major Issue 1: The random sampling methodology characterises the dataset well but is not sufficient on its own for a rigorous CNN performance evaluation

The current methodology constitutes an extensive evaluation that describes the dataset and human and CNN perception of it comprehensively, but what is additionally required is an intensive evaluation, targeting model output quality for each class specifically and in sufficient numbers to draw statistically meaningful conclusions. The random extraction of images produces a test set that reflects the inherent class distribution of DCMEX, and this is genuinely valuable for understanding the morphological composition of the campaign dataset. However, it is not adequate for characterising the performance of the CNN models. With CP comprising ~83% of 2D-S images, minority classes such as CBC, CA, WD, CC, and HPC are represented by as few as 1 or 2 images in the test set. Per-class metrics for very sparsely represented classes have very limited

interpretability (see Minor Issue 1). The evaluation as currently presented does not clearly test whether the models can reliably distinguish between classes, which is the central question for such a classification evaluation study.

I strongly advise the authors to complement their existing results with a class-specific evaluation. The easier and directly applicable approach is to adapt the methodology of Jaffeux et al. (2025, Section 2.3), who constructed evaluation sets of 200 images per class per CNN, extracted randomly from the classified datasets. This would allow a genuine per-class characterisation of model skill and produce results directly comparable to the original model article. The class-balanced set would then serve as the primary basis for conclusions about CNN performance.

If such a class-focused evaluation cannot be completed for some classes due to insufficient image availability, the manuscript should explicitly narrow its claims from "CNN evaluation" to a more limited assessment (e.g., "evaluation under DCMEX conditions" or "campaign-specific application with partial evaluation"). In that case, conclusions about model skill should be restricted to classes with adequate occurrence and to aggregate behavior, with clear statements that minority-class skill remains unresolved.

Relevant sections: Section 2.2.3, Sections 3.2.1–3.2.2, Section 4.2, and the conclusions (Section 5).

The nature of DCMEX means we have disproportionately sampled a select number of habits – CP / FA / Co and a couple others. We agree equal size classes for evaluation would be best for true evaluation of Jaffeux’s models. However to obtain classes of equal sizes, for example, currently we have labelled 31 images as CA (from our 10,000 2D-S sample), to get to a sample of 200 images (assuming the rest of the population is proportional in type of particles present), we would have to label approximately 54,000 more images. This is not practical at this stage of the analysis and have decided the following:

- *We have narrowed the scope of the paper, testing only the dominant habit categories from DCMEX dataset, and made this result clear throughout.*
 - *L:13-15: “The models constructed by Jaffeux et al. (2025) present an accessible and adjustable approach to identifying particles in optical array probe images from deep convective clouds.”*
 - *L258-259: “Ultimately, for the identification of deep convective 2D-S sized ice crystals, namely CP, FA and Co classes, the global model provides a satisfactory result in the classification of images”*
 - *L295-298 “Here we have tested the global CNN and 2D-S / HVPS specific CNNs on images from a field campaign targeting deep convective cloud, namely CP, FA and Co classes. Overall, the global model was successful at identifying both 2D-S and HVPS images, while the HVPS-specific model achieved an equivalent result on HVPS images.”*
 - *L393-396: “Unseen 2D-S and HVPS images from the DCMEX campaign evaluated the global, 2D-S-specific and HVPS-specific CNNs. The DCMEX campaign specifically targeted growing cumulonimbus turrets where riming dominates, this has strongly impacted the types of particles*

observed. Specifically, the dataset was dominated by rimed particles, aggregates and columns, thus these habits are used to evaluate the CNNs.”

- *L397: “For 2D-S images, the global CNN was best and is suitable for identifying particles imaged in deep convective clouds.”*
- *L403-406: “The HVPS-specific model attained a weighted F1 score of 73%, while the global model attained 69%. As values are very close, both models are suitable for particle identification of HVPS images from deep convective clouds.”*
- *L422-424: “From our evaluation, the predictions produced from the global (and optionally HVPS-specific) CNN(s) for 2D-S and HVPS images can be used with some confidence, especially with regard to the identification of ice particles and graupel in deep convective cloud”*
- *We have excluded classes of a sample size less than 50 in the final CNN results tables (figs 3 and 6), due to them being too small to provide reliable statistics.*

Major Issue 2: The paper identifies a fundamental limitation of the supervised classification framework but could be strengthened by drawing out a key community-level implication more explicitly.

Section 4.2 contains one of the most important findings of the paper: that the reason for CNN-to-human disagreement is the inherent ambiguity of OAP images coupled with unprecise class definitions (possible mismatches between those definitions and the training set images), and different views on hydrometeor classes among research teams. The demonstration is clear: human labellers themselves disagree on the same images, for the same reasons the CNNs make errors, and the CNN performance is therefore subject to lack of human consensus rather than by model capacity. This is an important and general result that extends well beyond DCMEX or the Jaffeux et al. models specifically.

Having identified this root cause, the paper draws only modest conclusions (clearer class boundary definitions, better diffraction training data) and does not follow through to the logical community-level implications. The conclusion section reads as a summary of results and lacks actionable scientific recommendations. Addressing this point would likely increase the manuscript's impact and long-term value for the community.

The authors could add a dedicated paragraph in the Discussion (Section 4.2) or Conclusions (Section 5) that explicitly frames the training set as a community consensus problem. Concretely, the international cloud probe workshop, that is held every four years and which gathers the leading experts in OAP measurement, represents an existing and appropriate forum for establishing a community-validated reference training dataset. The methodology demonstrated in the present study, where multiple expert labellers independently assign classes and a consensus is constructed from their disagreements, is precisely the process that could be formalised and scaled up at such a workshop. A community-validated reference training set, periodically reviewed and expanded, would:

1. Resolve the ambiguity in class boundary definitions by replacing individual judgement with collective expert consensus.
2. Provide a stable, citable benchmark against which future classification models could be evaluated consistently across research groups.
3. Transform the Jaffeux et al. framework into a real community standard, which seems to be one of the goals of the present study.

The paper's title and stated purpose are the evaluation of CNN tools for wider usage. If the fundamental obstacle to that uptake is the quality and consistency of the training set (as the paper demonstrates) then a recommendation to address that barrier at the community level should be a central conclusion.

Relevant sections: Section 4.2 (Discussion, CNN evaluation), Section 5 (Conclusions).

Thank you for this suggestion, this has now emphasised in a paragraph at the end of the CNN evaluation section:

L302-308: "For the future development of CNNs for image classification, input from the wider cloud physics community would be beneficial. Jaffeux et al. (2025) provides the foundational work for CNN architecture, and a set of viable classes for OAP images. To improve this, using a collective expert consensus to refine the definition of classes would clear up disagreements around the degree of riming, and the thresholds that separate habits e.g. aggregate classes. Likewise having more examples of unusual classes, specifically 2D-S diffracted images, will result in better classification and more accurate size distributions. Defining these classes and building a benchmark training dataset could take place during an in situ instrument workshop, such as Baumgardner et al. (2012). To create a community standard set of models that any researcher can implement and build on."

And a sentence in the conclusion:

L413-415: "For the further development of these CNNs, clearer definition of the boundaries between classes would be advantageous. This could be achieved through the expert consensus of class definitions, and the creation of a community standard set of models and training datasets."

D) Minor Issues

Minor Issue 1: Per-class accuracy is misleading on imbalanced data and should not be reported as a primary performance indicator

Per-class accuracy systematically overstates model performance when classes are rare, because of the large number of true negatives. The extreme case is illustrated in Figure 5a: the model does not identify the single CBC image in the HVPS test set (recall = 0%, precision = 0%), yet the reported per-class accuracy is 98.18%. The random-labelling control experiment (Figure A2) further illustrates this: random labels achieve a weighted accuracy of 80%, only marginally below the human-labelled result, despite near-zero F1 scores. The authors already make this observation (l. 224 and after), but do not draw the logical conclusion, which is that the per-class accuracy metric is irrelevant.

I suggest removing per-class accuracy from Figures 3, 5, and A2 and restricting to precision, recall, and F1 score, and retain accuracy only at the level of the full image sample. It is possible to add a clear methodological caveat, for example in Section

2.2.4, explaining why per-class accuracy is not a pertinent measure of model quality in this imbalanced setting.

Relevant sections: Section 2.2.4, Figures 3, 5, A2, Sections 3.2.1 and 3.2.2.

We agree with this, and have removed the accuracy metric from all figures (figs 3, 6, A2).

Description of why it is not used has been added to the methods section:

L156-159: "Accuracy is another metric (eq. C1), describing the fraction of predictions the model got correct. However, it is misleading, including the true negatives i.e. correctly identifying images outside the class, which dominates the metric for classes with a small sample size. The DCMEEX test dataset has uneven class sizes; consequently it is not used in this evaluation."

