https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

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

Ezri E. Alkilani-Brown[1], Declan L. Finney[1,2], Alan M. Blyth[1,2], Paul R. Field[1,3], Jonathan Crosier[4,5], and Chetan R. Deva[1]

[1]School of Earth and Environment, University of Leeds, Leeds, UK
[2]National Centre for Atmospheric Science, Leeds, UK
[3]Met Office, Exeter, UK
[4]Department of Earth and Environmental Sciences, University of Manchester, Manchester, UK
[5]National Centre for Atmospheric Science, Manchester, UK

Correspondence: Ezri E. Alkilani-Brown (e.alkilani-brown@leeds.ac.uk)

Abstract. Untangling the ice microphysical interactions within deep convective clouds presents an ongoing issue. Cumulonimbus have implications for localised precipitation and global radiative feedbacks. *In situ* flight campaign data is informative of these interactions and can consolidate our understanding. Identifying ice particle habits illustrates the evolving cloud on the micro-scale. In particular, the development and growth of graupel continues to be the least understood hydrometeor in numerical models. Consequently, in the ever-evolving machine learning landscape, a multitude of instrument and dataset specific ice habit identification algorithms are becoming commonplace. Here, we complete a key step of independently assessing several generalised and open source algorithms, to better understand their suitability for wider uptake. Evaluation and application of generalised convolutional neural networks (CNN), created by Jaffeux et al. (2025) has been undertaken on unseen two-dimensional stereo (2D-S) and High Volume Particle Spectrometer (HVPS) images from the Deep Convective Microphysics EXperiment (DCMEX). Models were not re-tuned to the dataset. Jaffeux et al.'s global CNN tested with human labelled 2D-S images obtained an accuracy of 72% and F1 score (harmonic mean of precision and recall) of 70%. While for HVPS images, the HVPS-specific CNN had an accuracy of 86% and F1 score of 73%, which was only marginally better than the global model. Then scaling up CNN application to the whole DCMEX dataset, graupel concentrations were inferred from rimed particle classification. The models constructed by Jaffeux et al. (2025) present an accessible, accurate and adjustable approach for particle identification of optical array probe images.

1 Introduction

Cumulonimbus, the largest convectively formed cloud with serious impacts on climate and weather. The resulting anvils can have a direct effect on the total cloud radiative effect (Hartmann et al., 2018; Sokol et al., 2024; Gasparini et al., 2021) and radiation attenuation (Oberthaler and Markowski, 2013). On a local scale, heavy precipitation is common (Raymond and Blyth, 1989; Blyth et al., 2015; Duffourg et al., 2016; Varble et al., 2021) and have the potential to develop into thunderstorms. At the core of these mesoscale implications are the microphysical mixed-phase interactions, which ultimately control the

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

resulting effect. Bulk microphysics parameterisations are the typical method to represent these processes in numerical models (Grabowski et al., 2019; Liu et al., 2023), however they routinely omit some complexities e.g. secondary ice (Field et al., 2017), and understanding of mixed-phase microphysics is continuously developing. Likewise, resolving deep convection has uncertainties associated with it (Bryan et al., 2003; Gilmore et al., 2004). Together, this results in a uncertain representation of cumulonimbus. 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. *In situ* measurements provide a key means to constrain model parameterisations and improve understanding of microphysical processes.

A recent campaign looking to answer some of these questions is the Deep Convective Microphysics EXperiment (DCMEX) (Finney et al., 2024). It was a summer field campaign, that took place over the Magdalena mountains in New Mexico, comprising of complementary *in situ* and remotely sensed measurements. Developing cumulonimbus were primarily targeted, with passes through detrainment layers and anvils where possible. This built up a picture of the microphysics, dynamics and aerosol of these clouds. Moreover, a strong focus on the measurement of cloud ice properties was pursued.

Understanding the intricacies of ice habits has proved pivotal in cloud properties (Ong et al., 2024), resulting precipitation (Oue et al., 2016; Jensen et al., 2018; Allabakash et al., 2019; Huang et al., 2023) and radiative processes (Ehrlich et al., 2008; Baran, 2012; Liu et al., 2014; Yang et al., 2015; Ren et al., 2021; Yi, 2022; Wolf et al., 2023). Beginning as an ice nucleating particle, through a secondary ice processes, or homogenous freezing, ice particles can grow via vapour deposition, accretion or aggregation. Classification of these particles have a long history, for example Nakaya (1954); Magono and Lee (1966). 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.

Accretion can culminate in graupel and hail particles. These particle names are often used interchangeably, but density can distinguish them. Size threshold is also used to separate particles, as defined in AMS, graupel is larger than $2mm$ (American Meteorological Society, 2024a) and hail greater than $5mm$ (American Meteorological Society, 2024b). However this references ground observations; currently there are no in-cloud size thresholds for these particles. Graupel is a product of riming, where supercooled cloud drops quickly freeze on the surface of an ice particle, associated with low density deposits. Whereas hail can accumulate droplets that freeze slowly, associated with high density deposits. Hail exhibits density close to that of bulk ice, around 0.8 to 0.9g cm^{-3} (List, 1958; Browning et al., 1963; Prodi, 1970). Graupel density is a lot more variable, typically lower than 0.5g cm^{-3} (Locatelli and Hobbs, 1974; Heymsfield, 1978; Heymsfield et al., 2018) but also reaching 0.7 to 0.8g cm^{-3}(Braham Jr, 1963; Knight and Heymsfield, 1983; Heymsfield and Kajikawa, 1987). The process of riming also presents difficulties for identifying graupel particles. Distinguishing when an ice particle has been rimed 'enough' to be classified as graupel is extremely nuanced. There have been several attempts at defining a riming degree e.g. Reinking (1975); Mosimann et al. (1994); Colle et al. (2014); Takami et al. (2022) often relying on an arbitrary scale or percentage coverage. A quantitative threshold for when graupel has been created has not been outlined. Instead, relying on a qualitative description, when riming has deformed the original ice particle beyond its original shape, this is a graupel particle.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

To view particles *in situ*, for decades optical array probes (OAP) have been utilised on the underwing of research aircrafts. These instruments consist of a linear array of photodetectors and corresponding laser opposite. The photodetectors are orthogonal to the direction of flight. If a particle is intercepted, the laser is occulted, reducing the light intensity received by the photodetectors. If this intensity decreases by a set threshold, the change is recorded, producing a 'shadow image' of a particle. Some instruments have one intensity threshold which produce black and white images, and others have several, capable of producing greyscale images. The two-dimensional stereo (2D-S) and High Volume Particle Spectrometer (HVPS) are an example of black and white image producing OAPs. One drawback to these images is the lack of depth and textural information about the particle. This is particularly a problem for understanding the orientation of the particle, for example a plate viewed from the side may appear like a column.

Nonetheless, OAP images provide a wealth of information about ice particle size and shape. Correspondingly, building accurate ice particle identification schemes for these images is ongoing. Early classification of OAP images has often used Fourier descriptors e.g. Rahman et al. (1981); Duroure (1982); Hunter et al. (1984); Moss and Johnson (1994), since dendrites, plates and columns have strong periodicities to identify them. To differentiate more complex particles, more advanced schemes are needed. 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). Furthermore, with the development of machine learning (ML), similar image descriptors have successfully been passed to ML algorithms to make a classification e.g. Garbrick et al. (1995); Grazioli et al. (2014); O'Shea et al. (2016). As ML has advanced, convolutional neural networks (CNNs) are becoming common, where image data can be directly parsed to an algorithm and classified, avoiding the feature extraction process, and which Touloupas et al. (2020) has shown to be more accurate.

CNNs are a type of supervised ML, where pre-defined classes are the objective of classification. A couple approaches can be taken in CNN development. Either creating a model from scratch, or using a pre-existing and non-specialised algorithm (Xiao et al., 2019; Wu et al., 2020) to be re-trained on each dataset. Likewise, CNNs are specialised to one type of image e.g. a high resolution multi-angled camera (Hicks and Notaroš, 2019) or holographic imager (Zhang et al., 2024) is inapplicable to any other instrument. This has resulted in a growing number of closed-source, hyper-specialised classification algorithms, applicable to one instrument on one dataset (e.g. Praz et al. (2017); Wu et al. (2021); Zhang et al. (2024)). With researchers repeatedly training and testing a new algorithm for every study, significant time and resources are required before answering any scientific question. 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.

Accordingly, this paper presents an evaluation and application of CNNs created by Jaffeux et al. (2025). First, a sample of unseen 2D-S and HVPS images from the DCMEX campaign have been classified by an independent group of human labellers, and parsed to respective 2D-S, HVPS and global CNN models. Then a focus of extracting graupel information from image

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

data has been outlined, before scaling up and applying the CNNs to the whole DCMEX dataset, to characterise the ice particles observed on the campaign.

2 Methods

2.1 DCMEX campaign

The DCMEX campaign (Finney et al., 2024) sampled around the Magdalena mountain range in New Mexico (US), during July and August of 2022. This landscape lends itself as a natural laboratory, with the orography forcing convection during the summer. The focus of DCMEX was to constrain the microphysical uncertainty surrounding cumulus clouds and their respective anvils. A combination of *in situ* aircraft, ground-based, and remote sensing data was gathered to investigate the microphysics and dynamics of these clouds, along with complementary aerosol composition. There were a total of 18 flights, in the following analysis only 17 are considered.

For investigating the physics and dynamics, two strategies were undertaken. Repeated sampling near congestus turrets tops, and sampling within or slightly above the Hallett-Mossop temperature range i.e. between -3 to $-10°C$. Cloud passes often doubled back on route, starting as close as possible to cloud base, then ascending to a higher altitude, following close to the top of developing thermals, to characterise the evolving particles. The Facility for Airborne Atmospheric Measurements BAe-146 aircraft was utilised during the campaign. The aircraft was unable to navigate through the most radar reflective regions of the clouds, and likewise sampling ceased prior to electrification. Thus sampling avoided the most convectively rigorous areas of cloud.

2.1.1 Imaging instruments

For *in situ* imaging of particles, the 2D-S and HVPS were operated during DCMEX. The 2D-S (Lawson et al., 2006) has a $10\mu m$ resolution and a 128 photodiode array, with a nominal sampling range of $10-1280\mu m$. The 2D-S also has two orthogonal views of the same sample of air, described as channel 0 (ch0) and channel 1 (ch1), providing complementary information. Anti-shatter tips (Korolev et al., 2011) were also fitted to the 2D-S, to reduce shattered particle artefacts. The HVPS (Korolev and Isaac, 2005) has a $150\mu m$ resolution and 128 photodiode array and resulting sample size range of $150-19200\mu m$.

2.2 CNN application

2.2.1 Jaffeux description

The initial publication from Jaffeux et al. (2022), has subsequently been updated in a second paper, Jaffeux et al. (2025). Thus models from the second paper have been applied. Jaffeux et al. have produced five CNN models, four instrument specific i.e. the CNN trained on images specific to that instrument for the 2D-S, HVPS, Precipitation Imaging Probe (PIP) and Cloud

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Imaging Probe (CIP). The other CNN, described as the global model, has been trained on images from all OAPs i.e. 2D-S, HVPS, PIP and CIP. Further description of training and model structure is outlined in their work.

Defined in Jaffeux et al. (2022), there are nine morphological classes: compact particles (CP) – heavily rimed and compact crystal shape / graupel; fragile aggregates (FA) – irregular, likely aggregated, unrimed, weak bridges; columns and needles (Co) – singular columns, needles or sheaths; hexagonal planar crystals (HPC) – singular stellar dendrites or plates; rimed aggregates (RA) – large, likely aggregated, heavily rimed; combination of bullets or columns (CBC) – aggregates of bullet rosette and/or columns; complex assemblages (CA) – aggregates of plates, columns, dendrites; capped columns (CC) and water droplets (WD) as named. Alongside defined particle classes are two miscellaneous classes of diffracted (Dif) and shattered particles (SP), to describe images which have been distorted by diffraction and images which contained shattered particles respectively. Due to instrument specifications 2D-S and HVPS do not use all classes, i.e. 2D-S images can be labelled as CP / FA / Co / HPC / CBC / CA / CC / WD / Dif, whereas HVPS images can be classified as CP / FA / Co / HPC / RA / CBC / SP.

