Dear Dr Bert Wouters,

Again, we would like to thank the three referees for taking further time to review our manuscript and provide valuable feedback which has helped to improve the paper. We are pleased to have this opportunity to respond, and we can confirm that all the issues they raise have been addressed.

In this response, we provide a detailed point-by-point response to each of the three Reviewers, with their comments (*verbatim*) in blue and our response in black with reference to specific lines/sections of the revised manuscript. As a result of their suggestions, we have undertaken further corrections, reduced the length of the paper and elaborated on potential applications of the method for the wider glaciological community. We shortened the manuscript by 3,271 words overall and length was reduced most by condensing the introduction for improve readability, focusing less on different parameter combinations (i.e., tile size, patch size, image bands), and by merging/moving figures to the supplement.

We thank you for your editorial work on our manuscript and look forward to hearing from you in due course.

Melanie Marochov

(on behalf of all authors)

Reviewer 1

Comments:

Most of the questions have been addressed carefully, and I appreciate the well-documented answers prepared by the author. Given this, I recommend this paper for publication. We thank the reviewer for their feedback and are delighted they would like to see our manuscript published.

However, I still have one concern regarding the response to my comment 9 about how the Phase 2 model can overcome the inaccurate boundaries in the Phase 1 model. I believe from Phase 1 to Phase 2, there is a huge improvement (Figure 13, 15, 16, and 17), and it deserves more detailed explanations. That is why I was looking for a theoretical explanation in my previous comment. However, the author only claims it is due to the robustness of the Phase 2 model. Again, it would be beneficial for me and other readers to know the mechanism behind the robustness.

My initial guess is that the robustness of the Phase 2 network is owing to the early stopping. The author mentioned that the training data are not all correct. Maybe the early stopping could prevent the Phase 2 network from being overfitted to the incorrect training information. In the first revision of the manuscript, we stated that the phase two models are robust to noise, and we try to further clarify the theory behind this in lines 115-121 of the revised manuscript. In general, it has been found that deep learning models can generalise to data well even if some training labels are incorrect due to the overall training process. This is because deep learning models are designed to minimise error and not overfit to training data (i.e., they are able to learn the overall trend in training data even if some labels are wrong). Thus, the phase two models can overcome some level of error in phase one predictions because *most* of the phase one predicted training labels are correct; they do not memorise the labels which are wrong but generalise to the overall trend. We have tried to clarify this in the revised manuscript.

Also, I am curious about how much percent of the training data is incorrect. I suppose that the F1 scores for Phase 1 CNN tile classification are tile-based ones. If I am correct, what the pixel-based F1 score would be after converting the tile labels to a full-size class raster (e.g., Figure 16c & Figure 17c)? F1 scores for both phase one and phase two are pixel-based and estimated from a sample of 10 million pixels which we state on line 371. Thus, phase one F1 scores give a representation of how correct phase two training labels are.

Reviewer 2

Marochov et al. made a huge effort to revise their manuscript and to address the reviewer's comments in every detail. The manuscript improved particularly with regard to more accurate explanations on technical details, additional test data and the implementation of an approach to extract the calving front. Additionally, the edge classification problem was addressed accurately and is clear now. Nevertheless, a few major concerns remain which are outlined below. We thank the reviewer for their constructive feedback which has helped to improve the revised manuscript.

1. The authors added two additional test sites and more test scenes covering a wider temporal variety. This really highlights the transferability of the developed approach. Could you please explain why you decided to remove the initial test site Scoresby Sund mentioned in the first version of the manuscript? As you already have the data it would be worth to include it as an additional test set. We decided to remove the Scoresby Sund site because although we had the original test data, the data for the revised test sites had to be completely re-digitised since reviewers rightly suggested that too much blank space existed in the original validation labels. As such, including the Scoresby site would also have prompted a complete re-digitisation of

the validation data for an increased number of acquisitions to provide a seasonal test of transferability as was done for the Helheim site and new out-of-sample sites. While it would have been ideal to include more sites, we were limited by the time it takes to manually delineate validation labels within the constraints of review deadlines (especially since the original Scoresby site spanned an entire Sentinel-2 tile).

