Response to Reviewer 1 comments

We thank Reviewer 1 for the valuable comments. We appreciate the time and effort the reviewer has put into providing feedback to help us improve the 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 work describes a machine learning (ML) algorithm that has been developed to process images captured by a balloon borne ice crystal imager (B-ICI) and classify ice crystal habit.

Overall, this manuscript does not look like it was ready for submission. While the work itself may very well be worthy of publication, the authors need to put considerably more thought into the manuscript's presentation of their work and what they want the research community to know about it. What follows below are comments to try to help the authors understand what content and clarification I feel is needed for the manuscript to be appropriate for publication, but these comments are not exhaustive.

The authors should generally give more consideration to how a reader in the community would interpret their writing.

The concepts in the manuscript, particularly those that set this work apart from existing work, are not well flushed out. The manuscript needs to clearly define the context under which the work contributes something novel. This is an area where the manuscript is significantly lacking. It should then perform some sort of investigation or analysis to establish the performance of the novel aspects of the technique. This could be a validation effort (showing the algorithm does what the authors claim it does), baselining performance against a standard method or even establishing clarity about where pitfalls exist and further work is needed.

Response: The novelty of this work, as we see it, lies in developing a machine learning model specifically tailored to our instrument, B-ICI. While classification models described in literature have been created for various other instruments using different techniques, our main challenge has been segmenting and identifying particles in our image data. As B-ICI captures particles on a plastic film *strip, the images are taken against the film as a background. On the one hand this ensures that particles are in focus, on the other hand it causes image noise (varying background) and non-particle features visible on the images. Thus, the segmentation model represents the core novelty. The classification model, acting on the output of the segmentation model, has been added to support future analyses planned with B-ICI datasets. To the best of our knowledge, there is no standard benchmark against which we can directly compare our results. Therefore we are presenting training metrics and comparing the model's output against results from manual analysis.*

It is already noted that there have been many segmentation models developed and applied to ice crystal images, so this is not, in itself, a novel contribution. Based on the content in the manuscript, I think there are potentially two novel aspects of the work that could be flushed out. In either case the authors need to do more work defining those research elements and showing progress toward them, including addressing potential criticisms.

The authors develop a "self-supervised" machine learning approach which might be novel. This is used to augment the training dataset in an effort to make the machine learning solution more generalizable to data from other projects and even potentially instruments. I think there are some concerns with this approach, which the authors would need to address through testing and evaluation in the manuscript. My principal concern is with how this influences the sample set used for training the processor. In particular, I worry it will emphasize those particles where the CNN is already performing well and it will ignore those where it does not perform well. Wouldn't that just reinforce its existing weaknesses? In addition, my understanding from the manuscript is that this technique did not appear to work for extending the processing algorithm to new instruments (line 225).

Response: Upon further reflection, we have recognized that the term 'semi-supervised' has been used incorrectly in the title of our manuscript, in the model description, and the conclusions. Instead, we have developed two separate fully supervised machine learning models, which are applied sequentially. A visual inspection step between the two to ensures that the segmented objects are indeed ice particles. The term 'semi-supervised' has been removed from the title and the manuscript.

Another possibly novel aspect would be in regard to the authors noting the lack of machine learning algorithms that can generalize across instruments. Typically the ML architecture needs to be retrained (or even re-hyperparameter optimized) between different instruments. This work could be used to motivate the development of more generalized solutions by highlighting this challenge and showing how performance degrades in time for the same instrument or across different instruments. The work does not necessarily have to solve this problem, but rather demonstrate to the community that the problem exists, is difficult and is worth solving.

Response: We tried to highlight in the Introduction of our original manuscript the lack of models that can be used across different datasets. We did that to motivate the need for a tailored model for our B-ICI datasets. Trying to apply freely available models to our dataset only confirmed this need. In the revised manuscript, we have made it clearer that we want to describe our model tailored to the special needs of B-ICI, which is not contributing to a more general model.