Minor Issue 2: Text

Several sentences can be improved, here are a few of them:

- L. 17: "Cumulonimbus, the largest convectively formed cloud with serious impacts on climate and weather." not a complete sentence.
L.17 Changed to: "Cumulonimbus is the largest convectively formed cloud, with serious impacts on climate and weather."
- L. 25: "Together, this results in a uncertain representation..." "a" should be "an".
L.25: Changed to "this results in an uncertain representation"
- L.137: "All 2D-S and HVPS images were created to the appropriate 200 by 200 pixel size...", "resized" instead of "created"
L135: Changed to: "All 2D-S and HVPS images were resized to the appropriate 200 by 200 pixel size for CNN use".
- L.168: "the mean across all classes provide an overall score...", "provides"
L.152: Changed to "provides"
- L.172-173: "The F1 score is a harmonic mean...", "the harmonic mean"
L.155: Changed to "is the harmonic"
- L.178-179: "By definition, CP and RA are the only classes that have had visible riming taking place...", "that exhibit visible riming"?
L.316-317: Changed to 'By definition, CP and RA are the only classes that exhibit visible riming'
- L.212-213: "Other misidentified CP classes", the formulation is unclear
This line has been removed
- L.244: "an precision score"
This line has been removed
- L.268: "A equivalent correction factor"
This line has been removed
- L.316-317: "some pair-wise label combinations for humans (fig. 1) was significantly higher..." "were"
This line has been removed
- L.351: "recall if a weak point", "is" instead of "if"?
L213: changed to "recall is a weak point"

McFarquhar, G. M., Baumgardner, D., Bansemer, A., Abel, S. J., Crosier, J., French, J., ... & Dong, J. (2017). Processing of ice cloud in situ data collected by bulk water, scattering, and imaging probes: Fundamentals, uncertainties, and efforts toward consistency. *Meteorological Monographs*, 58, 11-1.

Review 2

Review of “Evaluation and application of a convolutional neural network for graupel identification in DCMEX deep convective cloud” by Ezri E. Alkilani-Brown, Declan L. Finney, Alan M. Blyth, Paul R. Field, Jonathan Crosier, and Chetan R. Deva

Summary: This work employs convolutional neural networks (CNNs) for ice crystal habit classification, described by Jaffuex et al. (2025), to label in-situ ice crystal images collected from the Two-Dimensional Stereo probe (2D-S) and the High-Resolution Precipitation Spectrometer (HVPS) during the 2022 Deep Convective Microphysics EXperiment (DCMEX). The performances of three CNNs are evaluated: a universal neural network applicable to many optical array probes, and probe-specific networks for the 2D-S and HVPS. Sets of randomly-selected, manually-classified images are used for test sets. About 10 percent of the test images for each probe are manually classified by more than one person, and used to evaluate the uncertainty of manual classification. Performance is quantified using accuracy, precision, recall, and F1 score metrics, with emphasis given to the successful identification of graupel (rimed) particles. Finally, the universal CNN is used for 2D-S images, and the specific CNN for HVPS images, to classify (nearly) all crystals imaged during the DCMEX campaign and quantify the proportions of crystal habits observed in convective clouds.

This study works to fill a gap in the literature, namely in that although numerous algorithms for ice crystal classification have been developed and tested on sets of idealized crystal images, few have been broadly applied to the scale of an entire field campaign. The substantial uncertainty in habit that is seen even between manual classifications highlights the issue that a large portion of crystals collected during field campaigns are ambiguous in habit (and also that the class definitions can be ambiguous), at least when viewed at the resolution of an optical array probe. Likewise, the selection of random crystals for testing, rather than representative crystals, results in a notable decline in performance for the neural networks compared to the studies in which they were developed. Nevertheless, in the context of DCMEX, the neural networks are shown to be useful for distinguishing graupel particles from other particles, a distinction that is important for microphysical understanding.

Recommendation: The subject matter discussed in this manuscript represents a valuable contribution to the literature, in outlining the utility, but also the limitations, of applying CNNs for habit classification across the scale of an entire field campaign. However, it is difficult to determine the main takeaway from certain portions of the manuscript, such that I think reorganization, or more concise writing, is necessary. Additionally, although limitations are discussed, the skewing of the random test dataset toward compact particles is not ideal for evaluation. Given that I think significant changes are required regarding the methodology and organization of the paper, I recommend major revisions.

General comments:

Random sampling of the test set. Random sampling provides a valuable test set for how a CNN will perform for images from this particular field campaign, but likely does not produce results that are broadly applicable to cases with substantially different habit distributions. Ideally, I'd suggest randomly sampling and classifying images until ideally 100 images have been acquired for each class, matching Jaffeux et al. 2025. To more quickly obtain examples of less common habits, I'd suggest randomly sampling from specific time periods when these habits are more frequent, if any such periods are known to exist. Then, to balance the test set, take a subset of 100 images for the overpopulated classes. This could be done with the human-to-human comparison as well, if an additional person could be found to classify many crystals (although this may be prohibitively difficult to do, based on personal experience classifying crystals and the time commitment involved).

As a caveat, the low sample sizes in the current test set for CC in the 2DS, and CBC/HBC for the HVPS suggest to me that there aren't 100 samples even in the entire dataset, so it would make sense to omit those classes, except for the purpose of incorrect classifications by the CNN. While this balanced evaluation could be covered alongside the current unbalanced one, I think it would be reasonable to present only the balanced evaluation in figures for conciseness. Instead, the human-identified fractional habit distribution could be presented as part of Section 3.3 (DCMEX application), and compared to the CNN results.

The nature of DCMEX data means we have disproportionately sampled a select number of habits – CP / FA / Co and a couple of others.

For example, currently 31 2D-S CA particles have been identified out of 10,000 images. If we wanted 100 CA images and assume the whole dataset is proportional in the types of particles observed, we would have to label another 22,000 images. We consider this impractical at this stage of the analysis and have decided the following:

- We have narrowed the scope of the paper, testing only the dominant habit categories from DCMEX dataset, and made this result clear throughout. (manuscript examples are found on page 3-4 of this document)*
- We have excluded classes of a sample size less than 50 in the final CNN results tables (figs 3 and 6), due to them being too small to provide reliable statistics.*

Organization. This work appears to have two competing objectives, being the evaluation of CNNs, and the use of CNNs to identify habit in DCMEX. I think that these two objectives should be more cleanly separated for clarity, with each objective having its own dedicated section and associated discussion. By keeping everything together, the manuscript can also be written more concisely. To accomplish this, I'd suggest the following reorganization, although this is far from the only way to do so:

Merge the context of DCMEX and the 2DS/HVPS probes in section 2.1 into the portion of the introduction discussing the same topics.

Title Section 2 something along the lines of "CNN Evaluation", and include what is currently section 2.2, and sections 3.1 and 3.2. Merge the discussion in what is currently section 4.1 into what is currently section 3.1, and the discussion in what is currently section 4.2 into what is currently section 3.2. Relabel subsections as needed.

Title Section 3 something along the lines of “Graupel Identification in DCMEX” and include what is currently section 2.2.5, everything currently in section 3.3. Merge the discussion in what is currently section 4.3 into the new section 4. Relabel subsections as needed.

The conclusion then becomes Section 4.

I’d prioritize manuscript organization over the above major change, as I think the major findings of this work are an important contribution to the literature even as is, and simply need to be communicated more clearly.

We agree that changes to the structure of paper will make the findings clearer. Changes are as follows:

<i>Original structure</i>	<i>New structure</i>
1. Intro 2. Methods 2.1. DCMEX campaign 2.1.1. imaging instruments 2.2. CNN application 2.2.1. Jaffaux description 2.2.2. Data cleaning 2.2.3. Testing procedures 2.2.4. Evaluation metrics 2.2.5. Graupel approximation 3. Results 3.1. Human labelling 3.2. CNN evaluation 3.2.1. 2D-S 3.2.2. HVPS 3.3. DCMEX application 3.3.1. Graupel adjustment 3.3.2. Campaign overview 4. Discussion 4.1. Human confusion 4.2. CNN evaluation 4.3. DCMEX ice composition 5. Conclusion Appendix A. Random labelling Appendix B. Sample volume Appendix C. CNN stats calculation	1. Intro 2. Data 2.1. DCMEX campaign 2.2. Instrument description 3. CNN evaluation 3.1. <u>Methods</u> 3.1.1. Jaffaux description 3.1.2. Testing procedures 3.1.3. Evaluation metrics 3.2. <u>Results and discussion</u> 3.2.1. Human labelling 3.2.2. 2D-S 3.2.3. HVPS 4. Graupel identification in DCMEX 4.1. Graupel approximation 4.2. Campaign overview 4. Conclusions Appendix A. Random labelling Appendix B. Sample volume Appendix C. CNN stats calculation

