Response to Reviewer 2 comments

We are thankful for the work done reviewing our manuscript and appreciate the feedback and comments given by Reviewer 2. We are considering these while revising our manuscript.

Below we report our responses to each comment. We first repeat the reviewer's comment in normal font. Then, directly following the comment, we report our response in cursive font. Each response starts with 'Response:'. Considering all comments, we are revising the manuscript aiming for an improved data description, a clearer model description, and easier to understand messages.

This study proposed a machine learning (ML) algorithm for ice crystal image segmentation and classification. It is claimed to be a semi-supervised algorithm by the authors. In-situ imaging techniques can measure microphysical properties of individual cloud particles, which are valuable to study the climate and local influence of cloud. The traditional image analysis relys on manual processing which are time-consuming and labor-intensive. While over past decades, various machine learning (ML) methods were being applied to detect and classify of cloud particles collected by different devices. However, there is not a universal model that can be applied across different instruments. The nature of the data made the authors propose a 2-step model for processing and classifying them. A segmentation model that is responsible for filtering ice particles from the background, and a classification model that is used to classify filtered ice particles to 4 prescribed categories. An achievement of the ML algorithm is that it was successfully applied on the data collected by B-ICI for the first time. Based on the evaluation on 200 particles. two major results were concluded in this manuscript.

1. The filtered ice particles by the segmentation model showed high correlation with the hand-labled ice particles in terms of the maximum dimension and particle area. The maximum size distribution of model output and ground truth showed good agreement for particle larger than 100.

2. The agreement of columns, compacts and rosettes with ground truth was around 80%.

Comments and questions

In general, while the research topic and proposed ML model are of significant interest, the major concerns regarding the presentation of this manuscript, the novelty of its contributions and the sufficiency of analysis and interpretations currently prevent me from recommending publication in its present form. I believe the manuscript has potential, but it would require substantial revisions to address several issues which will be detailed discussed in the following review script.

Incomplete literature review and vague statement of motivation/research questions

In the introduction section, the review of machine learning (ML) methods applied to ice crystal images is not complete and up-to-date. Lindqvist et al. (2012) and Praz et al. (2018) used Principal Component Analysis (PCA) and logistic regression for classifying ice crystal images before Xiao et al. (2019). More recently, the field has seen significant advancements. Schmitt et al. (2024) proposed using a Visual Geometry Group (VGG) model for ice crystal classification, while Zhang et al. (2024) developed a specialized rotated object detection algorithm that can identify ice crystal aggregates down to the monomer scale. Additionally, Chu et al. (2025) introduced a semi-supervised algorithm to efficiently classify a large number of unlabeled images. This additional revision provides a more complete and chronological overview of the key studies, showing how the field has evolved from traditional methods to more advanced deep learning techniques.The current statement about a "universal model" (lines 37-40) is a bit confusing and doesn't clearly align with the rest of the manuscript's apparent purpose. To improve this, the authors need to re-consider their research questions or motivation in the introduction.

Response: We have extended the literature overwiev with newer papers published in the field, as suggested. In addition, while revising the Introduction section, we have clarified, that our paper is aiming for a model capable of processing data from B-ICI, for which the oil-coated film capturing ice particles is posing special challenges that a classification model alone cannot address. Therefore, the images have to be segmented first to identify ice crystals. Mentioning a "universal model" in a sentence (Lines 37–38, first version of manuscript) in the Introduction was not referring to work carried out and described in our manuscript. Instead, it should have been a motivation for developing a specific model for B-ICI.

Insufficient and imprecise description of data and CNN model

In Section 2.2, the description of image data is too brief. Although the authors cited the data analysis paper (Wolf et al. (2018, 2019)), in order to make it more "reader-friendly", the authors should provide a concise summary of the dataset directly in this section. Additionally, the authors should include a description of all datasets used in your study, including the one from the 2024 campaign, which was used for evaluation in Section 4. The summary of the dataset should include at least the name of the campaign and the number of particles. It would be better if the author can also describe the environmental conditions like in-cloud temperature and supersaturation briefly.

Response: To provide a better context and improve clarity in our manuscript, we are adding a table summarizing all data by providing campaign date, number of images, number of particles, and cloud altitude and mean temperature. For new campaign data that have not been included in Wolf et al. 2018 and 2019, more information would not benefit this manuscript, and further analysis of the measured ice clouds together with their environmental conditions will be described in future publications.