Three of the five CNN models will be evaluated on 2D-S and HVPS images from the DCMEX campaign; the 2D-S-specific CNN, the HVPS-specific CNN and global CNN. For this application, models have not been re-trained on the DCMEX data, making the entire DCMEX dataset an unseen test set.

To implement these CNNs, a minimum size diameter is required for particles to be classified; 2D-S requires a minimum $300\mu m$ (30 pixels) and HVPS $3000\mu m$ (20 pixels). It should be highlighted that Jaffeux et al. minimum diameter refers to the maximum extent of the particle, while our definition of diameter is an average of the extent of the crystal along / orthogonal to flight direction. All 2D-S and HVPS images were created to the appropriate 200 by 200 pixel size for CNN use. Also, the RA class is not present in the global model, instead images are categorised into CP. In addition, the CNNs produce a probability for each class for each image, and the highest probability class is considered to be the final label.

2.2.2 Data cleaning

From the DCMEX dataset, only particles which meet the 'entire-in' criteria have been used. I.e. the particle was not imaged on the top or bottom photodiode of the instrument. The sample volume has correspondingly accounted for this; see appendix B. Filtering of images with small inter-arrival times (Field et al., 2006) has also been applied, to remove potentially shattered particles. Particles with an aspect ratio larger than 10 were also omitted. This was due to numerous artefacts fitting this criteria, and predominantly excluded columns. Small diffraction effects were also accounted for, where particles with small holes (fewer than six pixels in extent) within the crystal boundary were also filled in.

2.2.3 Testing procedure

For robust evaluation of a CNN, a consensus label across multiple human labellers is best practice. This reduces the likelihood of miscategorising the particle. However, within the constraints of this study, only a subsample of the evaluation set had a multi-label consensus. The remainder of images were labelled by one person. The final test dataset was the combined group consensus and single label images. Nonetheless, analysis of the multiple labels assigned to one image has highlighted porous classes, and provided uncertainty bounds for the final human labels. From the 17 DCMEX flights, 10,000 random 2D-S images

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

were sampled across both channels, of this 9988 were identified as being non-artefact. Of 9988, a subsample of 900 were labelled by four co-authors. To construct a consensus from four labellers, a minimum of two labels must agree, however two pairs of conflicting labels were classed as non-consensus. Analytically, the probability of constructing a consensus is 50.6% for nine 2D-S labels and 56.7% for seven HVPS labels. From the 900 2D-S images, 845 had group consensus (93.8%). The other 9088 images were labelled by one person, totalling to 9933 2D-S images used in evaluation. For HVPS images, 1500 images were randomly sampled across DCMEX and 1445 were identified as non-artefact. From 1445 images, 200 were subsampled again for four labels; this resulted in 185 images (92.5%) with group consensus. Altogether with the other 1245 images labelled by one individual, 1430 HVPS images comprise the basis of our evaluation. The images are a representative sub-sample of particles observed during the DCMEX campaign.

An analogous experiment using a random assignment of labels has also been completed, in order to confirm human labelling results are distinct from random noise. The same sample of 2D-S and HVPS images i.e. respective 9988 and 1445 non-artefact images, were assigned four random labels, to emulate human labellers. From this, an equivalent 'random consensus' was obtained; 5137 images (51.4%) for the 2D-S and 880 (60.9%) for HVPS, these values sit close to the analytical probability.

2.2.4 Evaluation metrics

Four common CNN evaluation metrics were used: accuracy, precision, recall and F1 score. Equations are stated in the appendix (eq. C1 to eq. C4). Metrics are applied per CNN class, and the mean across all classes provide an overall score for performance. Accuracy describes the fraction of predictions the model got correct; this includes the true negative results i.e. correctly identifying images outside the class. Precision outlines the proportion of correctly identified images out of all images identified with that class. For example, low precision indicates the CNN is over predicting the class. Recall describes the proportion of all images in a particular class that were correctly identified by the CNN. This highlights what is being missed in predictions. The F1 score is a harmonic mean of precision and recall, depicting the balance of precision and recall for that class. This provides a better metric to understand CNN skill.

2.2.5 Graupel approximation

The defined classes from Jaffeux et al. present a unique opportunity to quantify graupel from OAP images. Previously, an automatic and systematic approach for identifying rimed particles from images has not been available. Formed from riming, the surface texture and deformation of shape can inform if graupel was observed. By definition, CP and RA are the only classes that have had visible riming taking place and are the "archetypes of dense ice particles" (Jaffeux et al., 2025). For our observations of growing cumulonimbus, if a particle shape was not recognisable enough to fit into a non-rimed class, we can assume significant riming has taken place, and can be used to approximate graupel. Subsequently, images identified as CP or RA will be used as an proxy for graupel.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

3 Results

Results are split into three sections, first outlining the difference of human opinions when labelling OAP images (sect. 3.1). Then evaluation of Jaffeux et al. CNN models with a DCMEX sub-sample of 2D-S and HVPS images (sect. 3.2). Lastly, scaling up the application of CNNs to the whole DCMEX dataset (sect. 3.3).

3.1 Human labelling

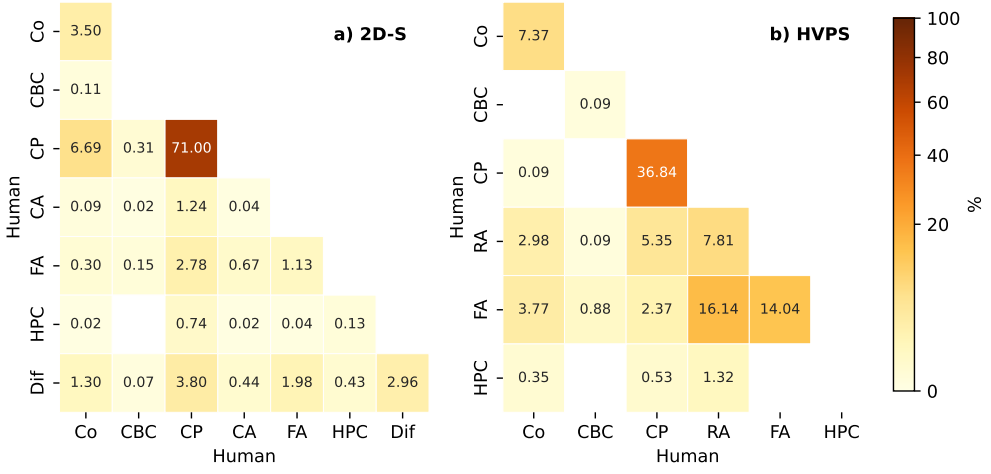

Figure 1. Confusion matrices for human labelling of a) 900 2D-S images and b) 200 HVPS images. Matrix is constructed from pair-wise combinations of four human labels for each image. Percent refers to the frequency of occurrence for each label combination e.g. 2D-S CP-CP matching labels accounts for 71% of all pairs.

Figure 1 depicts confusion matrices of human labelling disagreement for 2D-S and HVPS images. The equivalent figure for the random assignment of labels is in appendix A. 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. Here, CP was a prominent class across both 2D-S and HVPS images, all together responsible for 86.56% and 45.18% of at least one label for 2D-S and HVPS images respectively. Also conflicting CP labelling occurred in 15.56% of 2D-S and 8.34% of HVPS images. FA also proved to be a common class for HVPS images, responsible for 37.2% of at least one FA label, and 23.16% of data which had conflicting FA labelling. RA likewise was prominent in HVPS images, but had far more conflicting images (25.88%) than total ones agreeing (7.81%). Human disagreement was ubiquitous across both sets of images, with almost every possible label combination present. The highest disagreeing labels for 2D-S images include CP / Co (6.69%), CP / Dif (3.8%) and FA / CP (2.78%). HVPS conflicting labels were similar, FA / RA (16.14%), CP/RA (5.35%), FA/Co (3.77%) and RA/Co (2.98%).

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

3.2 CNN evaluation

3.2.1 2D-S

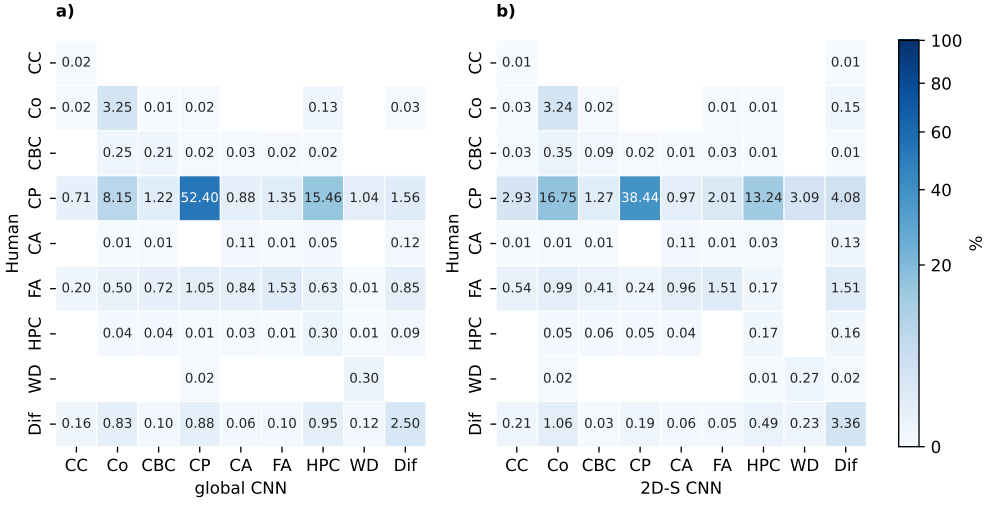

Figure 2. Confusion matrices for 9933 2D-S images, CNN versus human consensus. a) global CNN and b) 2D-S-specific CNN evaluation. Human labels are on the y axis, and CNN label on the x axis.

a)

	image count	accuracy	precision	recall	F1
CC	2	98.90%	1.80%	100.00%	3.54%
CA*	31	97.97%	5.70%	35.48%	9.82%
WD*	32	98.80%	20.41%	93.75%	33.52%
CBC*	55	97.55%	9.13%	38.18%	14.74%
HPC*	53	82.52%	1.72%	56.60%	3.34%
Co*	344	90.00%	24.94%	93.90%	39.41%
CP*	8222	67.63%	96.34%	63.31%	76.40%
FA*	629	93.71%	50.67%	24.17%	32.72%
Dif*	565	94.16%	48.53%	43.89%	46.10%
All macro avg.		91.25%	28.80%	61.03%	28.84%
All weighted avg.		72.01%	86.72%	60.63%	69.54%
*Subset macro		90.32%	32.18%	56.56%	32.24%
*Subset weighted		72.31%	86.73%	61.22%	69.97%

b)

	image count	accuracy	precision	recall	F1
CC	2	96.23%	0.27%	50.00%	0.53%
CA*	31	97.77%	5.16%	35.48%	9.02%
WD*	32	96.63%	7.56%	84.38%	13.88%
CBC*	55	97.73%	4.79%	16.36%	7.41%
HPC*	53	85.67%	1.21%	32.08%	2.33%
Co*	344	80.55%	14.43%	93.60%	25.00%
CP*	8222	55.16%	98.71%	46.44%	63.16%
FA*	629	93.06%	41.67%	23.85%	30.33%
Dif*	565	91.60%	35.65%	59.12%	44.47%
All macro avg.		88.27%	23.27%	49.03%	21.79%
All weighted avg.		61.19%	86.94%	47.21%	57.72%
*Subset macro		87.26%	26.15%	50.06%	24.97%
*Subset weighted		62.23%	86.96%	48.93%	59.17%

Figure 3. 2D-S CNN evaluation results, for a) global CNN and b) 2D-S-specific CNN. Weighted averages accounts for the sample size of each class. Habits with a sample size larger than 10 are included in the subset statistics and are marked with an asterisk.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

The CP class dominated the sample images, comprising 82.77% of all 2D-S images labelled. Consequently, this skewed results, and was reflected in the evaluation, directly impacting the assessment of other classes.