In the new version of this manuscript, a lot of sections have been merged:

- 2.2. includes both the instrument description and data cleaning procedures (old 2.1.1 and 2.2.2)
- 3.2.1 includes both human labelling results and discussion (old 3.1 and 4.1)

- *3.2.2 includes both 2D-S results and 2D-S relevant discussion from CNN evaluation (old: 3.2.1 and 4.2)*
- *3.3.2. includes both HVPS results and HVPS relevant discussion from CNN evaluation (old: 3.2.2. and 4.2)*
- *4.1. includes both graupel adjustment and graupel approximation (old 2.2.5 and 3.3.1)*
- *4.2. includes both campaign overview and DCMEX ice composition (old 3.3.2 and 4.3)*

Grammar. The grammatical structure could be improved. While I will not specifically point out everything, I will provide a couple examples. For one, at lines 219-221, the second sentence is a sentence fragment, and should be combined into the first sentence: “However, considering the weighted test scores across all and sub-setted habits, the global CNN performed almost universally better than the 2D-S-specific CNN. Apart from a very slightly higher weighted precision score.” The paragraph in which the above example is found then concludes at line 229 with another sentence fragment, “Respectively a higher weighted precision, recall and F1 scores of 88.60%, 88.30% and 88.28%,” and abruptly ends the section with a statistic without any comment on the importance of the statistic. Also note that the latter issue will be resolved by merging section 3.2 with the discussion in section 4.2, as suggested.

Thank you for highlighting this, the paper has been re-read to remove sentence fragments. Ones specifically outlined are below:

- *L219-221 changed from “However, considering the weighted test scores across all and sub-setted habits, the global CNN performed almost universally better than the 2D-S-specific CNN. Apart from a very slightly higher weighted precision score.”*
To L245-246: “Overall, considering the weighted test scores, the global CNN performed almost universally better than the 2D-S-specific CNN.”
- *L229 “Respectively a higher weighted precision, recall and F1 scores of 88.60%, 88.30% and 88.28%,”*
To L251-252: “Likewise, comparing to the original evaluation by Jaffeux et al. (2025), using the global model and all OAP images, they obtained a higher weighted precision, recall and F1 scores of 88.6%, 88.3% and 88.3% respectively.”

Time series. This is undoubtedly a personal preference, but it would be interesting to see time series plots of habit fractions for a couple of flights from the field campaign. I find that time series plots provide a useful visualization of the microphysical variability within clouds that cannot be expressed using cloud-mean quantities.

I agree it would be nice to see, but prompts further questions that are beyond the scope of the paper. And further analysis relating habits to location / altitude / temperature / LWC etc, would be essential to put the figure into context.

Quantifying model skill. In a scenario where the total numbers of crystals in each category are the same as in the test set, the F1 scores achieved by the random test in Appendix A provide a quantitative benchmark for evaluating the CNNs. If the CNN cannot exceed the F1 score from random classification, this indicates that the CNN has no skill for that class.

Specific comments:

Line 17: Change to “cumulonimbus is the largest convectively formed cloud, with serious impacts on climate and weather”.

L. 17 Changed to: “Cumulonimbus is the largest convectively formed cloud, with serious impacts on climate and weather.”

Line 24: Should be “complexities, e.g., secondary ice”. Note that there are other places in the manuscript where commas are missing before and after “e.g.” or “i.e.” but I will not list them here for brevity.

Commas have not been added, as in British English the comma before/ after e.g. or i.e. is not essential.

Line 26: Commenting that “Importantly, the development lifecycle of graupel within these clouds has not been well characterized” gives me the expectation that this issue will be discussed in the manuscript. However, such a characterization is noted to be beyond the scope of this work. While I agree that an in-depth analysis of graupel lifecycle is too much to include here, I think it is important to explain how using a CNN to identify graupel particles will be useful for understanding graupel lifecycle.

Also, the following sentence beginning with “Including the...” is a sentence fragment and should be combined into the quoted sentence.

L26: Changed from “Importantly, the development lifecycle of graupel within these clouds has not been well characterised. Including the understanding of liquid water removal from riming and the role of graupel in lightning production”

This has been changed to reduce the emphasis and increase clarity, to:

L26-27: “For example, the development lifecycle of graupel within these clouds has not been well characterised. This includes the understanding of liquid water removal from riming and the role of graupel in lightning production”

Paragraph lines 42-55: I’m not sure if the discussion of hail is needed, given that hail is not mentioned later in the manuscript. Further, near the end of the paragraph, it is noted that “a quantitative threshold for when graupel has been created has not been outlined. Instead, relying on a qualitative description...” which makes me expect that such a threshold will be discussed in this work. However, this study also uses a qualitative description, which leads to uncertainty in graupel classification among different human classifiers, as noted later in this work.

For the above reasons, I think this paragraph should be removed, and the previous paragraph should conclude with something along the lines of “Habits are often organized only according to the variance in depositional growth patterns with temperature and supersaturation. However, crystals that undergo aggregation and/or accretion will exhibit habits that cannot be predicted from environmental conditions alone, including limitless aggregating crystal monomers, graupel, and hail. As a result, knowledge of the ambient atmospheric conditions is insufficient to determine crystal

habits. Instead, the only surefire way to determine habits is to directly view the crystals through in-situ observations.”

Most of the paragraph describing hail and graupel has been removed, some parts have been incorporated into previous paragraph instead for conciseness. Still highlighting how there is no defined ‘riming threshold’ to produce graupel – as this come up in later human / CNN confusion.

Changed from L39-41: “Organising habits often focus on the variation with temperature and supersaturation. Beyond that, the combination of aggregating crystal monomers is infinite. Likewise, accounting for accretion makes the process of particle identification extremely complex.”

To L39-44: “Organising habits often focus on the variation with temperature and supersaturation, aggregation and riming bring further complexities, creating particles that cannot be predicted from environmental conditions alone. Classifying particles which have undergone aggregation and riming can be nuanced. For example, there is no quantitative riming threshold to produce a graupel particle. Instead, defining a riming degree has been used previously e.g. Reinking (1975); Mosimann et al. (1994); Colle et al. (2014); Takami et al. (2022), relying on an arbitrary scale or percentage coverage. Graupel can be informally classified when riming has deformed the ice particle beyond its original shape.”

Line 69: It may be worth noting that classification of habit from geometric features has also been done without principal component analysis, as described by Holroyd (1987), and applied generally to field campaign data by Schima et al. (2024). These references are provided below.

Changed from L69-71: “Alternatively, extracting the geometric features of particles e.g. perimeter, area and aspect ratio, then using principal component analysis to reduce dimensions before classification has been completed for Cloud Particle Imager (Lindqvist et al., 2012; Praz et al., 2018) and OAP images (Praz et al., 2018).”

To L58-60: “Alternatively, extracting the geometric features of particles e.g. perimeter, area and aspect ratio to make classifications (Holroyd, 1987; Schima et al., 2024). Or also applying principal component analysis to reduce geometric feature dimensions before classifications (Lindqvist et al., 2012; Praz et al., 2018).”

Holroyd, E. W. (1987). Some techniques and uses of 2D-C habit classification software for snow particles. *Journal of Atmospheric and Oceanic Technology*, 4(3), 498–511. [https://doi.org/10.1175/1520-0426\(1987\)004<0498:stauoc>2.0.co;2](https://doi.org/10.1175/1520-0426(1987)004<0498:stauoc>2.0.co;2)

Schima, J., McFarquhar, G. M., Delene, D., Heymsfield, A., Bansemer, A., Schnaiter, M., et al. (2024). A multi-probe automated classification of ice crystal habits during the IMPACTS campaign. *Journal of Geophysical Research: Atmospheres*, 129, e2024JD040895. <https://doi.org/10.1029/2024JD040895>.

Paragraphs lines 76-90: This is the part of the introduction that really highlights the motivation of this study, which is to evaluate the performance of the Jaffeux et al. 2025 algorithms using untested field campaign data, and then use the classifications to describe ice crystal habits in clouds. I think the importance of this work should be more clearly emphasized. As is noted, virtually all CNNs have been developed on hyper-

specific, idealized training images, but the work presented here is a critical test of the ability of the algorithms to adapt to the ambiguity of crystal habits in real clouds.

Thank you. The end of this paragraph has been rephrased for more emphasis.

From L83-87 “Generalised and open source algorithms, for wider use and development by the ice microphysics community, offer a means to reduce the burden arising from CNN development and lower the barrier for using these tools to answer science questions. Jaffeux et al. (2022, 2025) provide one of the first examples of an open-source, generalised set of CNNs for ice hydrometeor classification in OAP images. To ensure these models are beneficial to the research community, independent evaluation of these models is essential.”