2. The revised manuscript includes a lot of additional information on the developed approach and describes technical aspects in every detail. On the one hand, this is an advantage of the manuscript as the approach is very transparent. On the other hand, the manuscript has become rather long and focuses more on technical details. The authors have to be careful to not only provide methodical details on their approach but also to fit within the scope of The Cryosphere by addressing a broader cryospheric community. To make your manuscript more suitable for a wider cryospheric community I would recommend the following:

> *a. Highlight the advantages of your approach for the cryospheric community. So far, the discussion is solely technical. But you could add one section discussing the future advantages of your approach for the analysis of the cryosphere (e.g. calving front detection, change detection of class distributions, snow cover changes between different years etc.)* The first revision of the manuscript had a paragraph explaining the further potential applications of CSC within the cryosphere which we have tried to strengthen in the second revision (Section **4.2 CSC performance and wider application**). We have also added a small case study (Figure 10) as requested by the editor, which demonstrates that CSC outputs can be used to create time series data, in this case including calving front change and mélange area variation throughout the test imagery collected for the Helheim site during 2019.
>
> *b. Highlight the great performance of your classification algorithm but try not to confuse the reader with too many details about performance differences. For example, consider to shift some of the plots to the supplementary material. You could just show the data for the optimal model configuration in the manuscript and keep the rest in the supplementary materials.* We have followed your suggestions by condensing the amount of technical detail in relation to parameter testing and moving it to a small section titled **2.6 Optimal performance parameters**. Similarly, some plots have been removed/moved to the supplement and figures of optimal CSC outputs are shown in the results alongside the new case study.
>
> *c. Consider to shorten the text and densify the information on tile sizes, patches and different model configurations. For readers with no background in machine learning all those parameters might be confusing. Probably, a table including all those different*

parameters and the corresponding accuracies could help for a condensed and better overview. We have followed your suggestions as noted above and created the table you suggest (Table 1). Information on tile sizes, patch sizes and image bands are no longer in the results section but have been condensed into section 2.6.

d. The authors put a lot of effort into comparing different model configurations (tile and patch size) which is highlighted in several figures and graphs. In my opinion, it would be worth to merge some of the figures to shorten the manuscript. Please see the suggestions in the technical corrections below. Again, we have reduced the number of figures and merged some as suggested. The paper is now shorter, and we have tried to shift the focus from different model configurations to a more balanced overview of both the advantages of the approach in terms of glaciological applications and the technical considerations for future research in this field.

Technical corrections:

Figure 1: This is a really nice figure and helps to understand your classification approach much better. Well done! Thank you!

Figure 2: Nice idea to combine the class examples within this figure. Looks much nicer now. We appreciate the positive feedback.

Figure 6 & 7: Consider to merge those two figures into Figure 6a & 6b. It will be easier to see the differences between the cCNN and MPL approach. Thank you for this suggestion, the figures have now been merged.

Figures 11, 12, 13, 15, 16 & 17 demonstrate the classification results. The amount of figures might be a bit too heavy compared to the length of the paper. You could consider to merge the results into less figures or shift some results to the supplementary materials. As noted above, the number of figures has been reduced and some have been shifted to the supplement. In the revised version of the manuscript Figures 8 and 9 now show optimal classification/calving front detection examples of CSC applied to seasonal imagery and Figure 10 represents a small case study which highlights the usefulness of our multi-class approach.

Figure 14: You could condense the information and use one plot showing all three glaciers in the same plot with the best model configuration (RGB+NIR, Single). The remaining part could be moved to the supplementary. We have removed this plot completely as the information is now shown in Tables 1 and 2.

Figure 15d: It is interesting, that the model confuses mélange with glacier ice so heavily. Could you explain why this is the case? We expect this is to do with the spectral similarity between

the classes as well as textural properties. In the Helheim training area, most of the glacier ice is highly crevassed, whereas the section that was misclassified as mélange in the Store test image (originally Figure15d, now Figure S3d) appears to be less crevassed. I.e., since the CNN had not seen many samples of less crevassed glacier ice, it was less likely to predict correctly in this case. As a result, including additional training data from the Store site improved the model's ability to predict the class correctly (Joint training). We have also suggested in the revised manuscript that including more diverse training data from more than one glacier in future work would likely aid classification.