A significant issue is in the presentation of the CNN model itself where the focus tends to be poorly directed. In some cases, the manuscript focuses on unnecessary, pedantic and not entirely accurate details (like explaining convolutional layers – line 99) while poorly communicating the overall processing work flow and CNN architecture. The manuscript reads like there might just be a single scalar output (like using a series of CNN layers which feed into dense layers to output a scalar – based on the description at line 112) but the data (and use of the term "segmentation model" and pre-trained models) seems to suggest an all-convolutional NN such as a UNET architecture where the output is a 2D mask with the same dimensions as the input. I really can't tell which it is. In another place, the role of beta in the Tversky loss function is discussed. This represents a hyperparameter in the model training process and typically would be optimized. How is this done? What value do the authors use? It's mentioned that the model is trained with 3 values for beta but how are the results of those three outputs used and evaluated? Is there some conclusion that follows from this?

Response: Section 3 has been revised to address this issue and focus on necessary details in the model description. Specifically, Section 3.1 has been merged into Section 3.2, where the model parameters are now introduced within the context of the training process. It is important to note that β is not a parameter we aim to optimize. Instead, three separate models—corresponding to low, mid, and high β values—were deliberately created. The intention is to provide flexibility. B-ICI images can change depending on several conditions, the cloud, the illumination, and film preparation. Therefore, an optimized fixed value of β will not work for all B-ICI data. Having different models corresponding to low, mid, and high β values allows one to select the model that best captures particle edges and consequently makes the visual inspection step the easiest.

There needs to be a clearer description of the data that is used for inputs and labels. What are these exactly? I'm particularly confused by the label data. Are these actual segmentation masks with dimensions the same as the input images? Is the input to the classifier the original image or the output from the particle detection step?

Response: This point has also been clarified by rewriting Section 3 to more clearly distinguish between the source data used for the segmentation and classification models, respectively. A table has been introduced to provide additional *details on the labeled data used for the initial training of the models. For the segmentation the labeled data means manually generated mask images with the same dimensions as the corresponding raw images. This is used for training of the model. Running the trained model generates similar masks with the same image dimensions as the original images as output. For the classification, label means the class category. The classification model uses cropped images of ice crystals identified by the segmentation as input.*

The paragraph in the introduction (line 37) seems to imply that the authors are addressing the problem of a generalized ML solution for processing ice particle images across instruments. My guess, as a reader, is that the "semi-supervised" approach is the method by which the authors intend to address this. But ultimately, the authors note that the method developed here does not address this problem (line 225). If I'm not interpreting this content correctly, it suggests that the authors have not been explicit enough in their description of the scope and motivation for their effort.

Response: We are sorry that our mentioning that no existing model performs reliably across different instruments confused the Reviewers. This statement reflects the reason why we needed an instrument-specific solution. The statement can, however, also be interpreted as motivation for a generic model. To avoid this ambiguity, the introduction has been rewritten to clearly state that the model presented in this work is specifically designed for the B-ICI instrument.

"Semi-supervised" is the first word in the title but there is no mention of it in the abstract or conclusions. If this represents an important feature of this work, results connected to the technique (e.g. validation or performance comparison) need to be more thoroughly explained and demonstrated. If this is not a key element of the work presented in the paper, then it should be de-emphasized and not be called out in the title. The amount of space dedicated to describing and validating the performance of the "semi-supervised" approach is also very limited. How do you know that this approach is actually improving the model? As stated earlier, I would be concerned that this method actually reinforces the limitations of the model. Can you perform an analysis that shows this method allows the gradual improvement of the model?

Response: As we have pointed out earlier, we have recognized that the term 'semi-supervised' has been used incorrectly. We have referred to this term in the title of our manuscript, in the model description, and the conclusions. If we had used the term correctly, we should of course have had more explanation in the manuscript after referring to it in the title, as Reviewer 1 rightfully has pointed out. Semi-supervised learning refers to a training approach that uses a small amount of labeled data combined with a large amount of unlabeled data. In contrast, the work presented here involves two independently trained models using fully supervised learning, with a visual inspection step between them. Consequently, in revising the manuscript, we have changed the title and removed all *references to semi-supervised learning.*