To L72-76: “To counter this, having a set of generalised and open source algorithms for wider use and development by the ice microphysics community would be beneficial. This would reduce the burden arising from CNN creation and lower the barrier for using these tools to answer science questions. Jaffeux et al. (2022, 2025) provide one of the first examples of an open-source, generalised set of CNNs for ice hydrometeor classification in OAP images. An important step in model development is independent evaluation using field campaign data the model has not been trained on.”

Lines 111-112: I’ve typically seen these referred to as the horizontal and vertical channels, and I find that labeling them that way helps for my own visualization.

We would have done this, but unfortunately we do not know the orientation of the channels.

Lines 134-135: It’s probably worth noting that these diameter thresholds were required due to probe resolution constraints. The limitations of these relatively large thresholds should be discussed, as this leaves gaps in the classifications for particles smaller than 300 microns and 1.2-3 microns, which probably constitute a large portion of total crystal mass.

Added L132-133: “The size requirement is due to diffraction of particles, below the threshold size there is uncertainty in the ‘real’ particle shape.”

Added into discussion about the missed mass: L381-382: “With the peak concentration below 300 μ m, we can infer that a large portion of the total particle mass falls outside the CNN-classifiable range.”

Lines 145: I assume that columns with an aspect ratio greater than 10 were rare enough such that their omission makes negligible difference, but I’d like to see a comment on this, given that the habit statistics could be skewed if they are more frequent.

Yes this was the case, but didn’t impact the CNN classifiable columns.

L144-145 changed from: “This was due to numerous artefacts fitting this criteria, and predominantly excluded columns.”

To L108-109: “This was due to numerous artefacts fitting this criteria, but included approximately 2% of columns that were smaller than 300 μ m.”

Lines 149-150: The use of multi-label consensus for some labels, but the label from a single human for others, is a methodological inconsistency. For consistency, I’d recommend sticking to labels from a single person when evaluating the CNNs, given

that a large majority of the images were classified by one person. The comparison of human labelers in Section 3.1 effectively quantifies the uncertainty in using classifications performed by one person and thus remains a valuable analysis. In doing this, the need to describe the concept of a group consensus (lines 156-165) would be eliminated, allowing the manuscript to be shortened.

This has been changed, now all CNN evaluation has been done with 1 human labeller. Section 3.1.2 (testing procedure) outlines the new method with 1 labeller. All text referencing statistics has been corrected, and figures have also changed to reflect this.

- *Fig. 2 Confusion matrices for 9933 2D-S images, is now fig. 2 Confusion matrices for 9988 2D-S images.*
- *Fig. 3 2D-S CNN evaluation results, is now fig. 3 2D-S CNN evaluation results (for one human labeller)*
- *Old Figure 4. Confusion matrices for 1430 HVPS images, is now Figure 5. Confusion matrices for 1445 HVPS images*
- *Old Figure 5. HVPS evaluation results, is now Figure 6. HVPS evaluation results (for one human labeller)*
- *Fig A1 and A2 have been updated to match the human evaluation.*

This has also altered the HPC correction factor slightly, from 0.88 to 0.87 (on L331)- which has changed some of the values in table 1. But overall, not significant different.

Figure 1: I'm not entirely clear how the pair-wise comparison is performed. How are pairs chosen? My interpretation is that the results could vary depending on which pairs of humans are selected. Please elaborate on the methodology.

The text has been updated for clarity.

From L189-191: "Each image has four labels, which comprise a six pair-wise human label inter-comparison. The matrix depicts each combination of pair-wise labels, with the diagonal being perfect agreement, off-diagonal are disagreeing labels. The order of human labels is irrelevant i.e. Co - HPC is the same as HPC - Co, so a half matrix is displayed."

To L167-170: "Each image has four labels, with a possible six combinations of two label comparison; all six pairs of label intercomparisons have been used. The matrix visualises each combination of pair-wise labels, with the diagonal being perfect agreement, off-diagonal are disagreeing labels. As there are no 'ground truth' labels, the order of labels is irrelevant i.e. Co - HPC is the same as HPC - Co, so a half matrix is depicted."

Figure 2 and associated discussion: Based on the random consensus results, the accuracy metric is meaningless for evaluation purposes. I'd remove accuracy from all figures, remove accuracy statistics from the text, and give a sentence or two of discussion about why this was done.

Accuracy has been removed from all main figures (figs 3, 6 and A2). Description of why it is not used has been added to the methods section:

L156-159: "Accuracy is another metric (eq. C1), describing the fraction of predictions the model got correct. However, it is misleading, including the true negatives i.e. correctly identifying images outside the class, which dominates the metric for classes with a small sample size. The DCMEX test dataset has uneven class sizes; consequently it is not used in this evaluation."

Line 220: The better performance of the global CNN for the 2D-S is a surprising and important result. I'd add in the manuscript that this indicates that the specific model was likely overfit to the training data.

The global model outperforming the 2D-S-specific CNN is to be expected, as this was also reported in Jaffeux et al. 2025. The text below describes the differences between the global and specific models:

L246-249: "The specific model has a slightly higher precision score, which could be attributed to the specialist nature of the model, where weights and biases are constructed exclusively on 2D-S images, thus have a more rigid definition of classes, and the identification of particles within these classes. However, training the global model on the variety of OAP images improves the identification of 2D-S images, with slightly relaxed class boundaries."

Line 233: Include the note that the global CNN doesn't have an RA class here in the text as justification for merging the classes, rather than, or in addition to, within the description of Figure 4.

This is mentioned in the Jaffeux description section - L135-136: "Also, the RA class is not present in the global model, instead images are categorised into CP."

Also mentioned in the beginning of the HVPS section L264: "In the global model evaluation, RA was incorporated into CP."

It is also mentioned in both fig. 5 and 6 captions "Global CNN does not have a RA class thus RA labelled images are placed in CP".

As a consideration, it may also make sense to merge the RA and CP classes within the specific HVPS algorithm here, to provide a fair comparison between the general and specific algorithms. As further justification, the classes are already combined for the campaign-wide analysis.

Combining RA into CP (in fig. 1 below) slightly improves the weighted precision, recall and F1 scores compared to fig. 6 in the manuscript. It does not alter the conclusion that the HVPS-specific model is superior in HVPS image classification over the global model. The purpose of this evaluation is to test habit classification directly from model output, so it is beneficial to keep RA separate. Likewise, others who use the model may want distinct CP and RA categories, and would benefit from understanding how well the model can distinguish CP and RA.

a) global					b) HVPS				
	image count	precision	recall	F1		image count	precision	recall	F1
Co	66	14.76%	87.88%	25.27%	Co	66	21.81%	62.12%	32.28%
CP	752	96.09%	58.78%	72.94%	CP	752	95.05%	81.78%	87.92%
FA	620	92.86%	54.52%	68.70%	FA	620	81.62%	66.61%	73.36%
	Macro mean	67.90%	67.06%	55.64%		Macro mean	66.16%	70.17%	64.52%
	Weighted mean	90.96%	58.28%	68.92%		Weighted mean	85.90%	74.34%	79.09%

Figure 1: HVPS evaluation with class sample sizes greater than 60. RA has also been combined into CP for HVPS-specific model evaluation.

Line 251: For more justification to choose the specific model for the HVPS in this study, it appears to have markedly better performance for the CP class, which comprises the majority of crystals from the campaign.

Thank you for the suggestion, this has been added to the campaign overview section.

L347-349: “The HVPS-specific CNN has been chosen to classify HVPS images, due to slightly better performance than the global model, and higher F1 score for CP classification, which is important for graupel identification.”

Lines 263-268: I find the description of the correction factor hard to follow. My interpretation is that the correction factor is equivalent to the fraction of human-labeled CP images classified as HPC in testing, which from Figure 2 I calculated is 18.7%, divided by the total fraction of crystals classified as HPC over one second from the field campaign data. If my interpretation is not correct, it may be good to reword the description.

Thank you for highlighting this, this adjustment has been difficult to convey in writing alone. In fig 2a). boxes have been added to the CNN identified HPC column and human labelled CP / CNN identified HPC box, to highlight the values being discussed. Figure caption has likewise been updated:

“Confusion matrices for 9988 2D-S images, CNN versus one human labeller. a) global CNN and b) 2D-S-specific CNN evaluation. Human labels are on the y axis, and CNN label on the x axis. Dotted box in a) highlights all images identified as HPC by the model, within this, the solid box outlines human labelled CP but CNN identified HPC.”

Eq1 has also been updated from “Human(CP) $\subset$ CNN(HPC)” to “Human(CP) $\cap$ CNN(HPC)”. The subset notation was incorrectly used, and now is correctly the intersect, describing only images which are Human CP and CNN HPC.

