Dear Bert,

Thank you for your work on our manuscript and for inviting us to submit a revised version. We have uploaded this revised manuscript together with a track change version. We have carefully addressed the points raised by the referees and made changes in line with our responses. Below, we also provide definitive answers to how and where the referee's comments were addressed. For comprehensive answers to the questions, please see our responses (Reply on RC1/2/3).

Furthermore, I would like to inform you that I will be on a field campaign from 2 February to 19 April. I apologize in advance for not being able to reply to emails during this period. Our second contact (Martin Horwath) is informed about this, but he will not be able to make any major changes to the manuscript.

Best wishes,

Erik and all co-authors

Anonymous Referee #1, 01 Jun 2023

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.

All reference data applied in this study is available at http://dx.doi.org/10.25532/OPARA-282. In particular, this includes 898 manually delineated calving front positions provided in a georeferenced shapefile format, as well as 1220 machine learning ready raster subsets (pre-processed, 9 channels) with their corresponding manual delineated segmentation mask.

This is now also referenced in the *Code and data availability* section.

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.

We have incorporated 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.

Our results on spatial transferability are presented in a new section (section 3.2).

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).

Done.

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

Done.

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.

The phrase "digital twin" is 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.

The phrasing at this and other locations has been revised.

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

This information is now included in the main manuscript (Section 2.3.1).

L72: "Built" instead of "build"

Done.

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

The accuracy assessment is now its own section (Section 3).

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

Section 4.1 now includes this information.

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.

The discussion (Section 5) has undergone significant reworking. It now includes a comprehensive literature review and a more thorough contextualisation of our results.

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

Fixed.

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.

The sentence has been reworked 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.

Done.

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

Done.

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.

Done.

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 table with the binary classification metrics and the reworked confusion matrix has been moved to the supplement.

Anonymous Referee #2, 05 Jun 2023

General Comments

This paper uses a deep-learning-based method to produce 9243 calving front positions across 23 Greenland outlet glaciers from 2013 to 2021 and discusses the relationship between terminus variation and basal topography. Overall, I think this paper is well-written and the figures are well-presented. With that being said, I have reservations about its originality, impact, and the extent of its literature review. Based on these concerns, I would recommend rejecting the paper in its current form. Below, I provide a detailed evaluation and rationale for my recommendation.

1. Originality:

In recent years, there increasing number of deep learning-based studies to automate terminus extraction. Compared with the previous study, especially with the author's previous paper Loebel et al. (2022), what is the improvement of this study regarding the methodology?

See *Reply of RC2*.

The improvements to *Loebel et al. (2022)* have been emphasized more thoroughly throughout Section 2.

1. Impact:

The main objective of automating the terminus extraction is to produce as many termini as possible. However, compared with the CALFIN (Cheng et al, 2021), which produces 22 678 calving front lines across 66 Greenlandic glaciers from 1972 to 2019, this study seems not improve the temporal resolution, temporal coverage, and spatial coverage. A comparison between the product from this study and CALFIN would be helpful. For instance, which glaciers CALFIN did not cover but this study covers.

The results of this comparison, not only to CALFIN but also to AutoTerm and Termpicks, are presented in the new Section 4.3.

1. Literature Review:

The discussion about the terminus variation and basal topography is interesting, but studies have been investigating this for many years (Joughin et al., 2008, 2014; Kehrl et al., 2017; Bunce et al., 2018, Catania et al., 2018), suggesting that retrograde bed slopes can cause glacier dynamic instabilities (Meier and Post, 1987) and substantial retreat. However, these papers are missing in the manuscript. Therefore, I suggest the author include a literature review about how the glacier geometry influences the terminus variation, also remove the expression throughout this manuscript that this study is the first application analyzing the interaction between calving front variation and bedrock topography.

The discussion (Section 5) has undergone significant reworking. It now includes a comprehensive literature review and a more thorough contextualisation of our results.

Considering the points mentioned above, I believe that rejecting the paper would be appropriate. However, I recommend that the authors be given an opportunity to revise and resubmit their work, addressing the concerns mentioned above. By providing constructive feedback and clear expectations for revision, the authors might have the chance to strengthen their manuscript and overcome the present limitations.

Specific Comments:

Line 51: Why the usage of multispectral sensor information can increase the temporal solution, compared with using the single band? I was asking because using multi-band information could not yield more images.

See *Reply of RC2*.