The disagreement between consensus human labels and CNN predictions for 2D-S images is presented in fig. 2 confusion matrices. The full matrix is depicted, as human and CNN labels are independent. Again the diagonal represents agreeing labels and off-diagonal is the disagreeing labels. Each row of the matrix displays the total images in that class labelled by humans. Each column is the total images classified by the CNN for that respective class. The matrix is able to highlight what human / CNN combination of labels disagree the most. For 2D-S images, the CNN models experienced similar mislabelling combinations, especially in the CP class. Fig. 2 displays this; with the most mislabelled class CP / Co and CP / HPC confusion was omnipresent across both CNNs. 2D-S-specific had a higher proportion of misidentified CP/ Co, at 16.8% while the global model had a proportion of 8.2%. Likewise CP / HPC confusion was systemic across both CNNs, with 13.2% and 15.5% misidentification for the 2D-S and global CNNs respectively. Other misidentified CP classes, to a lesser extent were Dif, WD, CC, FA across both CNNs.

Results from the confusion matrix is supplemented with evaluation statistics, as shown in fig. 3. Rows show respective accuracy, precision, recall and F1 statistic for each class. Fig. 3 includes the macro and weighted statistics across all the habits, and the equivalent for a subset which excludes habits with a low sample size.

Both CNNs shared similar strengths and weaknesses, as depicted in fig. 3. Class specific precision and recall scores vary to a substantial degree. For example CP in the 2D-S-specific model has precision of 99% but a recall of 46%. Specifically, CC, CA, WD, CBC and HPC all exhibited very low precision scores, which is reflected in their respective F1 value. 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. Focusing on the results for all-habits, the global CNN attained a weighted F1 score of 69.54%, compared with the specific 57.72%; thus the application of the global model on 2D-S images was preferred over the 2D-S-specific CNN.

For the randomly labelled images (fig. A2a)) an equivalent 2D-S images, global CNN evaluation is displayed. Interestingly accuracy was high across all classes and achieved a weighted mean of 80.31% (which was higher than the global CNN all-habit human result of 72.01%). However weighted precision, recall and F1 scores were a lot lower, at 12.01%, 10.96% and 7.67% respectively. This highlights the importance of considering a range of statistical metrics beyond accuracy. In addition, comparing to Jaffeux et al. (2025) global model evaluation with all OAP images, they obtained predominantly higher results across the board. Respectively a higher weighted precision, recall and F1 scores of 88.60%, 88.30% and 88.28%.

3.2.2 HVPS

Similar to 2D-S image population, HVPS had an uneven distribution of classes; FA had the highest percentage (42.1%), then 37.83% for CP and 14.34% for RA class. In the global model evaluation, RA was incorporated into CP. Some classes have also been omitted, due to no sample images. Alike to 2D-S results, the variety of images has influenced the overall scores for CNN evaluation. Unlike the 2D-S, no unclassifiable i.e. shattered particles were observed in our sample.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

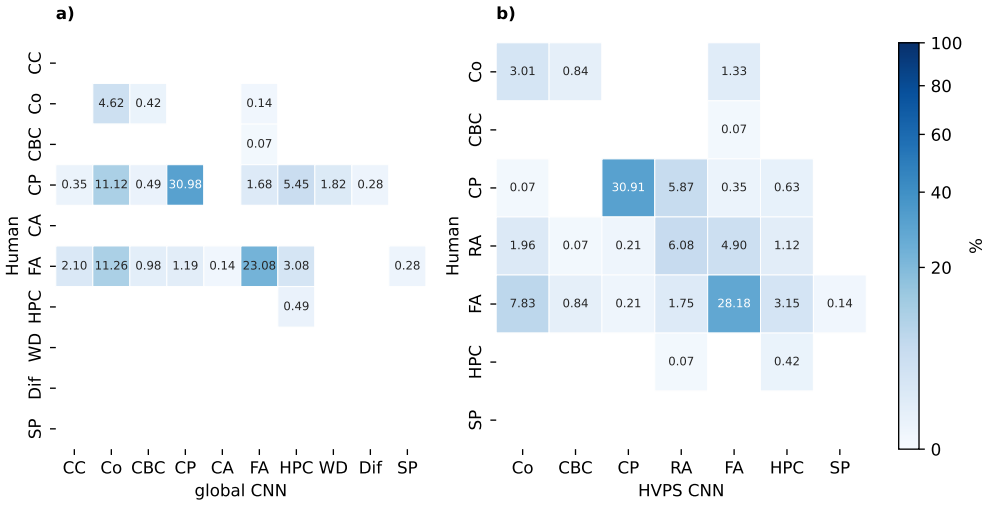

Figure 4. Confusion matrices for 1430 HVPS images, CNN versus human consensus. a) global CNN and b) HVPS-specific CNN evaluation. Global CNN does not have a RA class thus RA labelled images are placed in CP

a)

	image count	accuracy	precision	recall	F1
CBC	1	97.94%	0.00%	0.00%	
HPC	7	91.02%	5.43%	100.00%	10.29%
Co*	74	75.86%	17.10%	89.19%	28.70%
CP*	746	79.03%	96.30%	62.31%	75.66%
FA*	602	80.65%	92.44%	58.30%	71.51%
All macro avg.		84.90%	42.25%	61.96%	46.54%
All weighted avg.		79.62%	90.07%	62.15%	71.16%
*Subset macro		79.87%	68.70%	77.60%	62.35%
*Subset weighted		82.69%	90.66%	69.68%	76.65%

b)

	image count	accuracy	precision	recall	F1
CBC	1	98.18%	0.00%	0.00%	
HPC	7	95.03%	7.89%	85.71%	14.46%
RA*	205	84.03%	44.16%	42.44%	43.28%
Co*	74	87.96%	23.37%	58.11%	33.33%
CP*	541	92.65%	98.66%	81.70%	89.38%
FA*	602	79.55%	80.92%	67.17%	73.41%
All macro avg.		89.57%	42.50%	55.85%	50.77%
All weighted avg.		85.67%	78.97%	68.70%	72.77%
*Subset macro		86.79%	61.88%	68.23%	61.99%
*Subset weighted		86.85%	79.48%	73.31%	75.48%

Figure 5. HVPS evaluation results, for a) global CNN and b) HVPS-specific CNN. Global CNN does not have a RA class thus RA labelled images are placed in CP. Classes are omitted due to having no sample images. Weighted averages accounts for the sample size of each class. Habits with a sample size larger than 10 are included in the subset statistics and are marked with an asterisk.

Starting with one of the largest human labelled groups, FA, also a prominently mislabelled class. As depicted in fig. 4, FA images were identified as Co, across global and HVPS CNNs at 11.26% and 7.83% respectively. Similarly, both CNNs also labelled around 3% of FA images as HPC. Unlike fig. 1, FA and RA confusion was not prominent in HVPS CNN results, only accounting for 1.75% of results, or equivalent CP vs RA in the global model (1.19%).

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

For human labelled CP, models diverged significantly in predictions. Similar to 2D-S predictions, the global model most mislabelled CP as Co and HPC, 11.12% and 5.45% of images respectively. Whereas, HVPS-specific saw significantly less confusion with these class, and instead mislabelled 5.87% CP as RA - similar confusion to human labelling.

The statistics associated with HVPS, in fig. 5 also showed a mixed result. Again, extremes of values were observed in precision and recall across the different classes. For example HPC, in the global model attained an precision score of 5.43%, but recall of 100% describing that it identified all labelled HPC images, but also mis-classified a lot of other images as HPC too. Final accuracy and F1 scores for the global and HVPS-specific models are extremely close. Accounting for all-habits, the HVPS-specific model out-performed the global model, attaining a weighted accuracy and F1 score of 85.67% and 72.77% respectively, the global model accuracy was lower, at 79.62% but the F1 score was extremely close, at 71.16%. The sub-setted statistic further highlights this, excluding classes with small sample size, the global model performance exceeded the HVPS-specific CNN. Global accuracy was still slightly lower, but F1 score was 1.17% higher than the HVPS-specific CNN. Ultimately, the performance of both models on HVPS images was equivalent. In the subsequent application and analysis (sect. 3.3) the specific model has been applied to classify HVPS images, based on the all-habit F1 statistic.

In the randomly labelled HVPS sample (fig. A2b)), similar to panel a), random labelling attained high accuracy scores when evaluating the HVPS-specific model. In particular, reaching a weighted accuracy of 75.58%, close to the human equivalent 85.67%. Again however, weighted precision, recall and F1 scores appear much lower than human labelled. Likewise, comparing to Jaffeux et al. (2025), they again obtained higher results when evaluating the HVPS-specific CNN, with weighted precision, recall and F1 scores of 86.15%, 85.60% and 85.71% respectively.

3.3 DCMEX application

3.3.1 Graupel adjustment

CP misidentification as HPC in fig. 2 was observed previously in Jaffeux et al. (2022, 2025). This bias poses a significant issue for inferring graupel concentrations, thus adjustment to the labels of these classes should be accounted for. CP misidentification did not occur to a significant extent in the HVPS-specific model, largely only in the identification of 2D-S images; hence a correction factor has only been applied to 2D-S predictions.

$$\text{correction factor} = \frac{Human(CP) \subset CNN(HPC)}{CNN(HPC)} \tag{1}$$

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.

The correction has been applied per size bin, per second, only if HPC has been identified in predictions. The number of HPC particles has been artificially decreased by this factor, correspondingly, CP number is artificially increased by this

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

factor. Henceforth, corrected values are marked with CP† and HPC† notation, to display departure from original categorisation. Moreover, CP† values are inclusive of CP† from 2D-S and CP/RA from HVPS, and also described interchangeably as graupel.

Diffracted and shattered particles have not been adjusted across any class. These classes correspond to approximately 4% of all viable 2D-S, and 0.2% HVPS images in the DCMEX dataset. However, adjustment similar in application to HPC would result in a negligible change in concentration. Fundamentally diffracted particles are not in focus, so the diameter recorded is incorrect, accounting for this would involve a proportion change of equal probability across all classes and all sizes. Instead, these particles are ignored in final calculations, thus concentrations recorded should be interpreted as negligibly lower than reality.

3.3.2 Campaign overview

Examining 2D-S and HVPS images from the DCMEX campaign, two different models have been applied. The global CNN and HVPS CNNs have been applied to 2D-S and HVPS images respectively. Models have been chosen due to earlier evaluation, and have not been tuned to the DCMEX dataset. The graupel adjustment has also been applied to 2D-S concentrations, to account for CP misidentification. An overview of particle habits observed during the DCMEX field campaign is depicted in fig. 6, where concentrations are split by CNN classification and are aggregated over one second intervals. Distributions depict the mean concentration as a function of size bin, conditional on the presence of any particle of that class in a given one second measurement period. The sampling strategy undertaken in DCMEX is consequently reflected in the final results, where the largest sized particles in the HVPS sample range are limited. CC class has been omitted, due to insignificant numbers observed.