The full equation is now as follows:

$$\frac{\text{Human(CP)} \cap \text{CNN(HPC)}}{\text{CNN(HPC)}}$$

*In the main description of the correction factor, it has been changed from:
L265-269: “The simple adjustment factor used human labelled 2D-S images as a baseline; using the proportion of human identified CP images, but CNN labelled HPC out of the total CNN labelled HPC images (Eq. 1) i.e. the proportion of misidentified CP images. For the global model, due to the relatively small sample of human labelled HPC, a fixed correction factor of approximately 0.88 has been inferred across all size bins. A equivalent correction factor from Jaffeux et al. (2025), using the global model ‘all-OAP’ results would roughly be 0.15.”*

To

L326-334: “The simple adjustment factor used human labelled 2D-S images as a baseline, it is outlined in eq. 1 and depicted in fig. 2a). The factor is created from the total number of human identified CP and CNN identified HPC images (fig.2a) solid box), divided by the total CNN identified HPC images (fig.2a) dotted box). This factor is dataset specific. The sample of images used to test the CNNs is representative of the DCMEX campaign, i.e. there are large discrepancies in the proportions of classes, specifically very high CP and low HPC. This discrepancy is accounted for in the correction factor. For the global model, a fixed factor of approximately 0.87 has been inferred across all size bins. For Jaffeux et al. (2025), who purposefully had an equal proportion of classes, their equivalent correction factor using the global model ‘all-OAP’ results would be 0.15. This highlights how the factor is highly sensitive to class imbalances. If the global model is to be used without re-training, it is recommended that adjustments should take place on a campaign-by-campaign basis.”

Additionally, the authors’ fixed correction factor and the one from Jaffeux et al. (2025) are drastically different – why is this the case?

Thank you for bringing this up, this has now been highlighted in the text, see above for the description.

Lines 290-295: Explain why we might expect to see these differences in composition between the probes. I’m not surprised that the HVPS, which images larger crystals, has more classified aggregates, as the aggregation process more commonly occurs with larger crystals, and increases crystal size.

Note: I see that this is explained in the paragraph from lines 430-438. This is a place in which merging the sections would add clarity.

*With new structure, these sections are now merged. This is specifically discussed here:
L367-371: “These observations make physical sense, where 2D-S would be expected to observe a higher proportion of young pristine habits. With the 3000µm minimum size requirement, HVPS images capture ice particles which have circulated through the cloud for a longer period of time, or frozen raindrops. Older ice particles have a higher probability of undergoing riming or aggregation, which would remove the largest pristine particles, placing them into CP or aggregate classes instead.”*

Lines 297-298: What does “outside the scope of CNN classification” mean? Is there a specific minimum threshold concentration needed for the fraction of a given particle type to be considered nonzero?

This should have been stated more clearly. This section has been slightly re-worded, L297-298 is now changed to:

L379-382: “Seen in CP[†] and Co, there was no peak in concentration within 2D-S observations (above 300μm), so interpretation of the distribution is unclear. With the peak concentration below 300μm, we can infer that a large portion of the total particle mass falls outside the CNN-classifiable range.”

Figure 6 description: The phrase “Legend percentages display (number) proportions of habit concentration to the DCMEX total ice population...” reads as if these are proportions of the total particle count from the 2D-S and HVPS combined. However, I see that the labeled percentages for the ch0, ch1, and HVPS data are proportions of the ch0, ch1, and HVPS particle counts, respectively. Please clarify.

Thanks for highlighting this, see below for caption changes.

Also, do the errorbars refer to the standard deviation of the data among all 1-s time periods, or something else?

This is now clarified in the caption. Fig 6 has also now become fig 8. The caption has changed from:

“Mean ice crystal size distribution across DCMEX, where concentrations account for any measurement period that class was observed. Ch0 and ch1 refer to orthogonal 2D-S channels. [†]refers to post graupel adjustment values. Standard error bars illustrated in each size bin. Legend percentages display (number) proportions of habit contribution to the DCMEX total ice population (explicitly excluding Dif, SP and WD). CP / RA HVPS classes are included in CP[†]”

To: “Mean ice crystal size distribution across DCMEX, where concentrations account for any measurement period that class was observed. Ch0 and ch1 refer to orthogonal 2D-S channels. [†]refers to post graupel adjustment values. Error bars show standard error of the mean, calculated per size bin from all 1 s periods the class was observed in. Legend percentages display count proportions of habit contribution per instrument (and 2D-S channel), explicitly excluding Dif, SP and WD. CP / RA HVPS classes are included in CP[†]”.

Tables 1 and 2: Some of the sample sizes are very small. Be cautious about making conclusions from any flights with low sample sizes. I generally follow the rule of thumb that a minimum sample size of 30 is needed to draw any statistical conclusions.

We agree with this. Tables 1 and 2 just provide a quick overview of campaign days, with values from the table only discussed in one paragraph (L383-387). Days with less than 30 total images have not been mentioned.

Lines 316-317: It is noted that some pair-wise combinations were significantly higher than random when comparing the two. However, the distribution of habits is even in the random assignment case, whereas CP is heavily favored by both human classifiers in the human comparison, so this comparison is misleading. To create a more fair comparison, the set of labels given by one of the human classifiers could be randomly scrambled to create a random assignment case, ensuring a similar distribution of habits. Ideally, the human comparison would be roughly balanced to begin with, but if this cannot be done, the next best thing is to adjust the random classifications.

Yes some-pair wise combinations i.e. CP-CP are higher, this is due to the underlying random sample of images (there are more CP images than all other classes in DCMEX).

Adjusting the random labels to account for this cannot be done fairly, as would depend on having 1 ‘ground truth’ label, but in principal, all human labels are correct. This section doesn’t include any quantitative assessment, it was more to highlight that human labelling has more skill than random, and agree a higher percentage of the time. New text clarifies this.

Now changed L316-317 from: “Nonetheless, comparing to randomly assigned labels, some pair-wise label combinations for humans (fig. 1) was significantly higher than random (fig. A1), especially for matching diagonal labels. This indicates the humans have better skill than a random assignment of labels.”

To L173-176: “Where each labeller has interpreted Jaffeux et al. class definitions differently. Hence, matching labels only occurred 78 and 66% of the time for 2D-S and HVPS images respectively; if there were perfect agreement all values would be on the diagonal and total 100%. Nonetheless, this is significantly better than random labels (fig. A1), with agreement at 12% and 15% for 2D-S and HVPS images. Which indicates humans have more skill than random assignment.”

Lines 323-325: Did all the human classifiers look over the example images from Jaffeux et al. prior to their classifications? If not, I would expect better agreement if these images were analyzed ahead of time.

Yes, a crib sheet of example Jaffeux images and description for every category was presented to labellers. – that’s why it’s interesting to have human-human comparison, how individuals have interpreted the Jaffeux definitions.

This has now been explicitly added into the methods:

L144-145: “To complete human labelling, all participants were presented with the same crib sheet of example Jaffeux et al. images and description of each class.”

Lines 343-344: I wonder if there is a way to quantify this uncertainty from the human comparison data. Would it be possible to construct a margin of uncertainty for the precision, accuracy, and F1 score, knowing that the classifications may be wrong? One idea would be to perform a series of random resamples of the class labels according to the confusion matrices presented in Figure 1, and find the standard deviations of the scores across all resamples.

Quantifying human uncertainty would be ideal, but requires development of new statistical methods that we consider beyond the scope of this paper. The problem is there is no baseline ‘true’ label (that there is when comparing with the CNN). As technically all human labels are correct. So, we cannot subsample to get equal classes. Likewise, because of an uneven distribution of classes, we cannot assess labelling based on deviation from analytical random assignment. We will continue to think about possible metrics in our future work.

Figure 7: I like this visualization; it effectively showcases the ambiguity in the habit of real crystals. I’d argue that some images are inherently unclassifiable; at the resolution of the probe, and without any grayscale, sometimes two habits that are distinct in reality can produce the exact same image on an OAP. As a comparison, it may be informative to also show a few cases where both the human and CNN classify a crystal as CP.

We agree with this, a sentence on how particles may produce the same image has been added to the discussion (specifically relating to CP confused as HPC):

L226-227: “This emphasises the underlying issue that two distinct habits can appear extremely similar in a black and white OAP image.”

Old fig 7, now figure 4 has now been updated to include example of human CP / CNN CP. With caption correspondingly updated:

“Example confused 2D-S images from the human CP class, with respective Co and HPC labels from the global and 2D-S-specific CNN models.”

To “Example human labelled CP 2D-S images. Panels depict global and 2D-S-specific CNN misidentified Co and HPC images. CP that were correctly identified by both global and 2D-S-specific models are also depicted.”

Lines 385-389: As noted previously, it would make sense to just exclude the class for small sample size instead of taking the time to explain.

This has been completed. We have chosen to assess the classes with a sample size larger than 50 for both 2D-S and HVPS evaluation. Sections of text explaining this have been removed.