Section 2.1 and section 4.3 now also address this point.

Line 53: I believe the CALFIN can also resolve sub-seasonal terminus variations as it uses all the available Landsat images.

The phrasing at this and other locations has been revised.

Line 64: What modification did the author apply here? Does that modification improve the results? It would be better to have a more detailed description of this.

The descriptions in Section 2.3 have been expanded to include this information.

Section 2.2: It would be better if the author can change the name of this section (validation) to avoid confusion. Following the deep learning convention, there are three sets: the training set, the validation set, and the test set. This section is actually about the test set but is called validation.

Also, the descriptions of the validation set should be included in the manuscript.

Done.

Figure 8: It would be more helpful to change the second column to the time series of the bed elevation at the terminus. Time series might better reflect the verbal descriptions in the discussion section.

This figure (now Figure 10) has been reworked. It is now much easier to identify the bedrock slope at the calving front for a given date

Reference

Loebel, E., Scheinert, M., Horwath, M., Heidler, K., Christmann, J., Phan, L. D., ... & Zhu, X. X. (2022). Extracting Glacier Calving Fronts by Deep Learning: The Benefit of Multispectral, Topographic, and Textural Input Features. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-12.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I., & Rignot, E. (2021). Calving Front Machine (CALFIN): glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019. The Cryosphere, 15(3), 1663-1675.

Joughin, I., Howat, I., Alley, R. B., Ekstrom, G., Fahnestock, M., Moon, T., Nettles, M., Truffer, M., and Tsai, V. C.: Ice front variation and tidewater behavior on Helheim and Kangerdlugssuaq Glaciers, Greenland, J. Geophys. Res.-Earth, 113, F01004, https://doi.org/10.1029/2007JF000837, 2008.

Kehrl, L. M., Joughin, I., Shean, D. E., Floricioiu, D., and Krieger, L.: Seasonal and interannual variabilities in terminus position, glacier velocity, and surface elevation at Helheim and Kangerlussuaq Glaciers from 2008 to 2016, J. Geophys. Res.-Earth, 122, 1635–1652, https://doi.org/10.1002/2016JF004133, 2017.

Catania, G. A., Stearns, L. A., Sutherland, D. A., Fried, M. J., Bartholomaus, T. C., Morlighem, M., Shroyer, E., and Nash, J.: Geometric Controls on Tidewater Glacier Retreat in Central Western Greenland, J. Geophys. Res.-Earth, 123, 2024–2038, https://doi.org/10.1029/2017JF004499, 2018.

Bunce, C., Carr, J. R., Nienow, P. W., Ross, N., and Killick, R.: Ice front change of marine-terminating outlet glaciers in northwest and southeast Greenland during the 21st century, J. Glaciol., 64, 523–535, https://doi.org/10.1017/jog.2018.44, 2018.

Meier, M. F., & Post, A. (1987). Fast tidewater glaciers. *Journal of Geophysical Research: Solid Earth*, *92*(B9), 9051-9058.

Anonymous Referee #3, 18 Oct 2023

General comments

The authors present a deep learning method for automating the digitization of glacier calving fronts from Landsat 8 and 9 images, applied to 23 glaciers in Greenland from 2013 through 2021. They demonstrate the potential application of this method in studying aspects of outlet glacier behavior such as calving front seasonality and interactions with bed topography.

Overall the manuscript is clearly written, and the figures are high quality and do a good job of supporting the text. However, I am concerned that it is lacking important context from previous work, and is not sufficiently distinguishable from other automation methods:

First, how does this deep learning method compare to previous contributions to automation, both in methodology and in outcomes (accuracy)? In other words, as a data user, why should I choose the output from this particular automated method over any of the others that have been recently published? The authors highlight the temporal resolution of this dataset, but to me it doesn't appear to be substantially unique from other automated or even some manual datasets.

The results of this comparison, not only to CALFIN but also to AutoTerm and Termpicks, are presented in the new Section 4.3.

Next, for the highlighted seasonal and topographic applications, how do the results presented here agree with or differ from other studies? There is a wealth of literature on both terminus position seasonality and terminus-topography relationships, but none of it is referenced in the presentation of these data applications, and the authors suggest that their analysis is unprecedented, which is simply not true. Both section 3.2 and the discussion would benefit from a more thorough literature review and contextualization of the work presented.