Overarchingly within the size ranges classified, CP† dominated the composition of clouds. Reaching some of the highest concentrations across all sizes in all-habits, and comprised roughly 76% of 2D-S particles and 63% of HVPS particles. Co was also another major habit class for 2D-S images, with 9% of particles falling into Co classification, but was not significant for HVPS data. On the other hand, FA was the second largest class (32%) for HVPS images. Other habit classes CBC, HPC† and CA comprised a minor composition of all ice seen during the campaign. CA does not have HVPS values, as this class does not exist in the HVPS CNN.

Within the size requirements of CNN application, FA and CA were the only classes to display a peak in habit populations within 2D-S measurements. Interpretation of other size distributions are unclear, and peak concentration was outside the scope for CNN classification.

Depiction of graupel values and related measurements from the DCMEX campaign are outlined in tables 1 and 2 for 2D-S and HVPS data respectively. Across all flight days, it is unanimous that CP† dominated 2D-S measurements, comprising a minimum of 50% of all particles imaged per day, excluding the 3rd of August. CP particles were also common in HVPS observations, but overall had a larger variety of prominent daily habits. After CP†, pristine habits (Co and HPC†) were common amongst 2D-S images, while aggregates were prevalent to HVPS (FA and CBC).

Highest average graupel concentration varied between instruments; the 2D-S saw the highest average concentration on the 23rd at $5.5L^{-1}$, and HVPS on the 29th at $0.0181L^{-1}$. Also on the 29th, HVPS observations saw their highest concentrations at

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

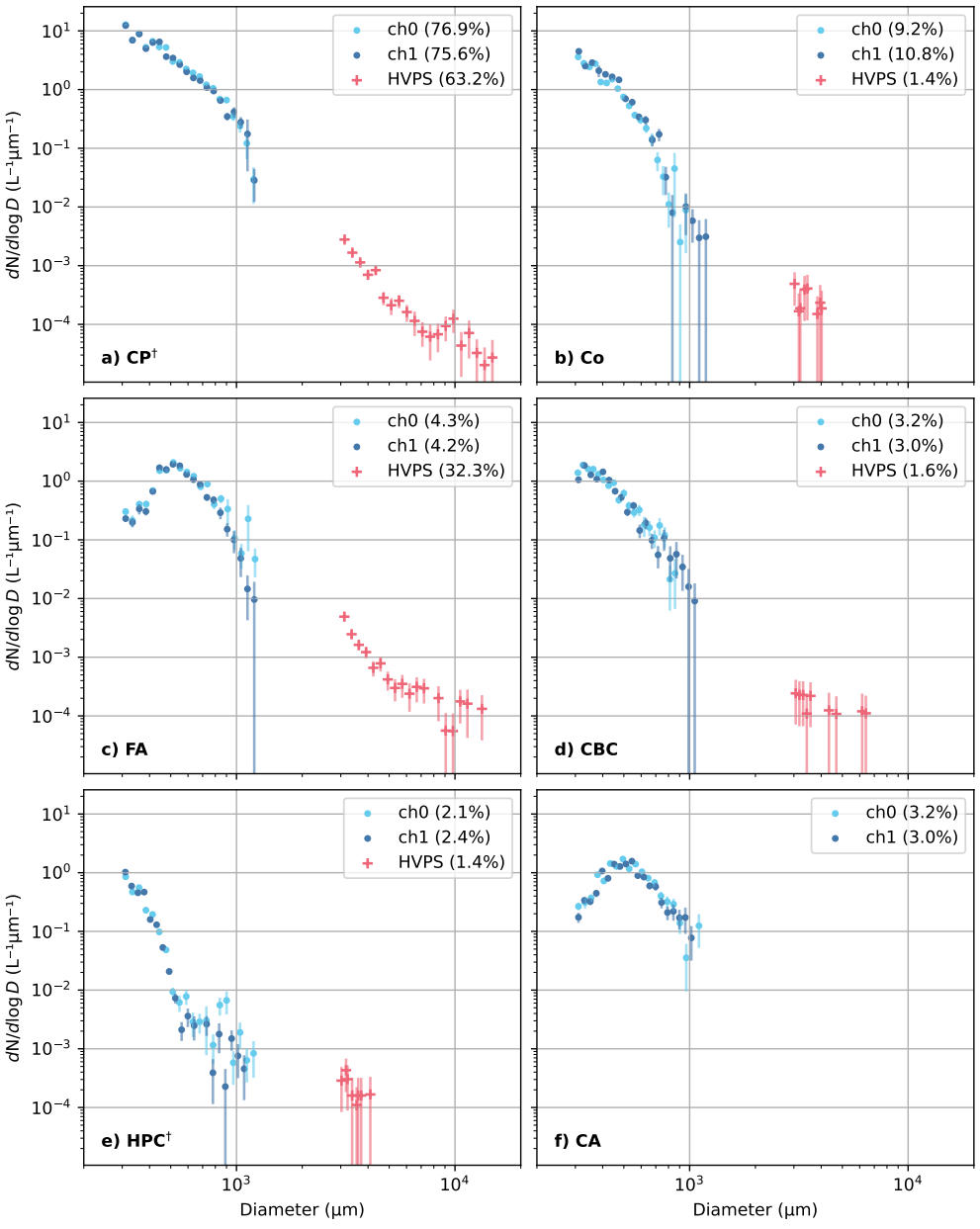

Figure 6. Mean ice crystal size distribution across DCMEX, where concentrations account for any measurement period that class was observed. Ch0 and ch1 refer to orthogonal 2D-S channels. †refers to post graupel adjustment values. Standard error bars illustrated in each size bin. Legend percentages display (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†.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Table 1. 2D-S Graupel and associated values observed during DCMEX campaign. Total images and top habits include only ice classes. CP† and HPC† reference graupel adjusted measurements. Graupel concentration and size refers to post adjustment CP†, of $300\mu m$ to $1280\mu m$ sized particles. Concentrations are from aggregated one second intervals. Mean values are of seen particles.

Date	In cloud distance (km)	Total images	Top habits (%)	Mean (max) conc. (L^{-1})	Mean (max) size (μm)
19th Jul	51.9	32359	CP† 82.5, Co 6.9, FA 2.7	3.1 (20.0)	408 (1210)
20th Jul	48.8	18805	CP† 68.9, Co 9.6, FA 7.5	3.7 (37.6)	391 (1040)
22nd Jul	65.0	10	CP† 55.2, Co 40.0, HPC† 4.8	0.1 (0.2)	430 (750)
23rd Jul	166.4	30654	CP† 84.1, Co 7.5, FA 3.0	5.5 (25.2)	425 (1250)
25th Jul	78.3	21854	CP† 53.0, Co 18.0, FA 9.3	2.9 (14.1)	375 (990)
26th Jul	137.7	1299	CP† 67.6, CBC 10.2, CA 9.9	0.4 (3.6)	420 (1180)
27th Jul	107.2	369	CP† 75.2, Co 11.6, CBC 4.9	1.3 (11.8)	453 (1050)
29th Jul	153.1	235	CP† 90.0, Co 4.3, HPC† 2.8	1.1 (3.7)	402 (1240)
30th Jul	122.6	4709	CP† 87.8, Co 8.2, HPC† 2.7	3.7 (32.9)	393 (1130)
1st Aug	221.7	3202	CP† 86.2, Co 10.6, HPC† 1.9	0.8 (15.1)	438 (1080)
2nd Aug	217.3	2657	CP† 85.1, Co 12.0, HPC† 1.8	1.5 (17.0)	448 (1180)
3rd Aug	6.6	49	Co 36.0, CP† 34.9, CBC 14.0	0.1 (0.3)	433 (770)
4th Aug	115.0	7	CP† 84.0, FA 14.3, HPC† 1.7	0.2 (0.3)	445 (660)
6th Aug	106.3	5884	CP† 77.7, Co 18.1, HPC† 2.2	1.1 (17.4)	415 (1060)
7th Aug	138.4	3799	CP† 86.2, Co 5.4, HPC† 2.4	1.4 (13.7)	408 (1020)
8th Aug	168.9	5457	CP† 86.8, Co 6.5, HPC† 3.1	2.6 (27.0)	376 (910)

$0.0394L^{-1}$, but the 2D-S highest recorded concentration was on the 20th at $37.6L^{-1}$. Considering the size of graupel particles; the 2D-S maximum observed size was on the 23rd at $1250\mu m$, while the HVPS occurred on the 7th, at $15.45mm$. Unpicking the relationships of size and concentrations across instruments is beyond the scope of this paper. Future work examining cases in detail will explore the sampling strategy and other physical mechanisms responsible for observations.

4 Discussion

4.1 Human confusion

Visualising disagreement between humans presents a unique opportunity to underpin uncertainty in ambiguous optical array probe images. However in the DCMEX dataset, riming was ubiquitous throughout campaign images, consequently habit classification has been challenging. This is best illustrated within human labelling, where each labeller has interpreted Jaffeux et al. class definitions differently. Hence, the group consensus for both sets of 2D-S and HVPS images did not reach 100%. Nonetheless, comparing to randomly assigned labels, some pair-wise label combinations for humans (fig. 1) was significantly

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Table 2. HVPS Graupel and associated values observed during DCMEX campaign. Total images and top habits include only ice classes. Graupel concentration and size refers to CP and RA combined value, of $3000\mu m$ to $19200\mu m$ sized particles. Concentrations are from aggregated one second intervals. Mean values are of seen particles.

Date	In cloud distance (km)	Total images	Top habits (%)	Mean (max) conc. (L^{-1})	Mean (max) size (μm)
19th Jul	51.9	60	CP 66.7, FA 30.0, CBC 3.3	0.0037 (0.0131)	4198 (10950)
20th Jul	48.8	78	CP 83.3, FA 11.5, Co 5.1	0.0068 (0.0271)	3794 (7200)
22nd Jul	65.0	9	CP 66.7, CBC 22.2, FA 11.1	0.0023 (0.0040)	4465 (8630)
23rd Jul	166.4	99	CP 50.5, FA 48.5, HPC 1.0	0.0037 (0.0147)	4132 (9600)
25th Jul	78.3	33	FA 60.6, CP 30.3, CBC 6.1	0.0019 (0.0022)	3595 (5700)
26th Jul	137.7	10	CP 50.0, FA 40.0, Co 10.0	0.0019 (0.0025)	4924 (6980)
27th Jul	107.2	39	CP 56.4, FA 38.5, CBC 5.1	0.0029 (0.0054)	4240 (8030)
29th Jul	153.1	50	CP 86.0, FA 14.0	0.0181 (0.0394)	4204 (9830)
30th Jul	122.6	29	CP 82.8, FA 17.2	0.0029 (0.0059)	4337 (9750)
31st Jul	125.3	202	FA 50.0, CP 47.5, CBC 1.0	0.0052 (0.0239)	4583 (12980)
1st Aug	221.7	43	CP 69.8, FA 25.6, CBC 2.3	0.0035 (0.0075)	3825 (9000)
2nd Aug	217.3	72	CP 83.3, FA 16.7	0.0049 (0.0206)	3704 (10880)
3rd Aug	6.6	20	HPC 50.0, CP 30.0, FA 20.0	0.0025 (0.0026)	3178 (3380)
4th Aug	115.0	2	CP 100.0	0.0024 (0.0025)	3375 (3450)
6th Aug	106.3	30	CP 80.0, FA 13.3, Co 6.7	0.0031 (0.0093)	3580 (4580)
7th Aug	138.4	40	CP 70.0, FA 22.5, CBC 5.0	0.0039 (0.0105)	5121 (15450)
8th Aug	168.9	13	CP 100.0	0.0043 (0.0113)	3371 (4280)

higher than random (fig. A1), especially for matching diagonal labels. This indicates the humans have better skill than a random assignment of labels.