This change has clearly been stated in the text:

L207-208: “For classes with a statistically significant sample size, i.e. larger than 50, fig. 3 displays the associated evaluation statistics”.

Line 390: “Despite not attaining perfect performance scores” is redundant given its mention later in the paragraph.

This has been removed.

Lines 423-426: Figure 6 is very separated from this part of the text; this is one reason why I’ve suggested merging this section with Section 3.3. Also, while I agree with the approach of combining the FA and CP classes, I’m not convinced that there is enough evidence to confidently claim that it is “the correct approach.”

With structure of paper altered, fig 6 is now better placed. I assume you mean combining RA and CP, agreed it is a bit strong.

Now re-worded from L423-425: “Alignment of CP[†] and FA distribution curves between 2D-S and HVPS demonstrate consistency across instruments; evidently combining RA into CP class for HVPS was the correct approach in understanding the distribution of rimed particles. This alignment also further consolidates the graupel adjustment applied to CP.”

To L372-375: “Alignment of CP[†] and FA distribution curves between 2D-S and HVPS demonstrate consistency across instruments. This also highlights how combining HVPS RA into CP class may be important for understanding the distribution of rimed particles, and consolidates the graupel adjustment applied to CP.”

Lines 444-445: For classes with extremely low precision, such as CBC, the size distributions will look like those from other classes, as the overwhelming majority of particles placed in that class will be incorrect classifications. For these classes, I would argue we cannot infer anything at all from the size distribution plots. The most robust result is by far the CP size distribution.

We agree with this. Old Fig 6, now fig 8 has had CBC, HPC and CA panels removed.

Discussion of these habits have been removed from section 4.2 campaign overview.

Lines 449-451: Theoretically in-cloud sampling time does relate to the number of particles observed, but this effect is overcome by variations in number concentration and size between cases.

This sentence has been removed.

Lines 461-463: While there were issues with CP categorization, I think that this statement is misleading, as it implies classifications for other classes were relatively accurate. Instead, both CNNs performed best for CP crystals per the F1 metric, which should be emphasized here, while classes such as Co were much more poorly represented.

Agreed, now broadened. The problem is the poor performance of Co is explainable, where CP images could also fit into Co, this is down to the underlying ambiguity of the images, which has potentially skewed results.

L461-463 changed from: “There were issues with the categorisation of CP images, partly explained through the loose definition of CNN classes, human uncertainty when labelling ambiguous images, and also fundamental misidentification as HPC from the CNNs.”

To L398-401: “There were issues with the categorisation of images, partly explained through the loose definition of CNN classes and human uncertainty when labelling ambiguous images. Importantly for graupel identification in 2D-S images, is also fundamental CNN misidentification as HPC, which can be adjusted for”.

Line 478: I don't think the images from the 2D-S and HVPS provide sufficient information to determine a degree of riming with any accuracy. Higher-resolution probes with grayscale imaging, like the PHIPS and CDP, are needed.

Agreed, and further detail has been added to this sentence. But we have not added in info about grayscale imaging, as these models are explicitly classifying black and white images.

From: “In DCMEX, riming was ubiquitous, but there is no defined riming degree to separate classes.”

To L416-417: “In DCMEX, riming was ubiquitous, but there is no defined riming degree to separate classes, or enough distinction within black and white OAP images.”

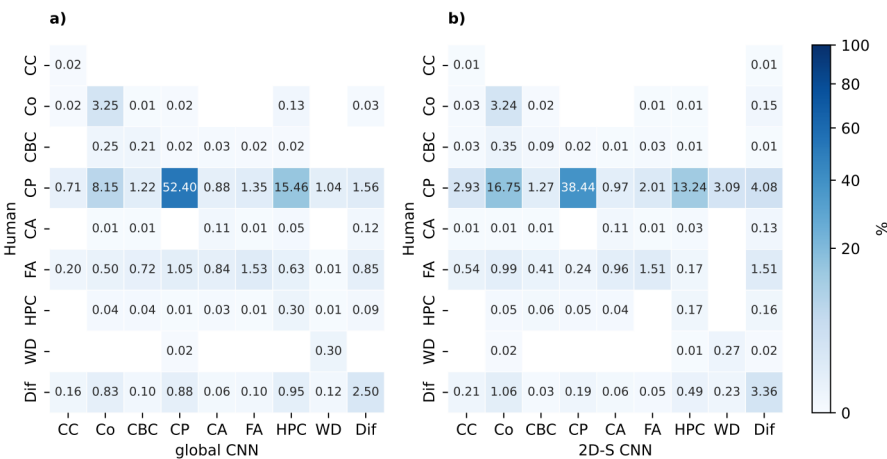
Context on PHIPS probe if curious: Abdelmonem, A., Järvinen, E., Duft, D., Hirst, E., Vogt, S., Leisner, T., and Schnaiter, M.: PHIPS–HALO: the airborne Particle Habit Imaging and Polar Scattering probe – Part 1: Design and operation, Atmos. Meas. Tech., 9, 3131–3144, <https://doi.org/10.5194/amt-9-3131-2016>, 2016.

Appendix C: I do think that the accuracy calculation should be kept here, even with my suggestion of removing accuracy from the rest of the manuscript, since it is a frequently used metric.

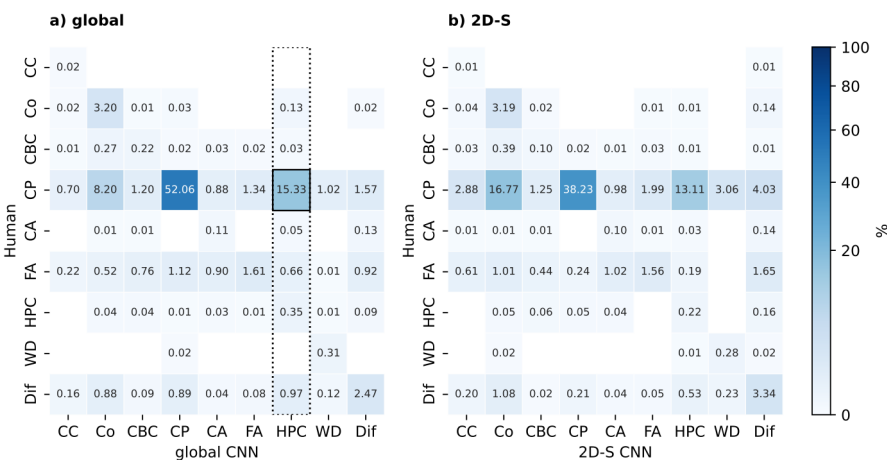
Agreed, it has been mentioned briefly in the text (L156-159), accuracy calculation has been kept in appendix.

Old / new figures comparison

Old fig 2. 9933
2D-S images,
from mix of
single label and
consensus.



New fig 2. 9988
2D-S images
labelled by 1
person.



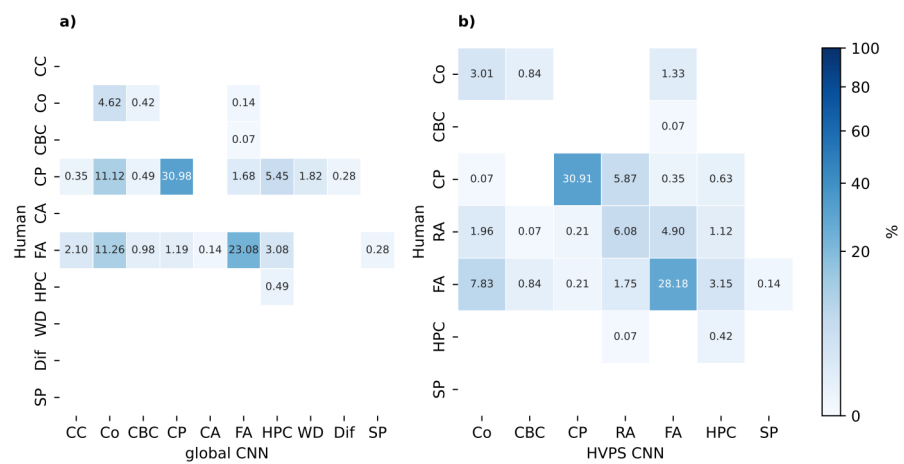
Old fig 3. 2D-S
stats, from mix of
single label and
consensus.

