Dear Referee #1,

We thank you for your very useful and constructive review. We will follow your recommendations in the revised version and give detailed answers to your comments below. Some of your points were also raised by the other Referees and we thus refer occasionally also to those answers.

Again, many thanks for helping to improve our manuscript!

Best wishes,

Erik Loebel and all co-authors

General comments

The authors present an exciting deep learning method for tracing glacier calving fronts in Landsat 8 and 9 images. The manuscript presents the method based on a specialized Artificial Neural Network, the resulting dataset of 9243 calving front traces for 23 of Greenland's outlet glaciers, and an example of how the data may be useful for examining glacier dynamics. The method produces calving fronts that are on average within 80 meters of manually traced calving fronts, which is less than the uncertainty of manual calving front delineations according to a study by Goliber et al. (2022).

The authors thoughtfully developed the deep learning method. They considered different illumination conditions and terminus morphologies when training the model. There is good documentation of the time and storage requirements for training the model. I applaud the contribution to open-source code and datasets, which are valuable to the glaciological community. The 698 manual delineations used for training the model would also be valuable to the community and I recommend submitting them to a new or existing data repository.

Thank you for your positive and constructive feedback. We fully agree with the recommendation to publish our reference data set. We assume that this data set has considerable value for glaciological analyses and will also prove highly beneficial in subsequent machine learning applications.

In particular this reference data set includes 898 (we used 698 for model training and 200 for model testing) manually delineated calving front positions, which we will provide in a georeferenced shapefile format, as well as 1026 machine learning ready input raster subsets (pre-processed, 9 channels) with their corresponding, manual delineated, segmentation mask. Raster subsets are available in both png and georeferenced tif format. The data is already submitted to the TU Dresden *Open Access Repository and Archive* (OpARA). Reference (with doi) will be included in the revised version.

Overall, the manuscript is well-presented and concisely written. The figures are particularly well-constructed and compelling. However, I think the main text currently lacks detail on the deep learning method. I think the information included in Appendix A and B should be included in the main manuscript since it is relevant to understanding how the ANN algorithm was developed.

This is in line with a comment from Referee #3. We welcome this recommendation and will incorporate appendix A and B into the main body of the manuscript, specifically in Section 2.

Spatial transferability of the method is mentioned throughout the manuscript and described as an advantage to using this method compared to other existing automated calving front tracing methods. The deep learning model is tested on glaciers from regions outside of Greenland (e.g., Antarctica, Svalbard, and Patagonia). I would be really interested in formal discussion of how the method, trained on Greenland's outlet glaciers, performed with the glaciers in other regions specifically. How does the accuracy calculated for those test glaciers compare to the accuracy of the Greenland test glaciers? Discussing this would provide appropriate support for the spatial transferability of the method.

This comment is much appreciated. When developing our method, we have put a lot of emphasis on model generalization and spatial transferability. In fact, we also process and publish time series of glaciers outside the training dataset.

Although some results are already presented in the figures (e.g. Figure 3 (b) and Figure 7 (j)), we fully agree that a proper discussion of this topic would strengthen the manuscript.

Results for this discussion are already at hand. We calculated the model accuracy separately for (1) glaciers outside the training data set, (2) glaciers outside Greenland and (3) glaciers within the training data set. The results will be presented and comparatively discussed in a new section. In addition, we have created a figure showing example validation results specifically for glaciers from Antarctica, Svalbard and Patagonia (similar in style to Figure 3).

In general, this manuscript presents a valuable contribution to the field and I would like to see this work published after these more major comments and the minor comments listed below are addressed.

Specific Comments

In general, proof read for compound adjectives that need to be hyphenated, e.g., Greenland-wide (L176).

Many thanks for the comment. We will double check the text and the compound adjectives it contains.

L10: You should include a statement about the accuracy of your method that you calculated here.

The calculated accuracy estimates will be included in the abstract.

L14-15: The phrase "digital twin" of Greenland ice sheet is not clearly defined. Unnecessary in abstract unless explained in more detail. It's not discussed throughout the paper so I don't think it's appropriate to include here or in the conclusion without further elaboration.

We agree, the phrase "digital twin" will be removed from the abstract.

L52-55: This is not the first automated method that captured sub-seasonal resolution time series of calving front change (see Liu et al., 2021). Reframe the language here.

This point has also been raised by the other referees and we agree. The wording (as well as at L166-168 and L173) will be revised.

L69: What is the fixed window size and how was it chosen?

The window size is 512 px by 512 px. Together with the 30 m resolution of Landsat-8, these dimensions were determined to be optimal for capturing the glaciers in Greenland without resampling the imagery. Only Humboldt Glacier, Nioghalvfjerdsbrae and Zachariae Isstrøm do not fit into this window size. This information (some is already in Appendix B) will be included in the main manuscript.