However, there was still disagreement for some classes. CP and RA are an example of this. Despite being defined as the only classes which have explicitly experienced riming, a clear description on a 'riming threshold' would clear up ambiguity. This is especially the case for 2D-S CP / Co and HVPS RA / FA conflicts. In the definition of Co riming explicitly should not be present on these crystals, and are exclusively grown from deposition; labellers are then conflicted whether the column is 'too-rimed' to fit into the Co class, but the crystal surface is not deformed enough to match with CP definition. However, some training images from Jaffeux et al. contain columns which are not exclusively depositional growth, which could imply this class is broader than what is explicitly defined.

Likewise the differentiation of RA and FA is important for the quantification of graupel; but experiences similar difficulties in HVPS images. Both are aggregate classes, but unpicking the nuance of riming on low resolution images explains the high conflict in human image labelling, as a fine line distinguishes classification. Similarly, disagreement between FA and CP

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

is another important combination to highlight, across 2D-S and HVPS images. Despite being quite distinct in terms of the processes that define the classes, overlap in images was common.

Human disagreement can be explained through the features of the crystal. For example a particle which has a predominantly round and compact shape, but contains some protrusions consistent with aggregation. Objectively the image could fit into both, but with binary classification, a decision must be made to which parts of the particle are more important in categorisation.

Lastly, the disagreement in labelling diffracted 2D-S crystals presents a unique challenge. The diffracted class does not have any explicit definition, only if particles display patterns consistent with 'diffraction effects'(Jaffeux et al., 2022) they should be identified as this class. This is open-ended, but difficult to resolve with a single unifying and comprehensive definition. Diffraction patterns can vary hugely, as observed in Vaillant de Guélis et al. (2019), who simulated diffraction pattens of four particle shapes. Consequently, the CNN training dataset may not contain an exhaustive possibility of diffracted crystal appearances. This was observed in DCMEX images, where the variety of diffracted images did not align exactly with training images, consequently predictions across global and 2D-S-specific models struggle to capture all human labelled dif images. To correct this bias, one option is to create a new CNN, trained on synthetic data of all diffraction pattens, which could be applied before classification to remove distorted particles.

Ultimately the ambiguous nature of OAP images coupled with uncertain human perception of ice particle features means that there will often be some limitations in the assessment of CNN performance. Using fewer, more distinct classes would reduce these limitations but at a compromise of the breadth of variety for some studies. Overall, we feel that a good balance has been struck with the Jaffeux et al. (2022) classes, as authors applying these models can merge classes if required to reduce ambiguity in their study. The consequence of ambiguity is that the maximum skill the CNN can achieve in this evaluation is not perfection, but constrained by the ambiguity of human labelling.

4.2 CNN evaluation

Starting with one of the most prevalent classes CP; all CNN models exhibit extremely high precision (i.e. all greater than 96%) for the CP class, indicating they have a strict definition of features it associates with this class. However, recall if a weak point, i.e. the CNNs are missing the identification of some CP particles. Low recall values for 2D-S images can be explained through misidentification as Co and HPC. The separation of CP and Co should be distinct, but false negatives could be explained by disagreements in human labelling. The degree of riming has played an important part in this. There has been no clear description of when riming has reached the threshold for ice to be classified as graupel e.g. Reinking (1975); Mosimann et al. (1994) . So, crystals may appear rimed and slightly rounded, but have retained a semi-identifiable rectangular shape, the CNN has instead placed the image into the Co class, depicted in fig. 7. Surprisingly, this misidentification of CP as Co is present in HVPS images classified by the global model, despite the considerable difference in HVPS Co and CP images. As the global CNN has trained on OAP images of different resolutions, the features it associates with Co transcend resolution and origin of the images, thus result in mislabelling.

Following the same logic, mislabelling other particles as Co is present across a couple of HVPS classes. Despite Co being defined as exclusively vapour deposition grown, training data also included images which had been slightly altered i.e. riming

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

and very minor aggregation had occurred to the particles, but still overall identifiable as columns. A combination of vague images in porous classes and mismatch between class definition and training images can explain low Co precision and low recall for FA, CP and RA across all CNNs. Also depicted in fig. 8, RA and FA are commonly misidentified. This can be attributed to the coarse resolution of the images, and the nuance of riming degree distinguishing them; with confusion experienced by human labellers and HVPS CNN alike.

However, misidentification of CP particles as HPC in global and 2D-S CNN presents a more serious matter and less easily resolved problem. Highlighted in Jaffeux et al. (2025), the lacking surface details contribute to unclear distinction between classes. From the CNN perspective, both CP and HPC exhibit rounded 'lobe' patterns on the boundary of the crystal image, shown in fig. 7, consequently the feature maps of CP and HPC created during the convolution process must be highly similar. Resolving this misidentification is challenging, images in the classes are porous, but are distinct in formation and ice particle properties, this is especially important where CP is used to infer graupel concentrations. One approach to correct misidentified HPC is to apply a post inference adjustment, as completed in this work. Alternatively merging of ambiguous classes would remove confused predictions, but for example the resulting CP/HPC class would be difficult to interpret physically.

HVPS identification of CP as RA was likewise common in the HVPS CNN, but culminates in unproblematic results. These

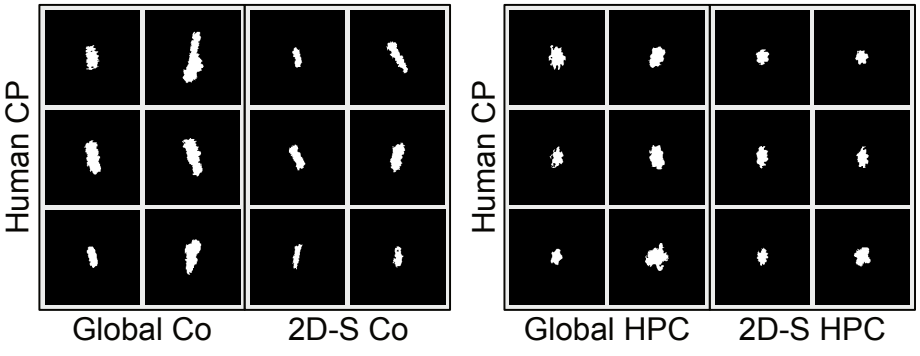

Figure 7. Example confused 2D-S images from the human CP class, with respective Co and HPC labels from the global and 2D-S-specific CNN models.

classes have large overlap, with size being the only distinguishing feature. This confusion was also experienced during human labelling, which may partially explain why the CNN made an 'incorrect' inference, where the underlying 'true' human label may be debatable.

Lastly, the 2D-S human identified Dif class; which does not encompass many images, but may have serious implications for CNN predictions. With the potential that CNN models have not been trained to pick out all diffraction patterns, consequently diffracted images are labelled as viable crystals. Over large datasets, this could result in incorrect size distributions and concentration for a given class.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

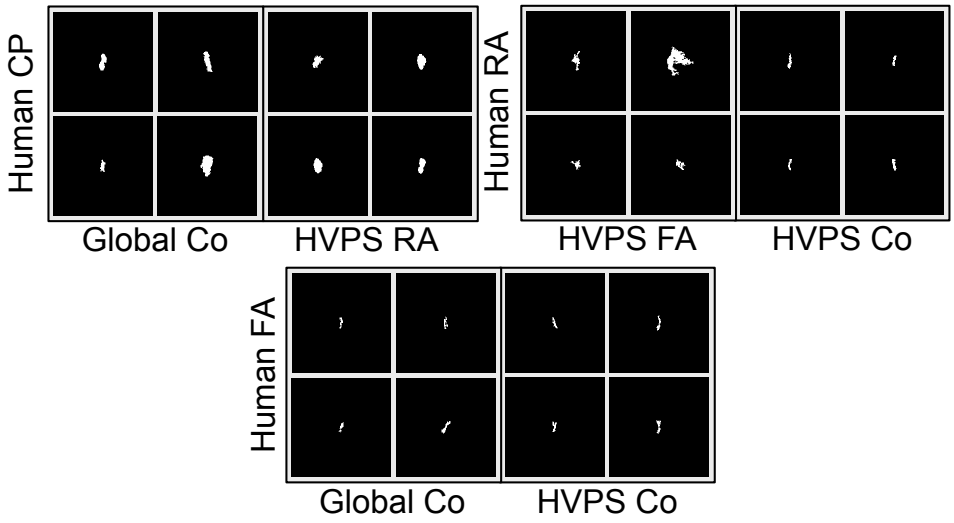

Figure 8. Example confused HVPS images from the human CP, FA and RA classes, with respective misidentified labels from the global and HVPS-specific CNN models.

Classes which did not compose the majority of images sampled, namely CC, CBC, CA, WD for 2D-S and HPC across both 2D-S and HVPS, have explainable extremes in precision and recall. The uneven class sizes have skewed results for the classes of very small size. For example the CC class for the 2D-S global CNN, where only two CC images were labelled, obtained a precision score of 1.8% and recall of 100%. This means the CNN identified both CC images, but also misidentified many others into this class. The interpretation of precision and recall of classes with a sample size less than 100 images should be considered with caution.

For 2D-S images, despite not attaining perfect performance scores, the global model has proven to be a superior CNN (over the 2D-S-specific model) in the identification of particle habits. Systematically attaining a higher weighted accuracy, recall and F1 score than the 2D-S-specific CNN. The specific model has a slightly higher precision score, which can 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. Although the global model did not attain a perfect accuracy or F1 score, this cannot be expected. This deviation can be explained through disagreements in human labelling, coupled with permeable class definitions and the ambiguous nature of images. Accordingly, this explains the lower precision, recall and F1 scores, when compared to the original evaluation from Jaffeux et al. (2025). Having defined the class boundaries, and consistent labels across the training and test images, the evaluation from Jaffeux et al. establishes the best performance the models could attain. For the DCMEX sample tested, the distribution of classes were uneven, which again may have contributed to lower weighted statistics. Nonetheless, in this evaluation of the global model with all 2D-S images,

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

obtained an accuracy of 72% and F1 score of 70% which proves to be a satisfactory result for the classification of 2D-S images. On the contrary, both global and HVPS-specific models achieved similar test statistics for HVPS image classification. Considering all-habits, the HVPS-specific CNN slightly outperformed the global model. The resolution of the images may have contributed to this. To match required CNN image size, coarser resolution HVPS images have more padding around the particles than 2D-S, which may have made the learning process harder. As the global model trained on all OAP images, the weights and biases from higher resolution images may have skewed final HVPS classification. The misidentification of particles as Co is a prime example; HVPS columns are very slender and fragile, but application of the global CNN misplaces the larger FA and CP particles into the Co class, seen in fig. 8. Nonetheless, comparable to our 2D-S evaluation, obtaining perfect accuracy and F1 scores is not possible, accounting for the nuance of class definitions and unclear images. Therefore, both the global and HVPS-specific CNNs achieve comparable results for habit identification in HVPS images. Ultimately, the global (and optionally HVPS-specific) CNNs created by Jaffeux et al. (2025) for habit identification of 2D-S and HVPS images can be recommended to be implemented in future research.

In this evaluation we have used the CNN predictions as expected to be commonly applied, i.e. using the CNN class of highest probability. However, the CNN provides the probability of each class for every image. As such, exploring the range of probabilities may capture the extent of ambiguity in human labelling. If the CNN could skilfully capture this ambiguity, it would be valuable to explore how such information could be used in analysis of observations. Likewise, we have neglected to evaluate respective models with CIP and PIP images, so applying the global CNN (or other instrument-specific CNN) cannot be implied with any certainty.

4.3 DCMEX campaign ice composition

Application of the CNNs provide interesting (and the first ever) insight to ice crystal composition across the DCMEX campaign. CP class and by extension, graupel particles have significant presence throughout all flights in the campaign. Distributions also give strong indication of the sampling strategy employed, with lowest concentrations of the largest particles. Size distributions in fig. 6 illustrate sensible predictions by the global and HVPS-specific CNN. 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. Similarly, Co, CBC and HPC† alignment is promising, but with fewer HVPS, there is slightly more uncertainty in interpretation. Also promisingly, HPC† graupel adjustment appears proportional, with good alignment of 2D-S and HVPS.