a)	image count	accuracy	precision	recall	F1	b)	image count	accuracy	precision	recall	F1
CC	2	98.90%	1.80%	100.00%	3.54%	CC	2	96.23%	0.27%	50.00%	0.53%
CA*	31	97.97%	5.70%	35.48%	9.82%	CA*	31	97.77%	5.16%	35.48%	9.02%
WD*	32	98.80%	20.41%	93.75%	33.52%	WD*	32	96.63%	7.56%	84.38%	13.88%
CBC*	55	97.55%	9.13%	38.18%	14.74%	CBC*	55	97.73%	4.79%	16.36%	7.41%
HPC*	53	82.52%	1.72%	56.60%	3.34%	HPC*	53	85.67%	1.21%	32.08%	2.33%
Co*	344	90.00%	24.94%	93.90%	39.41%	Co*	344	80.55%	14.43%	93.60%	25.00%
CP*	8222	67.63%	96.34%	63.31%	76.40%	CP*	8222	55.16%	98.71%	46.44%	63.16%
FA*	629	93.71%	50.67%	24.17%	32.72%	FA*	629	93.06%	41.67%	23.85%	30.33%
Dif*	565	94.16%	48.53%	43.89%	46.10%	Dif*	565	91.60%	35.65%	59.12%	44.47%
All macro avg.		91.25%	28.80%	61.03%	28.84%	All macro avg.		88.27%	23.27%	49.03%	21.79%
All weighted avg.		72.01%	86.72%	60.63%	69.54%	All weighted avg.		61.19%	86.94%	47.21%	57.72%
*Subset macro		90.32%	32.18%	56.56%	32.24%	*Subset macro		87.26%	26.15%	50.06%	24.97%
*Subset weighted		72.31%	86.73%	61.22%	69.97%	*Subset weighted		62.23%	86.96%	48.93%	59.17%

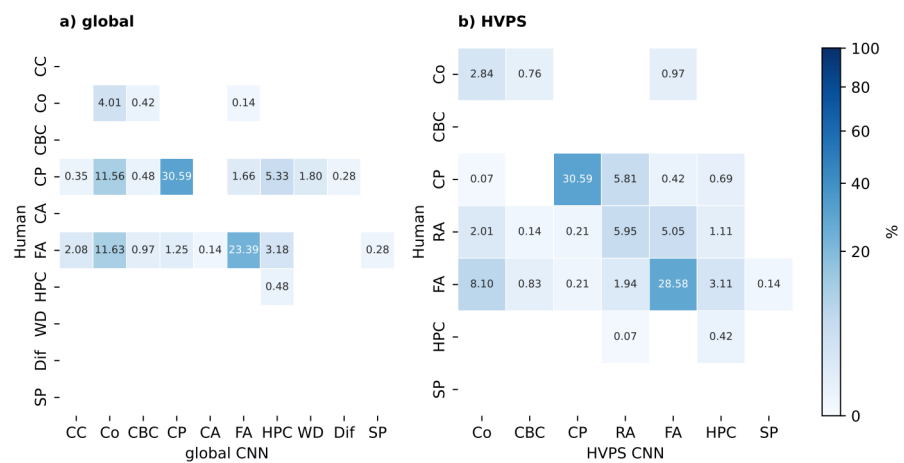
New fig 3. 2D-S
stats, labelled by
1 person.
Sample size
larger than 50.

a) global					b) 2D-S				
	image count	precision	recall	F1		image count	precision	recall	F1
CBC	60	9.44%	36.67%	15.02%	CBC	60	5.26%	16.67%	8.00%
HPC	58	2.00%	60.34%	3.87%	HPC	58	1.56%	37.93%	3.00%
Co	341	24.41%	93.84%	38.74%	Co	341	14.18%	93.55%	24.62%
CP	8221	96.14%	63.25%	76.30%	CP	8221	98.66%	46.44%	63.15%
FA	672	52.61%	23.96%	32.92%	FA	672	42.74%	23.21%	30.09%
Dif	570	47.50%	43.33%	45.32%	Dif	570	35.16%	58.60%	43.95%
Macro mean		38.68%	53.57%	35.36%	Macro mean		32.93%	46.07%	28.80%
Weighted mean		86.85%	60.32%	69.50%	Weighted mean		87.19%	46.96%	57.80%

Old fig 4.
1430 HVPS
images, from mix
of single label
and consensus.



New fig 5. 1445
HVPS images
labelled by 1
person.



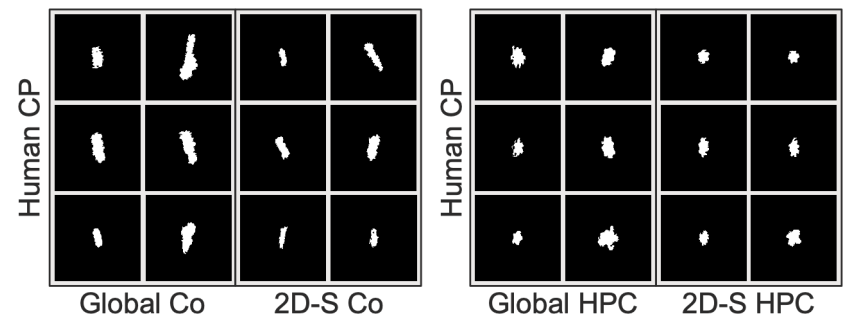
Old fig 5. HVPS CNN stats, from mix of single label and consensus.

a)						b)					
	image count	accuracy	precision	recall	F1		image count	accuracy	precision	recall	F1
CBC	1	97.94%	0.00%	0.00%		CBC	1	98.18%	0.00%	0.00%	
HPC	7	91.02%	5.43%	100.00%	10.29%	HPC	7	95.03%	7.89%	85.71%	14.46%
Co*	74	75.86%	17.10%	89.19%	28.70%	RA*	205	84.03%	44.16%	42.44%	43.28%
CP*	746	79.03%	96.30%	62.31%	75.66%	Co*	74	87.96%	23.37%	58.11%	33.33%
FA*	602	80.65%	92.44%	58.30%	71.51%	CP*	541	92.65%	98.66%	81.70%	89.38%
						FA*	602	79.55%	80.92%	67.17%	73.41%
All macro avg.		84.90%	42.25%	61.96%	46.54%	All macro avg.		89.57%	42.50%	55.85%	50.77%
All weighted avg.		79.62%	90.07%	62.15%	71.16%	All weighted avg.		85.67%	78.97%	68.70%	72.77%
*Subset macro		79.87%	68.70%	77.60%	62.35%	*Subset macro		86.79%	61.88%	68.23%	61.99%
*Subset weighted		82.69%	90.66%	69.68%	76.65%	*Subset weighted		86.85%	79.48%	73.31%	75.48%

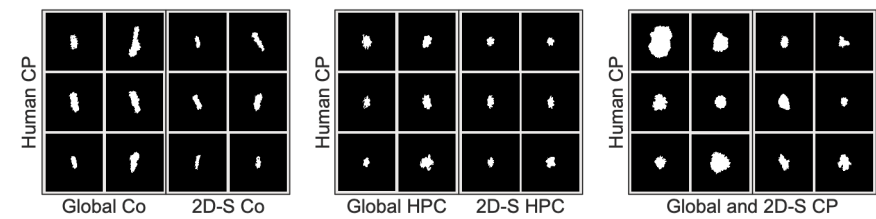
New fig 6. HVPS CNN stats, labelled by 1 person. Sample size larger than 50.

a) global					b) HVPS				
	image count	precision	recall	F1		image count	precision	recall	F1
Co	66	14.76%	87.88%	25.27%	RA	209	43.22%	41.15%	42.16%
CP	752	96.09%	58.78%	72.94%	Co	66	21.81%	62.12%	32.28%
FA	620	92.86%	54.52%	68.70%	CP	543	98.66%	81.40%	89.20%
					FA	620	81.62%	66.61%	73.36%
	Macro mean	67.90%	67.06%	55.64%		Macro mean	61.33%	62.82%	59.25%
	Weighted mean	90.96%	58.28%	68.92%		Weighted mean	79.73%	68.29%	72.92%

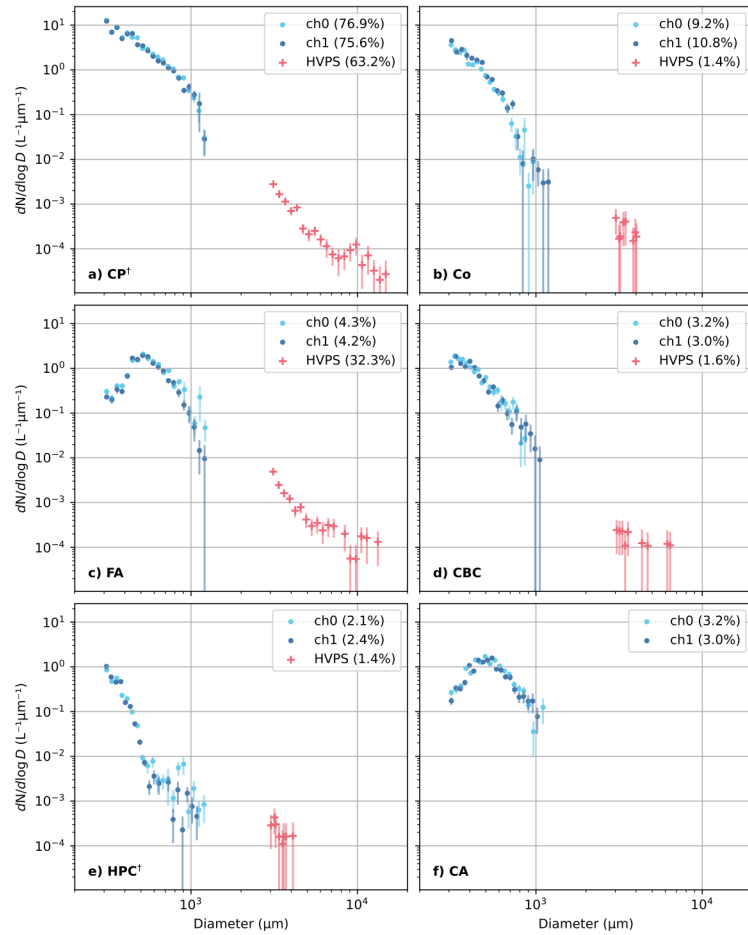
Old fig 7.