The model description (Section 3) lacks references. Most of the descriptions of CNNs are based on past work in the field of computer vision, while not much relevant work is cited here. In addition, the expressions are not accurate. For example, the expression: "The CNN processes this matrix by evaluating each pixel along with its neighboring pixels, assigning values to local pixel groups," is vague. "Assigning values" does not clearly describe how convolution between a kernel and the corresponding receptive field is calculated. I suggest the authors improve the description of the CNN to make it more precise. Line 148, please cite the paper for those models: ResNet, VGGNet, and AlexNet.

Response: With Section 3.1 we wanted to provide an overview and information applicable to both the segmentation and the classification model. Together with some vagueness in the descriptions, this has, however, reduced clarity in our model descriptions. Consequently, while revising the manuscript, we have ommited Sect. 3.1., Overview of ML model training*, and merged its relevant content into the following sections that describe our models. We have changed descriptions to be more concise and precise and have added references.*

Insufficient description of segmentation model training/evaluation

The current description of the segmentation model is too brief. The authors need to provide more details about its architecture, justify the choice of EfficientNet-B7, and clarify the training and evaluation process. Most segmentation models follow an encoder-decoder structure. EfficientNet is just a backbone which can be applied to different tasks. It would be better if you revealed more informa-

tion about the structure of the segmentation model and justified your choice of EfficientNet-B7 instead of simply mentioning "through repetitive training."

Response: In the revised manuscript, we are now better describing the structure of the segmentation model. Before selecting EfficientNet-B7, we had evaluated 29 different models, With "through repetitive training" we referred to the repetitive evaluations in running all these models, each time training the same dataset with varying batch-size and learning rate. We judged the performance using precision and recall values from the training for each evaluated model version and found that EfficientNet-B7 was marginally better than the other models. We have revised the manuscript such that the choice of EfficientNet-B7 is more clearly justified and that the process leading to that choice is described too. All tested model versions have been listed as well.

The description of the beta factor in Section 3.2 is indeed vague and incomplete. The author needs to explain why it's necessary for their dataset and provide more details about how it was used in their model training. For example, since the beta factor is used for balancing the FN and FP values, the reason why the authors want to balance them in their datasets should be specified here.

Response: The goal was not to find an optimal beta value for the segmentation model, but to have three models with different sensitivities, from which the user can select one as needed when processing new data. B-ICI image quality can vary depending on several conditions, the cloud, the specific illumination settings and set up, and the film and its oil coating. Therefore, an optimized and fixed value of β will not work for all B-ICI data. Instead, our approach was to have different models corresponding to low, mid, and high β values so that one can select the model that is best at identifying ice particle and rejecting non-particle features on the specific image data under processing. Thus, the chosen best beta value will make the following visual inspection step the easiest.

I am confused in lines 178-179 about why the performance on fresh data is inspected visually instead of evaluated according to some mathematical metrics. Why is human vision here more trustworthy than statistics?

Response: Due to the nature of our image data, the image processing still cannot be fully automated. For example, the holes along the edge of the film tape (visible in Figure 3) appear periodically and are often accompanied by darkened corners or marks. Despite this regularity, the segmentation model struggles to reliably distinguish these features from actual ice particles. Similarily, the model is not 100% reliable in distinguishing other non-particle features from ice particles. As a result, an intermediate visual inspection step—between segmentation and classification—remains necessary to manually correct the model-generated mask images. While the performance has improved with the progressive inclusion

of new data into the training dataset and subsequently retraining the segmentation model, this enhancement is still insufficient to allow the model to run entirely without manual inspection of the model output to guarantee good quality of the data before it is further analysed. Statistical comparisons could be applied on a dataset where confidence in the segmentation is already very high. This is not the case for B-ICI data yet. Based on our experience, even though initial training may yield a certain precision, the model still performs with a lower precision when applied to new datasets. Applying the segmentation model to a new dataset followed by this manual inspection is much quicker than entirely relying on manually generating mask images.

As far as I understand, the choice of the hyperparameter beta in this study is based on human vision by iteratively inspecting the trained segmentation model output of the validation set. Then the question is why not using a grid search method on the hyperparameter beta based on mathematical metrics?