The discussion (Section 5) has undergone significant reworking. It now includes a comprehensive literature review and a more thorough contextualisation of our results.

Finally, can the authors demonstrate that this automated method, and its resultant higher temporal density of calving fronts, improves upon prior analyses? It would be helpful to move beyond demonstrating the capabilities to demonstrating how they are better than existing approaches. Figure 6 starts to get at this with the comparison between TUD and ESA-CCI, but I think this could be explored in more detail.

This is demonstrated in the comparison presented in the new Section 4.3.

Specific comments

L31-35: The authors cite several studies that use manually delineated fronts, then state that these products "often lack temporal resolution, making seasonal analysis … difficult." However, several of the cited products do perform robust seasonal analyses. Schild and Hamilton (2013) had near-daily termini for five glaciers. King et al. (2020) had seasonal data, though centerline positions rather than full calving fronts. Goliber et al. (2022) specifically addressed how their combined dataset could be used to better study seasonality than any of the individual component datasets. Black and Joughin (2023) was explicitly about analyzing sub-seasonal (weekly or monthly) front variability throughout Greenland – in fact, for their weekly dataset, they reported an average of 50 fronts per glacier per year (including polar night), versus 45 in this manuscript (9243 fronts / 23 glaciers / 9 years). While it is true that manual digitization is laborious and will struggle to keep up with the volume of new imagery (L33), that cannot be used to imply that those efforts have not produced comparable datasets. The authors go on to use language such as "unprecedented temporal resolution, resolving [the glaciers'] sub-seasonal calving front variability for the first time" (L52-53), "[this time series] clearly surpasses the potential of manually delineated data products" (L166-168), and "unprecedented temporal resolution" (L173). Papers cited within this manuscript, as well as others, have repeatedly demonstrated comparable temporal resolution of calving front data products, including through manual digitization.

The formulations at these and other locations have been reworked. The phrase "*unprecedented temporal resolution*" has been removed from the manuscript entirely.

L73-75: I'm glad these various messy situations were considered for training data. Did the training data also focus on glaciers that are heavily crevassed near the terminus? This can also be a tricky condition – and different from dense mélange – perhaps wrapped up in "morphological features"?

See *Reply of RC2*.

These crevassed conditions are now mentioned in the revised text (section 2.2).

L78-80: Did this system end up being spatially transferable? The results for looking at the non-Greenlandic glaciers are not addressed.

Our results on spatial transferability are presented in a new section (section 3.2).

L98-101: I agree with the overall conclusion that the quality of automated fronts from this method is comparable to the quality of manually delineated termini. The comparison with the error estimate from Goliber et al. (2022) is a helpful reference, but a direct comparison is difficult due to differences in the error calculation method. Goliber et al. (2022) use the Hausdorff distance (they say "greatest minimum distance between two lines") rather than averaging the minimal distances along the front as is done here, so I suspect their error estimate would be a bit larger than what is reported here. It would also be helpful to see how error estimates from this method compare to other automated methods (it looks like this is at least done for CALFIN in Table C1, but this should be mentioned in the main body text about error estimation too).

The Hausdorff distance estimate is now included in the accuracy assessment in Section 3.1.

L105-106: What percentage of imagery was completely clouded (and therefore unused)?

Section 4.1 now includes this information.

L106-107: What qualifies as a "failed calving front extraction"? What are the criteria for failure? Did the authors manually check all results from the automation?

Section 4.1 now includes this information.

Figure 8/Discussion: This is a nice presentation of the relationship between topography and calving fronts over seasons and years. The expanded interpretation in the discussion is helpful, especially for highlighting correlations between topography and seasonal amplitude. However, these interpretations should be placed in the context of other topography-terminus studies (how are they similar/different, what is new here, etc.). Catania et al. (https://doi.org/10.1029/2017JF004499 and others) would be a good starting point.

The discussion (Section 5) has undergone significant reworking. It now includes a comprehensive literature review and a more thorough contextualisation of our results.

Appendix B: Since the core of this manuscript is the deep learning method, the methodology should be in the main body rather than tucked away in an appendix.

We have incorporated appendix A and B into the main body of the manuscript, specifically in Section 2.

Technical corrections

L72: change “build” to “built”

Done.

L140: add BedMachine version here (I see v5 in the references, but it is helpful to see up front since there have been some big changes between versions).

Done.

L172: remove “are” after Nioghalvfjerdsbræ

Fixed.