For the proportions of habits seen across instruments, CP has dominated throughout this campaign across both instruments. The second most common habit for 2D-S is Co, all other classes are approximately similar. Many observations of columns could be attributed to the sampling strategy, targeting the 'Hallet-Mossop' zone of the cloud i.e. around $-5°C$, which is the temperature regime for column growth. FA is the second most common class for HVPS images. While pristine Co and HPC† particles did not reach the maximum size possible for HVPS images. Both these observations make physical sense, where 2D-S would be expected to observe a higher proportion of young pristine habits. With the $3000\mu m$ minimum size requirement,

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

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.

The different class distributions have distinct appearances. HPC† is a particularly unusual shape, explainable through post-inference adjustment of CP/HPC. Where applying a fixed adjustment value across all size ranges has had a disproportional effect on the largest particles, which had the lowest initial count. The peaks of size distributions are interesting within FA and CA classes. This could be attributed to them being aggregate classes, where by definition they reach larger sizes than pristine particles. Unusually however, this does not apply to the other aggregate class CBC. One limitation of using these CNNs is the maximum size requirements for classification. Seen in CP†, Co, CBC and HPC†, there was no peak in concentration within 2D-S observations. The only inference to conclude is that these classes have very small particles, and understanding minimum sizes of habits is not viable.

Breakdown of DCMEX flight days, in tables 1 and 2, corroborate the mean habit distributions. The sampling strategy used during DCMEX avoided the areas of highest radar reflectivity, thus habit distributions and graupel concentration is representative of early to middle stage growth of convective cloud, not capturing the later development of the largest particles. Likewise, the total in-cloud distance does not relate to the number of particles observed, owed to the sampling strategy and minimum size requirements of the CNNs.

This work only provides broad information about the types and proportions of particles observed during DCMEX. Further understanding of graupel and ice particles observed during DCMEX will be undertaken in future work, exploring cases in more detail.

5 Conclusion

Evaluation of three CNN models from Jaffeux et al. (2025) has been undertaken by independent researchers from a different institution. 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 will have strongly impacted the types of particles observed, thus the types of particles used to evaluate the CNNs.

For 2D-S images, the global CNN was best. Considering all-habits, the global CNN obtained an accuracy and F1 score of 72% and 70% respectively, compared with the 2D-S-specific CNN 61% and 58%. 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.

In contrast, both global and HVPS-specific CNNs achieved similar classification performance on HVPS images. Accounting for all-habits, the HVPS-specific CNN slightly outperformed the global model, with respective accuracy and F1 scores of 86% and 73%, while the global model attained 80% and 71%. Again, there are caveats to the evaluation results, where unclear OAP images and imprecise class definition explains the imperfect performance scores. The discrepancy between HVPS-specific and global CNNs was attributed to the different resolution of OAP instruments, in which the global model was trained on. Images

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

of different resolutions i.e. some smaller than HVPS, resulted in weights and biases better trained to classify images of higher resolution, thus explains slightly worse performance. However, if habits with a low sample size (less than 10) were removed, the global model F1 score was 1.17% higher than the HVPS-specific. Highlighting how both models are suitable for HVPS identification.

When scaling up the application of the respective CNNs to all DCMEX images, the habit distribution and types of particles encountered were reasonable and realistic. Moreover, using the combined rimed classes, CP and RA across 2D-S and HVPS images, we have been able to infer graupel concentrations within the particle size bounds of the instruments and CNNs.

For the further development of these CNNs, clearer definition of the boundaries between classes would be advantageous. By nature, OAP images are challenging to interpret, lacking depth and detail, with many particles fitting between one or more classes. In DCMEX, riming was ubiquitous, but there is no defined riming degree to separate classes. Consequently, there was mismatch between defined classes and training images. Moreover, this image classification scheme is the first to account for diffracted OAP images. However, the Dif class may be constricted. Relying solely on real images for training, they have not seen the full scope of all possible diffracted particle patterns, so are potentially bias in predictions. Instead using synthesised images of all patterns may counter this issue.

Overall, these models present an exciting and generalised tool in the classification of OAP images, which has the potential to reduce the computational burden required to utilise these observations to answer scientific questions. 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 graupel particles.

Appendix A: Random labelling

Further information regarding the random labelling of images is outlined below. For the pair-wise assignment of labels, analytically, the probability of choosing any two labels (order of label is important) is 1.23% for the 2D-S and 2.05% for HVPS. As order of labels is not important the off-diagonal is doubled; which is approximately shown in fig. A1, from one random assignment of labels. From this pair-wise distribution of labels, the set of random labels generated can be used as an exemplar. fig. A1 caption references 'images', as random labels were assigned without knowledge of the content of OAP images. From the random consensus of labels, two example metric comparison tables are depicted in fig. A2. 2D-S images - global model and HVPS images - HVPS-specific model were chosen as they were the final assigned models in the DCMEX application. The random assignment of labels has resulted in classes of approximately the same sample size. As this is the same image sample that humans labelled, the uneven distribution of classes was captured by the CNN and is reflected in the final assessment. For example CP was the biggest group for 2D-S images (fig. 3), so in corresponding statistics shown in fig. A2a), the CP row has slightly different results to other classes. Fig. A2 highlights the nature of the accuracy statistic. Despite a completely random assignment of labels, both CNNs result in high accuracy of 80% and 75% for respective global and HVPS-specific CNNs. However, the precision, recall and F1 score show a more authentic reflection of a random label assessment of the CNNs. With no skill in image labelling, the models do a poor job at image identification.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

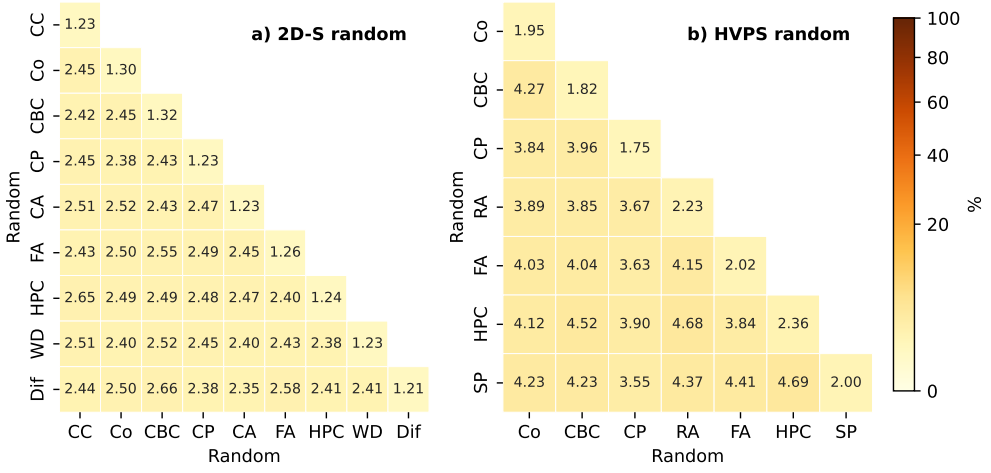

Figure A1. Confusion matrices for random labelling of a) 9988 '2D-S images', b) 1445 'HVPS images'. Matrix is constructed from pair-wise combinations of four random labels for each 'image'. Percent refers to the frequency of each label combination.

a) 2D-S random global CNN

	image count	accuracy	precision	recall	F1
CC	556	88.28%	12.90%	1.44%	2.59%
Co	594	78.18%	11.19%	12.79%	11.94%
CBC	602	86.82%	17.95%	3.49%	5.84%
CP	551	46.78%	10.50%	52.63%	17.50%
CA	578	87.02%	9.17%	1.73%	2.91%
FA	555	87.09%	14.00%	3.78%	5.96%
HPC	580	74.87%	11.48%	18.28%	14.11%
WD	569	87.93%	11.94%	1.41%	2.52%
Dif	552	84.95%	8.61%	4.17%	5.62%
Macro mean		80.21%	11.97%	11.08%	7.66%
Weighted mean		80.31%	12.01%	10.96%	7.67%

b) HVPS random HVPS CNN

	image count	accuracy	precision	recall	F1
Co	126	77.61%	15.53%	12.70%	13.97%
CBC	118	84.66%	5.26%	0.85%	1.46%
CP	111	63.41%	11.91%	29.73%	17.01%
RA	137	74.66%	15.87%	14.60%	15.21%
FA	117	60.91%	12.79%	33.33%	18.48%
HPC	145	79.66%	14.58%	4.83%	7.25%
SP	126	85.68%	50.00%	0.79%	1.56%
Macro mean		75.23%	17.99%	13.83%	10.71%
Weighted mean		75.58%	18.17%	13.30%	10.59%

Figure A2. CNN evaluation results from 'random consensus' of labels, for a) 5137 2D-S images classified by the global CNN and b) 880 HVPS images classified by the HVPS-specific CNN. Weighted mean accounts for sample size of each class.

Appendix B: Sample volume

Calculation of sample volume for OAP instruments, is a product of true air speed (TAS), effective array width (EAW) and depth, as defined in eq. B4 minimum of depth of field or arm separation.

$$SV = TAS \times EAW \times depth \tag{B1}$$

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

$$EAW = (N-1)w - W_p \tag{B2}$$

This effective array width accounts for 'entire-in' imaging of particles i.e. the particle was not imaged on the first or last diode. Where N is the number of photodiodes, w is the resolution of the instrument photodiode and W_p is the width of particle parallel to the photodiode array.

$$DoF = \frac{cD_o^2}{4\lambda} \tag{B3}$$

Eq. B3 is from O'Shea et al. (2021). Where D_o is particle diameter, λ is laser wavelength (assumed $0.8\mu m$ for 2D-S and HVPS), c is a dimensionless constant, usually set as 8 (Lawson et al., 2006). Particle diameter may have multiple definitions e.g. McFarquhar et al. (2017), which culminate in differences in concentrations between researchers. In this work the diameter is defined as the mean of the particle extent along / orthogonal to flight direction, not the maximum length.

$$depth = min(\text{arm separation}, DoF) \tag{B4}$$

The depth of area is important to be constrained by the maximum arm separation i.e. $63mm$ for the 2D-S and $200mm$ for the HVPS. The largest sized particles DoF exceeds instrument arms.

Appendix C: CNN statistics calculation

Equations C1 through C4 outline the CNN metrics used in evaluation. Figure C1 visualises these equations for the example CP evaluation. There are four possible labels associated with each CNN prediction. For the example in fig. C1 CP evaluation, true positive (TP) represents the proportion of CNN correctly identified CP images; true negative (TN) is the CNN correctly identifying non-CP images as not CP; false positive (FP) is the CNN incorrectly identifying non-CP images as CP, and false negative (FN) is the CNN incorrectly identifying CP images as not CP.

Metrics use the labels defined above to assess each class independently.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{C1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{C2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{C3}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{C4}$$

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

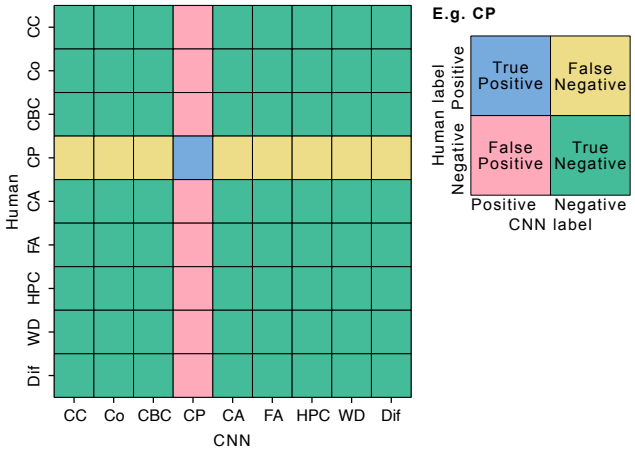

Figure C1. Example interpretation of CNN confusion matrix

Data availability. Human labelled DCMEX images are available at https://doi.org/10.5281/zenodo.20612982.