L371: "trains to learn". Please re-phrase. This has been corrected.

L486: What does "bergy" mean? Or just a typo. This was a typo and has been corrected.

L591: Why does the joint model provide higher classification accuracies but the single model higher accuracies for the calving front extraction? This seems to be contradictory. F1 scores show accuracy for the entire test image so while small classification differences at the calving front may not significantly affect F1s for the image classification as a whole, differences of just a few pixels (e.g., in areas of shadow) can impact calving front detection more acutely. Thus, small changes in the way the Single and Joint models predict pixels at the calving front can lead to some variation in error between the two approaches even though the additional training data for the Joint approach improved overall classification performance.

L644: Here you mention that the developed approach might be suitable for lake mapping but earlier it is mentioned that the approach has difficulties with classes being smaller than the tile size. Are those lakes always large enough to be captured by your model? Additionally, the class "lake" is not included in the classification or is it defined as open water enclosed by glacier ice? We suggest that smaller scale features such as lakes could be identified by using CSC to isolate target classes (line 633). For example, if we are purely interested in supraglacial lakes, the glacier ice class could be used as a search area input for an additional model which is designed to detect lakes (this could be a simple MLP), rather than searching a whole image which may produce noisier outputs.

Supplement:

*1. Why did you use additional Helheim scenes for the joint training method even though the single model was trained on Helheim anyways? (see scenes with * in Table S1)* It allows a comparison and appreciation for the addition of fine-tuning for out-of-sample sites. We did not expect Joint results for the Helheim site to show significant improvements over Single training for the exact reason you suggest, but thought it was valuable to test this anyway.

2. Why do you provide only a confusion matrix for the single training but not the combined training approach? It would be interesting to see the performance differences to justify the necessity of a joint and single model approach. Moreover, I would assume that the most robust (spatially transferrable) model would be achieved by including several training areas from the beginning (instead of only Helheim) over different glaciers which would make the additional joint training unnecessary. We have now included confusion matrices for both Single and Joint outputs using optimal parameters to allow comparison. We agree that the addition of more training data from more glaciers may improve the performance of the phase one CNN on out-of-sample images which we state on line 643. Nonetheless, we believe that the ability of CSC to accurately classify out-of-sample images when trained only on data from Helheim (F1s over 90%) and produce calving front errors comparable to manual delineation/previous deep learning methods shows the benefits of the workflow.

Please don't be discouraged by the length of my review. I know that a lot of work went into the revised version and the provided comments might require some further effort. Nevertheless, my comments are mostly suggestions and not a must hopefully helping to improve your manuscript. Thank you for your suggestions, they have been valuable, and we think they have helped improve the manuscript.

Reviewer 3

The authors have made significant improvements to the manuscript, and have resolved many of the previously addressed issues. I believe the manuscript provides a valuable addition to the literature by providing a novel approach for delineating calving fronts only using tile labels and including multiple surface classes. We thank the reviewer for their helpful comments which have helped to improve the clarity of the paper.

My remaining comments are with regards to improved clarity of the paper:

Section 2.1: it should be emphasized that the output of Phase 1 is not used as the input of Phase 2. Rather, the output of phase is the training LABEL of the next phase. Both phases use the original image as the input. This can also be clarified in Figure 1 by adding an arrow from the original image to the input of phase 2. This has been clarified in section 2.1 and Figure 1 has also been adapted as suggested.

Lastly, the authors previously noted in the response letter that quantifying the average uncertainty of manual delineations is beyond the scope of the work, while noting that there is "Clearly, there is a 3-5 pixel error in manual digitising." That is exactly my point. Including an approximate (but justified!) uncertainty range for the corresponding manual digitising provides

much better context for the baseline "expectation" of the uncertainty of results with the given images and sites. We have included an estimate of manual digitisation error on line 407 in the revised manuscript and agree that it provides a good baseline expectation for what automated techniques can achieve.