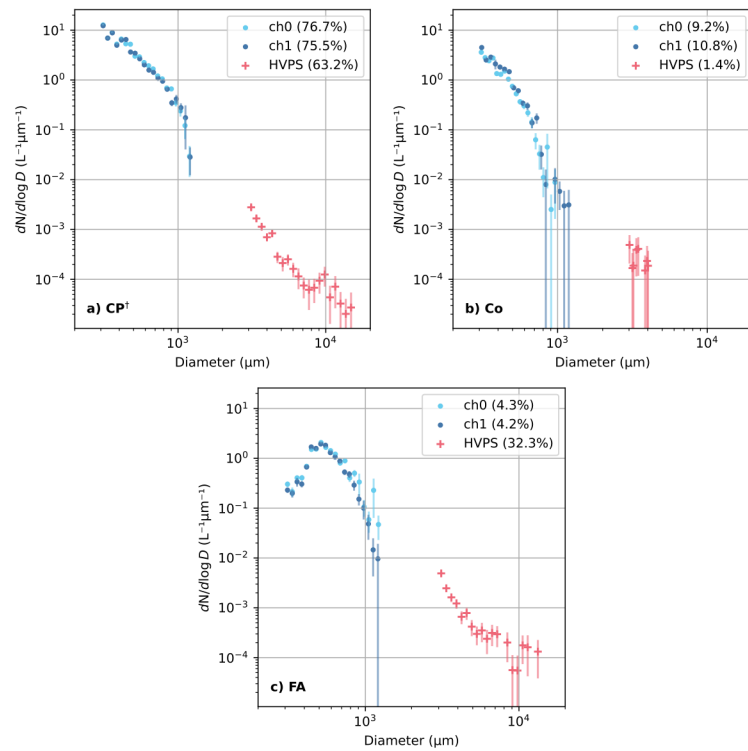
New fig 4. Inclusion of human and CNN agreeing CP images.



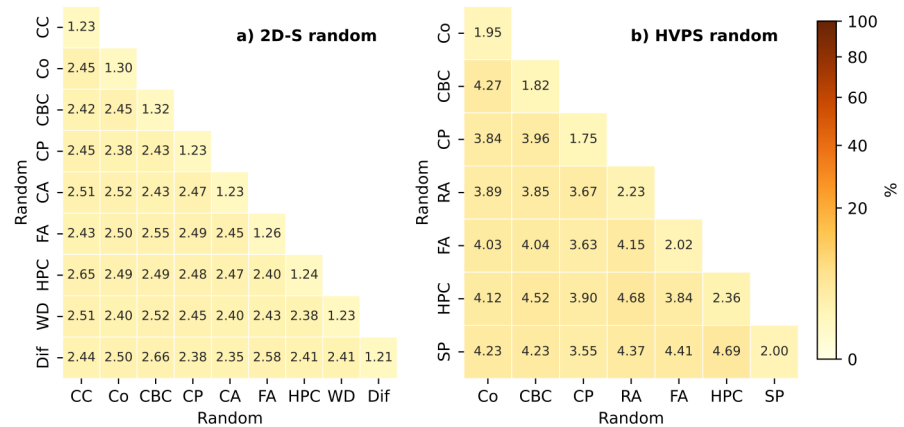
Old fig 6.



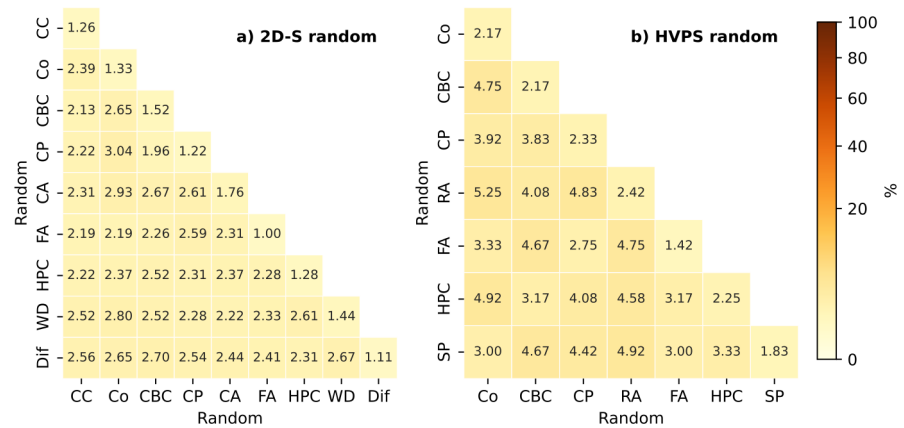
New fig 8. Only the three top habits observed during DCMEX



Old fig A1. pair-wise combinations of four random labels for a) 9988 2D-S 'images' and b) 1445 HVPS 'images'.



New fig A1. pair-wise combinations of four random labels for a) 900 2D-S 'images' and b) 200 HVPS 'images'.



Old fig A2. CNN evaluation results from four label 'random consensus' of a) 5137 2D-S images and b) 880 HVPS

a) 2D-S random global CNN						b) HVPS random HVPS CNN					
	image count	accuracy	precision	recall	F1		image count	accuracy	precision	recall	F1
CC	556	88.28%	12.90%	1.44%	2.59%	Co	126	77.61%	15.53%	12.70%	13.97%
Co	594	78.18%	11.19%	12.79%	11.94%	CBC	118	84.66%	5.26%	0.85%	1.46%
CBC	602	86.82%	17.95%	3.49%	5.84%	CP	111	63.41%	11.91%	29.73%	17.01%
CP	551	46.78%	10.50%	52.63%	17.50%	RA	137	74.66%	15.87%	14.60%	15.21%
CA	578	87.02%	9.17%	1.73%	2.91%	FA	117	60.91%	12.79%	33.33%	18.48%
FA	555	87.09%	14.00%	3.78%	5.96%	HPC	145	79.66%	14.58%	4.83%	7.25%
HPC	580	74.87%	11.48%	18.28%	14.11%	SP	126	85.68%	50.00%	0.79%	1.56%
WD	569	87.93%	11.94%	1.41%	2.52%	Macro mean		75.23%	17.99%	13.83%	10.71%
Dif	552	84.95%	8.61%	4.17%	5.62%	Weighted mean		75.58%	18.17%	13.30%	10.59%
Macro mean		80.21%	11.97%	11.08%	7.66%						
Weighted mean		80.31%	12.01%	10.96%	7.67%						

*New fig A2. CNN
evaluation
results from one
random label of
a) 9988 2D-S
images and b)
1445 HVPS
images*

a) 2D-S random global CNN

	image count	precision	recall	F1
CC	1098	8.85%	0.91%	1.65%
Co	1139	11.14%	12.82%	11.92%
CBC	1159	15.02%	3.02%	5.03%
CP	1078	10.57%	53.06%	17.64%
CA	1190	9.05%	1.51%	2.59%
FA	1059	9.80%	2.83%	4.40%
HPC	1107	11.37%	17.98%	13.93%
WD	1094	11.56%	1.55%	2.74%
Dif	1064	10.38%	5.08%	6.82%
Macro mean		10.86%	10.97%	7.41%
Weighted mean		10.88%	10.82%	7.37%

b) HVPS random HVPS CNN

	image count	precision	recall	F1
Co	201	12.23%	11.44%	11.83%
CBC	201	8.00%	1.00%	1.77%
CP	223	13.62%	27.35%	18.18%
RA	194	10.05%	10.31%	10.18%
FA	221	15.61%	35.75%	21.73%
HPC	204	11.69%	4.41%	6.41%
SP	201	0.00%	0.00%	
Macro mean		10.17%	12.89%	11.68%
Weighted mean		10.30%	13.43%	11.95%