Author contributions. EEAB wrote the manuscript, with support from DLF and AMB. EEAB, DLF, AMB and PRF conceptualised this work and labelled the sample images. CRD provided insight to interpretation of CNN analysis. JC provided and pre-processed image data. This work was reviewed by all authors.

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

Acknowledgements. We would like to thank Jaffeux for the creation and open access nature of these CNN models, along with sharing code for initially setting up. We extend our gratitude to the researchers across the University of Leeds and University of Manchester who conducted the DCMEX field campaign. This work also used JASMIN, the UK collaborative data analysis facility.

Financial support. This research has been supported by the Natural Environment Research Council (NERC), grants NE/T006420/1 and NE/T006439/1. Ezri E. Alkilani-Brown was funded by Leeds-York-Hull NERC Doctoral Training Partnership (Panorama) under grant NE/S007458/1.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

References

Allabakash, S., Lim, S., Chandrasekar, V., Min, K. H., Choi, J., and Jang, B.: X-Band Dual-Polarization Radar Observations of Snow Growth Processes of a Severe Winter Storm: Case of 12 December 2013 in South Korea, Journal of Atmospheric and Oceanic Technology, 36, 1217 – 1235, https://doi.org/10.1175/JTECH-D-18-0076.1, 2019.

American Meteorological Society: Graupel. Glossary of Meteorology, https://glossarytest.ametsoc.net/wiki/Graupel, 2024a.

American Meteorological Society: Hail. Glossary of Meteorology, https://glossarytest.ametsoc.net/wiki/Hail, 2024b.

Baran, A. J.: From the single-scattering properties of ice crystals to climate prediction: A way forward, Atmospheric Research, 112, 45–69, https://doi.org/https://doi.org/10.1016/j.atmosres.2012.04.010, 2012.

Blyth, A. M., Bennett, L. J., and Collier, C. G.: High-resolution observations of precipitation from cumulonimbus clouds, Meteorological Applications, 22, 75–89, https://doi.org/https://doi.org/10.1002/met.1492, 2015.

Braham Jr, R. R.: Some measurements of snow pellet bulk-densities, Journal of Applied Meteorology and Climatology, 2, 498–500, iSBN: 0021-8952, 1963.

Browning, K. A., Ludlam, F. H., and Macklin, W. C.: The density and structure of hailstones, Quarterly Journal of the Royal Meteorological Society, 89, 75–84, https://doi.org/https://doi.org/10.1002/qj.49708937905, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49708937905, 1963.

Bryan, G. H., Wyngaard, J. C., and Fritsch, J. M.: Resolution Requirements for the Simulation of Deep Moist Convection, Monthly Weather Review, 131, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2, 2003.

Colle, B. A., Stark, D., and Yuter, S. E.: Surface Microphysical Observations within East Coast Winter Storms on Long Island, New York, Monthly Weather Review, 142, 3126 – 3146, https://doi.org/10.1175/MWR-D-14-00035.1, 2014.

Duffourg, F., Nuissier, O., Ducrocq, V., Flamant, C., Chazette, P., Delanoë, J., Doerenbecher, A., Fourrié, N., Di Girolamo, P., Lac, C., Legain, D., Martinet, M., Saïd, F., and Bock, O.: Offshore deep convection initiation and maintenance during the HyMeX IOP 16a heavy precipitation event, Quarterly Journal of the Royal Meteorological Society, 142, 259–274, https://doi.org/https://doi.org/10.1002/qj.2725, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2725, 2016.

Duroure, C.: Une nouvelle méthode de traitement des images d'hydrométéores données par les sondes bidimensionnelles, Journal de recherches atmosphériques, 1982.

Ehrlich, A., Wendisch, M., Bierwirth, E., Herber, A., and Schwarzenböck, A.: Ice crystal shape effects on solar radiative properties of Arctic mixed-phase clouds—Dependence on microphysical properties, Atmospheric Research, 88, 266–276, https://doi.org/https://doi.org/10.1016/j.atmosres.2007.11.018, 2008.

Field, P. R., Heymsfield, A. J., and Bansemer, A.: Shattering and Particle Interarrival Times Measured by Optical Array Probes in Ice Clouds, Journal of Atmospheric and Oceanic Technology, 23, 1357–1371, https://doi.org/10.1175/JTECH1922.1, 2006.

Field, P. R., Lawson, R. P., Brown, P. R. A., Lloyd, G., Westbrook, C., Moisseev, D., Miltenberger, A., Nenes, A., Blyth, A., Choularton, T., Connolly, P., Buehl, J., Crosier, J., Cui, Z., Dearden, C., DeMott, P., Flossmann, A., Heymsfield, A., Huang, Y., Kalesse, H., Kanji, Z. A., Korolev, A., Kirchgaessner, A., Lasher-Trapp, S., Leisner, T., McFarquhar, G., Phillips, V., Stith, J., and Sullivan, S.: Secondary Ice Production: Current State of the Science and Recommendations for the Future, Meteorological Monographs, 58, 7.1 – 7.20, https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0014.1, 2017.

Finney, D. L., Blyth, A. M., Gallagher, M., Wu, H., Nott, G. J., Biggerstaff, M. I., Sonnenfeld, R. G., Daily, M., Walker, D., Dufton, D., Bower, K., Böing, S., Choularton, T., Crosier, J., Groves, J., Field, P. R., Coe, H., Murray, B. J., Lloyd, G., Marsden, N. A., Flynn, M., Hu,

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

K., Thamban, N. M., Williams, P. I., Connolly, P. J., McQuaid, J. B., Robinson, J., Cui, Z., Burton, R. R., Carrie, G., Moore, R., Abel, S. J., Tiddeman, D., and Aulich, G.: Deep Convective Microphysics Experiment (DCMEX) coordinated aircraft and ground observations: microphysics, aerosol, and dynamics during cumulonimbus development, Earth Syst. Sci. Data, 16, 2141–2163, https://doi.org/10.5194/essd-16-2141-2024, 2024.

Garbrick, D., Chandrasekar, V., and Xiao, R.: Neural network based classification procedure for 2D-PMS ice crystal images, in: Conf. on Cloud Physics, pp. 59–64, 1995.

Gasparini, B., Rasch, P. J., Hartmann, D. L., Wall, C. J., and Dütsch, M.: A Lagrangian Perspective on Tropical Anvil Cloud Lifecycle in Present and Future Climate, Journal of Geophysical Research: Atmospheres, 126, e2020JD033 487, https://doi.org/https://doi.org/10.1029/2020JD033487, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020JD033487, 2021.

Gilmore, M. S., Straka, J. M., and Rasmussen, E. N.: Precipitation Uncertainty Due to Variations in Precipitation Particle Parameters within a Simple Microphysics Scheme, Monthly Weather Review, 132, 2610–2627, https://doi.org/10.1175/MWR2810.1, 2004.

Grabowski, W. W., Morrison, H., Shima, S.-I., Abade, G. C., Dziekan, P., and Pawlowska, H.: Modeling of Cloud Microphysics: Can We Do Better?, Bulletin of the American Meteorological Society, 100, 655 – 672, https://doi.org/10.1175/BAMS-D-18-0005.1, 2019.

Grazioli, J., Tuia, D., Monhart, S., Schneebeli, M., Raupach, T., and Berne, A.: Hydrometeor classification from two-dimensional video disdrometer data, Atmospheric Measurement Techniques, 7, 2869–2882, https://doi.org/10.5194/amt-7-2869-2014, 2014.

Hartmann, D. L., Gasparini, B., Berry, S. E., and Blossey, P. N.: The Life Cycle and Net Radiative Effect of Tropical Anvil Clouds, Journal of Advances in Modeling Earth Systems, 10, 3012–3029, https://doi.org/10.1029/2018MS001484, publisher: John Wiley & Sons, Ltd, 2018.

Heymsfield, A., Szakáll, M., Jost, A., Giammanco, I., and Wright, R.: A Comprehensive Observational Study of Graupel and Hail Terminal Velocity, Mass Flux, and Kinetic Energy, Journal of the Atmospheric Sciences, 75, 3861 – 3885, https://doi.org/10.1175/JAS-D-18-0035.1, 2018.

Heymsfield, A. J.: The Characteristics of Graupel Particles in Northeastern Colorado Cumulus Congestus Clouds, Journal of Atmospheric Sciences, 35, 284 – 295, https://doi.org/10.1175/1520-0469(1978)035<0284:TCOGPI>2.0.CO;2, 1978.

Heymsfield, A. J. and Kajikawa, M.: An Improved Approach to Calculating Terminal Velocities of Plate-like Crystals and Graupel, Journal of Atmospheric Sciences, 44, 1088 – 1099, https://doi.org/10.1175/1520-0469(1987)044<1088:AIATCT>2.0.CO;2, 1987.

Hicks, A. and Notaroš, B. M.: Method for Classification of Snowflakes Based on Images by a Multi-Angle Snowflake Camera Using Convolutional Neural Networks, Journal of Atmospheric and Oceanic Technology, 36, 2267–2282, https://doi.org/https://doi.org/10.1175/JTECH-D-19-0055.1, 2019.

Huang, H., Tao, R., Zhao, K., Wen, L., and Chu, Z.: Potential of Snowfall Nowcasting Using Polarimetric Radar Data and Its Link to Ice Microphysics: Study of Two Snowstorms in East China, Journal of Geophysical Research: Atmospheres, 128, e2022JD037 654, https://doi.org/https://doi.org/10.1029/2022JD037654, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022JD037654, 2023.

Hunter, H. E., Dyer, R. M., and Glass, M.: A Two-Dimensional Hydrometeor Machine Classifier Derived from Observed Data, Journal of Atmospheric and Oceanic Technology, 1, 28–36, https://doi.org/https://doi.org/10.1175/1520-0426(1984)001<0028:ATDHMC>2.0.CO;2, 1984.

Jaffeux, L., Schwarzenböck, A., Coutris, P., and Duroure, C.: Ice crystal images from optical array probes: classification with convolutional neural networks, Atmospheric Measurement Techniques, 15, 5141–5157, https://doi.org/10.5194/amt-15-5141-2022, 2022.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Jaffeux, L., Breiner, J., Coutris, P., and Schwarzenböck, A.: Convolutional neural networks for specific and merged data sets of optical array probe images: compatibility of retrieved morphology-dependent size distributions, Atmos. Meas. Tech., 18, 2311–2331, https://doi.org/10.5194/amt-18-2311-2025, publisher: Copernicus Publications, 2025.

Jensen, A. A., Harrington, J. Y., and Morrison, H.: Impacts of Ice Particle Shape and Density Evolution on the Distribution of Orographic Precipitation, Journal of the Atmospheric Sciences, 75, 3095 – 3114, https://doi.org/10.1175/JAS-D-17-0400.1, 2018.

Knight, N. C. and Heymsfield, A. J.: Measurement and Interpretation of Hailstone Density and Terminal Velocity, Journal of Atmospheric Sciences, 40, 1510 – 1516, https://doi.org/10.1175/1520-0469(1983)040<1510:MAIOHD>2.0.CO;2, 1983.

Korolev, A. and Isaac, G. A.: Shattering during Sampling by OAPs and HVPS. Part I: Snow Particles, Journal of Atmospheric and Oceanic Technology, 22, 528–542, https://doi.org/10.1175/JTECH1720.1, place: Boston MA, USA, 2005.

Korolev, A. V., Emery, E. F., Strapp, J. W., Cober, S. G., Isaac, G. A., Wasey, M., and Marcotte, D.: Small Ice Particles in Tropospheric Clouds: Fact or Artifact?, Bulletin of the American Meteorological Society, 92, 967–973, http://www.jstor.org/stable/26218567, 2011.