Response: The hyperparameter beta was used to have different sensitivities of the segmentation model to simplify the visual inspection step when processing new data (not the validation set) as explained above. The values of beta used were not based on iterativelly running and inspecting the segmentation model output of the validation set. Instead, the beta parameters used for the strict, medium, and loose models were selected as 0.7, 0.05, and 0.025, respectively, as these values produced noticeably different sensitivities.

Until Section 3.2, the process of training the segmentation model seems like a fully-supervised method to me. The training of the segmentation model starts from inputting an annotated dataset according to Figure 6. The further training of the model with unseen data is still based on human supervision. Therefore, the segmentation model of this manuscript cannot be defined as a 'semi-supervised' model. Typo: line 160 "amountof"

Response: As explained in our responses to Reviewer 1, the term 'semi-supervised' was used incorrectly. To correct this, we have revised the title and removed all references to semi-supervised learning from the manuscript. Thank you, we have corrected the typo.

Insufficient description of classification model training/evaluation

Line 204, it would be better if the authors could provide the number of particles used in training.

Response: A table has been created that shows the number of particles in each category used in the training.

Line 213, typo "number"

Response: Thank you, typo fixed.

Lines 214-219, the details of each component in the confusion matrix are not described.

In Line 219, the decimals in Figure 9 are defined as 'confidence', which is vague to the audience what they represent.

Normally, the decimals on the diagonal are the precision of each category. Plus, you mention the lowest one is 85%, which I did not find in Figure 9. In general, the process of training the classification model is also fully supervised.

Response: In the revised manuscript, we have improved the description of the confusion matrix. Indeed 'confidence' is not well defined here, thus we have avoided this term. The reported values on the diagonal represent precision for the respective category. The lowest precision is 88%, writing '85%' in the text was a mistake.

Insufficient model performance evaluation

The major problem here is that the segmentation model and classification model were evaluated only on a 200-particle dataset, which is too small to achieve a statistically significant conclusion, although the way of evaluation (size distribution comparison and confusion matrix) is good.

Response: We have extended our dataset for performance evaluation. For this, we have used measurements that took place on 2021-03-17, not used previously for training. From that day, we have added 205 particles that have been segmented and classified both manually as well as by using the ML models. Now altogether there are 385 particles used in the performance evaluation.

Lack of novelty

There have been many classification models developed for ice crystal images. Therefore, a simple 4-category classification model in this manuscript is not a novel contribution. However, the segmentation model developed for an oil-coated ice crystal image system could be a potential novel contribution to the community.

Response: Indeed, there are many other models developed for ice particle images. As we have pointed out in responses above, each instrument needs its own model. This is what literature so far seems to confirm. In the revised Introduction section we have tried to make this clearer, and, as mentioned earlier, included more references to these previous works. Early on in this project, we

tested to process our own data with some of these previous models only to find that the results were not satisfactory. Therefore, we have developed a new classification model for B-ICI. The choice of a model with four categories was in line with previous preliminary analysis (Wolf et al., 2018) and with our needs for continuing this analysis with new data acquired since then. We agree that the main novelty of this paper lies in the segmentation model required to preselect particles in B-ICI data before the classification can be performed. Thus, this paper presents the method and tailored models to process and analyse B-ICI data more efficiently.

The segmentation model has the potential to filter out ice particles from films among other artifacts. As mentioned before, the authors still need to add more analysis about the evaluation of the segmentation model. Another problem is that the authors claimed the method is semi-supervised while the flowchart indicates that each node of training still involves human supervision.

Response: As described above, we have added more data to improve the performance evaluation. Our method is not semi-supervised, it was our mistake to use this term, which we have corrected now, as stated above. We are sorry for the confusion this has created.

Code and data availablity

I strongly suggest to make code and data available to public through associated GitHub/Zenodo repository even at the preprint stage.

Response: All codes are available on github: `https://github.com/Omnitok/amt-2025-2818.git` *The training dataset is too large for github and will be published on Researchdata.se.*

Quality of figures

All figures except Figure 6 suggest adjusting the font size of the text to make it clearer for the audience.

Response: Thank you, font sizes have been increased on figures. The size of the figures have also been adjusted for improved clarity.