L72: “Built” instead of “build”

Will be fixed, thank you.

L86-100: This section discussing the method performance should be moved to the Results or Discussion section.

We thank you for this advice. However, this section does not concern the validation of our actual results, but rather the assessment of our machine learning method. The validation of machine learning methods is so closely intertwined with the method itself that we believe it is better to describe these two parts together. Validating our model on independent test data is not a result of our study, but a prerequisite to continuing with our processing and starting with the inference. To make this clear, we propose to rename Section 2.2 from "Validation" into "Accuracy Assessment". For the results and discussion sections, we want to focus on the application to Greenland, the resulting data product, and the glaciological implications. We note that the reviewers had no criticism about this structure. If you strongly prefer a reorganization, please let us know.

L106: Elaborate on how the completely clouded Landsat scenes are filtered.

We recognize that the current manuscript lacks detailed information on the filtering of clouded Satellite scenes and failed extractions in our processing workflow. Your comment is therefore very appreciated and also in line with Referee #3. We will rework and expand Section 3 to include information of the following processing steps:

1. Landsat scenes are downloaded using the USGS EarthExplorer (earthexplorer.usgs.gov). Scenes with cloud cover larger than 20% and all *Systematic Terrain Correction* (L1GT) scenes are checked manually. If the glacier front is not visible, the satellite scene is not downloaded for further processing.
2. After the ANN processing, failed calving front extractions are discarded. Calving front extraction fails when the longest feature (which is derived by applying the GDAL contour algorithm with threshold 0.5 on the segmented image) does not intersect the static mask which results in no shapefile being produced.
3. Finally calving fronts are filtered after time series generation using the rectilinear box method. Here we separate all entries with an area difference of larger than 1 km^2 to both the previous and the next entry. Separated entries are checked manually.

The manual cloud cover check in step 1 was done to reduce download traffic and time. Depending on the glacier 51% (for Ingia Isbræ) to 63% (for Helheim Glacier) of the available satellite scenes are discarded before download. Step 2 and 3 reduced the data product from 10587 to 9243 entries, i.e. discarded about 13% of data.

If data download is no issue (for example when having a local archive), step 1 could be skipped. This will result in significantly more discarded glacier fronts in step 2 and 3.

L136: Include citations for how glacier geometry impacts terminus retreat. At the very least, Felikson et al., 2020 (https://doi.org/10.1029/2020GL090112) should be cited here since it directly discusses the impact of bed topography on glacier retreat.

This is a major shortcoming of the current version and the issue has also been raised by the other referees. The discussion will be revised significantly. We will include a literature review on geometric controls on calving front change and link our results to those of existing studies. Please have a look at our *Reply of RC2* (page 4).

L140: Looks more like 2016 and 2017, not 2018 showed the rapid retreat for Ingia Isbræ.

Thank you very much for pointing this out, this will be corrected.

L164-166: Is it that the algorithm performs better at overcoming challenging cloud, illumination, and mélange issues than manual delineations? The way this sentence is currently structured implies that. I think this sentence could be removed altogether since the sentence that follows already emphasizes the high temporal resolution of the time series.

We see the problem and appreciate the comment. The sentence will be reworked entirely so it does not imply that the method outperforms manual delineation.

Figures and Tables

Fig. 2. In the caption, write out TU Dresden or just refer to it as the testing dataset for this study. I think it's fine to exclude the testing glaciers from other regions. Adding a location in parentheses after each of the excluded glaciers would make it more clear why they aren't included in this map. E.g., Drygalski Glacier (Antarctica), Storbreen Glacier (Svalbard), etc.

Many thanks, this is indeed a very good suggestion which we will implement.

Fig. 5. I recommend adding a colorbar for the green shading.

We will add a colorbar for Figure 5.

Fig. B1. This figure could remain in the Appendix or Supplementary Material even if the description of methodology in Appendix B is moved to main text.

We will move this Figure B1 together with the reworked Table C1 (see comment below) and time series of all other glaciers (not shown in Figure 7) into the supplementary material. Thank you for this suggestion.

Table C1. Is the right side really a confusion matrix if only done for TUD? Listing the fraction/percentage of total pixels would be more meaningful here than the raw pixel numbers. As of now, I draw much more from the mean and median errors listed on the left side than the Confusion Matrix. Consider separating the Confusion Matrix portion of this table into its own table. Clearly define TP, TN, FP, FN in the caption.

The confusion matrix is only calculated for the TUD validation set. We have presented it to enable the calculation of commonly used binary classification metrics (like accuracy, F1-Score, recall, precision). Giving the values as percentage is a very good suggestion which we will implement.

We also agree that the distance errors are more meaningful. We will move the table with the distances (which will be expanded to include the mean and median Hausdorf distance, see comment from referee #3) to the main manuscript. The table with the binary classification matrix and the reworked confusion matrix will be moved to the supplement.