Lawson, R. P., O'Connor, D., Zmarzly, P., Weaver, K., Baker, B., Mo, Q., and Jonsson, H.: The 2D-S (Stereo) Probe: Design and Preliminary Tests of a New Airborne, High-Speed, High-Resolution Particle Imaging Probe, Journal of Atmospheric and Oceanic Technology, 23, 1462 – 1477, https://doi.org/10.1175/JTECH1927.1, 2006.

Lindqvist, H., Muinonen, K., Nousiainen, T., Um, J., McFarquhar, G. M., Haapanala, P., Makkonen, R., and Hakkarainen, H.: Ice-cloud particle habit classification using principal components, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/https://doi.org/10.1029/2012JD017573, 2012.

List, R.: Kennzeichen atmosphärischer Eispartikeln, Zeitschrift für angewandte Mathematik und Physik ZAMP, 9, 217–234, https://doi.org/10.1007/BF02033027, 1958.

Liu, C., Yang, P., Minnis, P., Loeb, N., Kato, S., Heymsfield, A., and Schmitt, C.: A two-habit model for the microphysical and optical properties of ice clouds, Atmospheric Chemistry and Physics, 14, 13 719–13 737, https://doi.org/10.5194/acp-14-13719-2014, 2014.

Liu, Y., Yau, M.-K., Shima, S.-i., Lu, C., and Chen, S.: Parameterization and Explicit Modeling of Cloud Microphysics: Approaches, Challenges, and Future Directions, Advances in Atmospheric Sciences, 40, 747–790, https://doi.org/10.1007/s00376-022-2077-3, 2023.

Locatelli, J. D. and Hobbs, P. V.: Fall speeds and masses of solid precipitation particles, Journal of Geophysical Research (1896-1977), 79, 2185–2197, https://doi.org/10.1029/JC079i015p02185, 1974.

Magono, C. and Lee, C. W.: Meteorological classification of natural snow crystals, Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics, 2, 321–335, 1966.

McFarquhar, G. M., Baumgardner, D., Bansemer, A., Abel, S. J., Crosier, J., French, J., Rosenberg, P., Korolev, A., Schwarzoenboeck, A., Leroy, D., Um, J., Wu, W., Heymsfield, A. J., Twohy, C., Detwiler, A., Field, P., Neumann, A., Cotton, R., Axisa, D., and Dong, J.: Processing of Ice Cloud In Situ Data Collected by Bulk Water, Scattering, and Imaging Probes: Fundamentals, Uncertainties, and Efforts toward Consistency, Meteorological Monographs, 58, 11.1–11.33, https://doi.org/https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0007.1, 2017.

Mosimann, L., Weingartner, E., and Waldvogel, A.: An analysis of accreted drop sizes and mass on rimed snow crystals, Journal of Atmospheric Sciences, 51, 1548–1558, 1994.

Moss, S. J. and Johnson, D. W.: Aircraft measurements to validate and improve numerical model parametrisations of ice to water ratios in clouds, Atmospheric Research, 34, 1–25, https://doi.org/https://doi.org/10.1016/0169-8095(94)90078-7, 1994.

Nakaya, U.: Snow crystals: natural and artificial, Harvard University Press, ISBN 0-674-18275-8, 1954.

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Oberthaler, A. J. and Markowski, P. M.: A Numerical Simulation Study of the Effects of Anvil Shading on Quasi-Linear Convective Systems, Journal of the Atmospheric Sciences, 70, 767 – 793, https://doi.org/10.1175/JAS-D-12-0123.1, 2013.

Ong, C. R., Koike, M., Hashino, T., and Miura, H.: Responses of Simulated Arctic Mixed-Phase Clouds to Parameterized Ice Particle Shape, Journal of the Atmospheric Sciences, 81, 125 – 152, https://doi.org/10.1175/JAS-D-23-0015.1, 2024.

O'Shea, S., Crosier, J., Dorsey, J., Gallagher, L., Schledewitz, W., Bower, K., Schlenczek, O., Borrmann, S., Cotton, R., Westbrook, C., and Ulanowski, Z.: Characterising optical array particle imaging probes: implications for small-ice-crystal observations, Atmos. Meas. Tech., 14, 1917–1939, https://doi.org/10.5194/amt-14-1917-2021, aMT, 2021.

O'Shea, S. J., Choularton, T. W., Lloyd, G., Crosier, J., Bower, K. N., Gallagher, M., Abel, S. J., Cotton, R. J., Brown, P. R. A., Fugal, J. P., Schlenczek, O., Borrmann, S., and Pickering, J. C.: Airborne observations of the microphysical structure of two contrasting cirrus clouds, Journal of Geophysical Research: Atmospheres, 121, 13,510–13,536, https://doi.org/https://doi.org/10.1002/2016JD025278, 2016.

Oue, M., Galletti, M., Verlinde, J., Ryzhkov, A., and Lu, Y.: Use of X-Band Differential Reflectivity Measurements to Study Shallow Arctic Mixed-Phase Clouds, Journal of Applied Meteorology and Climatology, 55, 403 – 424, https://doi.org/10.1175/JAMC-D-15-0168.1, 2016.

Praz, C., Roulet, Y. A., and Berne, A.: Solid hydrometeor classification and riming degree estimation from pictures collected with a Multi-Angle Snowflake Camera, Atmos. Meas. Tech., 10, 1335–1357, https://doi.org/10.5194/amt-10-1335-2017, aMT, 2017.

Praz, C., Ding, S., McFarquhar, G., and Berne, A.: A Versatile Method for Ice Particle Habit Classification Using Airborne Imaging Probe Data, Journal of Geophysical Research: Atmospheres, 123, 13,472–13,495, https://doi.org/https://doi.org/10.1029/2018JD029163, 2018.

Prodi, F.: Measurements of Local Density in Artificial and Natural Hailstones, Journal of Applied Meteorology and Climatology, 9, 903 – 910, https://doi.org/10.1175/1520-0450(1970)009<0903:MOLDIA>2.0.CO;2, 1970.

Rahman, M. M., Quincy, E. A., Jacquot, R. G., and Magee, M. J.: Feature Extraction and Selection for Pattern Recognition of Two-Dimensional Hydrometeor Images, Journal of Applied Meteorology and Climatology, 20, 521–535, https://doi.org/https://doi.org/10.1175/1520-0450(1981)020<0521:FEASFP>2.0.CO;2, 1981.

Raymond, D. J. and Blyth, A. M.: Precipitation development in a New Mexico thunderstorm, Quarterly Journal of the Royal Meteorological Society, 115, 1397–1423, https://doi.org/https://doi.org/10.1002/qj.49711549011, 1989.

Reinking, R. F.: Formation of Graupel, Journal of Applied Meteorology (1962-1982), 14, 745–754, 1975.

Ren, T., Li, D., Muller, J., and Yang, P.: Sensitivity of Radiative Flux Simulations to Ice Cloud Parameterization over the Equatorial Western Pacific Ocean Region, Journal of the Atmospheric Sciences, 78, 2549 – 2571, https://doi.org/10.1175/JAS-D-21-0017.1, 2021.

Sokol, A. B., Wall, C. J., and Hartmann, D. L.: Greater climate sensitivity implied by anvil cloud thinning, Nature Geoscience, 17, 398–403, https://doi.org/10.1038/s41561-024-01420-6, 2024.

Takami, K., Kamamoto, R., Suzuki, K., Yamaguchi, K., and Nakakita, E.: Relationship between newly fallen snow density and degree of riming estimated by particles' fall speed in Niigata Prefecture, Japan, HYDROLOGICAL RESEARCH LETTERS, 16, 87–92, https://doi.org/10.3178/hrl.16.87, 2022.

Touloupas, G., Lauber, A., Henneberger, J., Beck, A., and Lucchi, A.: A convolutional neural network for classifying cloud particles recorded by imaging probes, Atmospheric Measurement Techniques, 13, 2219–2239, https://doi.org/10.5194/amt-13-2219-2020, 2020.

Vaillant de Guélis, T., Schwarzenböck, A., Shcherbakov, V., Gourbeyre, C., Laurent, B., Dupuy, R., Coutris, P., and Duroure, C.: Study of the diffraction pattern of cloud particles and the respective responses of optical array probes, Atmospheric Measurement Techniques, 12, 2513–2529, https://doi.org/10.5194/amt-12-2513-2019, 2019.

Varble, A. C., Nesbitt, S. W., Salio, P., Hardin, J. C., Bharadwaj, N., Borque, P., DeMott, P. J., Feng, Z., Hill, T. C. J., Marquis, J. N., Matthews, A., Mei, F., Öktem, R., Castro, V., Goldberger, L., Hunzinger, A., Barry, K. R., Kreidenweis, S. M., McFarquhar, G. M., McMurdie, L. A.,

https://doi.org/10.5194/egusphere-2026-3385
Preprint. Discussion started: 9 July 2026

Pekour, M., Powers, H., Romps, D. M., Saulo, C., Schmid, B., Tomlinson, J. M., Heever, S. C. v. d., Zelenyuk, A., Zhang, Z., and Zipser, E. J.: Utilizing a Storm-Generating Hotspot to Study Convective Cloud Transitions: The CACTI Experiment, Bulletin of the American Meteorological Society, 102, E1597 – E1620, https://doi.org/10.1175/BAMS-D-20-0030.1, 2021.

Wolf, K., Bellouin, N., and Boucher, O.: Sensitivity of cirrus and contrail radiative effect on cloud microphysical and environmental parameters, Atmospheric Chemistry and Physics, 23, 14 003–14 037, https://doi.org/10.5194/acp-23-14003-2023, 2023.

Wu, Z., Liu, S., Zhao, D., Yang, L., Xu, Z., Yang, Z., Zhou, W., He, H., Huang, M., Liu, D., Li, R., and Ding, D.: Neural Network Classification of Ice-Crystal Images Observed by an Airborne Cloud Imaging Probe, Atmosphere-Ocean, 58, 303–315, https://doi.org/10.1080/07055900.2020.1843393, 2020.

Wu, Z., Liu, S., Zhao, D., Yang, L., Xu, Z., Yang, Z., Liu, D., Liu, T., Ding, Y., Zhou, W., He, H., Huang, M., Li, R., and Ding, D.: Optimized Intelligent Algorithm for Classifying Cloud Particles Recorded by a Cloud Particle Imager, Journal of Atmospheric and Oceanic Technology, 38, 1377–1393, https://doi.org/https://doi.org/10.1175/JTECH-D-21-0004.1, 2021.

Xiao, H., Zhang, F., He, Q., Liu, P., Yan, F., Miao, L., and Yang, Z.: Classification of Ice Crystal Habits Observed From Airborne Cloud Particle Imager by Deep Transfer Learning, Earth and Space Science, 6, 1877–1886, https://doi.org/https://doi.org/10.1029/2019EA000636, 2019.

Yang, P., Liou, K.-N., Bi, L., Liu, C., Yi, B., and Baum, B. A.: On the radiative properties of ice clouds: Light scattering, remote sensing, and radiation parameterization, Advances in Atmospheric Sciences, 32, 32–63, https://doi.org/10.1007/s00376-014-0011-z, 2015.

Yi, B.: Diverse cloud radiative effects and global surface temperature simulations induced by different ice cloud optical property parameterizations, Scientific Reports, 12, 10 539, https://doi.org/10.1038/s41598-022-14608-w, 2022.

Zhang, H., Li, X., Ramelli, F., David, R. O., Pasquier, J., and Henneberger, J.: IceDetectNet: A rotated object detection algorithm for classifying components of aggregated ice crystals with a multi-label classification scheme, EGUsphere, 2024, 1–27, 2024.