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 occurence and to aggregate behavior, with clear statements that minority-class skill remains unresolved.

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

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.

Section 4.2 (Discussion, CNN evaluation), Section 5 (Conclusions).

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.

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

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. 25: "Together, this results in a uncertain representation…." "a" should be "an".
- L.137: “All 2D-S and HVPS images were created to the appropriate 200 by 200 pixel size...”, “resized” instead of “created”
- L.168: “the mean across all classes provide an overall score...” , “provides”
- L.172-173: “The F1 score is a harmonic mean...”, “the harmonic mean”
- L.178-179: “By definition, CP and RA are the only classes that have had visible riming taking place…”, “that exhibit visible riming”?
- L.212-213: “Other misidentified CP classes”, the formulation is unclear
- L.244 : “an precision score”
- L.268 : “A equivalent correction factor”
- L.316-317: “some pair-wise label combinations for humans (fig. 1) was significantly higher…” “were”
- L.351:”recall if a weak point”, “is” instead of “if” ?

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.