Dear Bert,

Thank you for your work on our manuscript. We have addressed the remaining points raised by the referees. Please find our revised manuscript together with a track change version. Below, we also provide answers to how and where the referee's comments were addressed.

Best wishes,

Erik and all co-authors

Anonymous Referee #2, 05 Mar 2024

Most of the comments have been carefully addressed, and I appreciate the response provided by the author. However, I would like to comment on a limitation regarding the integration of multispectral bands as follows.

The authors demonstrate that integrating multispectral bands can yield more accurate results compared to using single-band inputs, particularly in resolving challenging conditions such as ice-mélange. Nonetheless, there are several limitations to consider. Firstly, the exclusion of the panchromatic band from Landsat-8 impacts the spatial resolution of the terminus products. While I understand that this omission may be necessitated by the requirement of identical resolution for the input bands in the artificial neural network (ANN) used, it raises concerns about potentially sacrificing higher spatial resolution data. Secondly, the requirement of nine bands for input in the ANN restricts the integration of other satellite products. For instance, SAR images like Sentinel-1 typically do not offer multiple bands, and while Sentinel-2 images provide 12 bands, only four of them have the highest 10-meter resolution. Consequently, while the integration of multispectral bands from Landsat-8 provides more termini compared to CALFIN or AutoTerm, it excludes the utilization of other valuable datasets entirely, which in turn affects the temporal resolution of the terminus data. Particularly during winter when optical images lack coverage, SAR images can provide essential winter traces, which are crucial for studying seasonality.

Although I do not currently propose solutions for these limitations, I recommend that the authors at least acknowledge and discuss them in their work.

We thank the reviewer for raising this point, this is indeed worth discussing.

In principle, it is no problem to integrate input data with different resolutions into our processing. In fact, that's what we did with the 100 m resolution TIR bands. This is not a limitation to the use of multispectral information. The reason we chose not to integrate the 15 m panchromatic band is that either the input tiles would have been too large and thus the computational effort too high (if we decided for larger tiles), or we would have lost the spatial context within individual ANN predictions (if we decided for a smaller area).

The nine input bands of our model are tailored to use Landsat-8 and Landsat-9 imagery. Even the integration of Sentinel-2 imagery would require retraining of the model, as Sentinel-2 only captures imagery in VIS and SWIR wavelengths and therefore does not acquire TIR data. Therefore, we fully agree with the reviewers statement that our approach excludes the use of other satellite sensors. In contrast to the approach of *Zhang et. al 2023* (AutoTerm), where the authors developed a model capable of analyzing multi-sensor imagery, our approach uses only Landsat imagery, but exploits the full sensor information. There are advantages and limitations to both approaches.

We now emphasize these aspects and limitations in several places in our manuscript.

In Section 2.1:

[...] "The 15 m resolution band 8 is excluded due to the otherwise too high computational cost." [...].

In Section 2.3.1:

[...] "The 30 m ground sampling distance, and thus the exclusion of panchromatic band 8, is a compromise between the spatial context provided within a single subset, the computational effort and the resolution of the calving front predictions." [...].

In Section 4.3:

[...] "AutoTerm has the most mapped and unique fronts as well as the highest sampling rate. This is mainly due to its data basis which includes Landsat, Sentinel-2 and Sentinel-1. Compared to our approach, which is limited to the use of multispectral Landsat data, this is a clear advantage." [...].

In Section 4.3:

[...] "As a final point, we want to emphasise the potential of combining different glacier front products. Particularly for data sets based on optical data, this not only increases the overall sampling rate, but also allows data gaps to be filled during the polar winter." [...].

Anonymous Referee #3, 01 Mar 2024

The authors present a revised manuscript detailing their deep learning method for automated delineation of glacier calving fronts. The method is applied to a set of glaciers in Greenland, but the authors demonstrate spatial transferability to other glaciers in Greenland as well as in Svalbard, Patagonia, and Antarctica. The manuscript discussion has been greatly improved from the previous submission with respect to providing context from previous work on seasonal calving front variations and interactions with bed topography, and I thank the authors for their work to address that and other comments.

I do have some remaining general concerns about the comparisons with other calving front datasets and the evidence provided for distinguishing this method from other automated calving front delineation methods:

The authors have addressed comments concerning the novelty and capabilities of their method by adding a comparison of their model output with that of some other "big data" calving front datasets. The TermPicks dataset (Goliber et al., 2022) is used as a point of reference for comparing this method to manually delineated datasets. This is appropriate as TermPicks is a large compilation of several previously published datasets: it is comparable in size to automatically delineated datasets, it spans the space and time necessary for a comparison with this work, and it provides internal error estimates which can also be compared with this and other methods. However, the analysis presented here (pages 14-15) emphasizes differences in temporal resolution/sampling rate and in the capacity to capture seasonal changes, and the author response (RC-3; not the manuscript) specifically criticizes the uneven spatial and temporal sampling of TermPicks. While this is a shortcoming of the TermPicks dataset, this is due to it being a compilation of other datasets with disparate spatial and temporal requirements, and not necessarily a weakness of manual delineation more generally as is implied. Black and Joughin (2023), which is already cited elsewhere in the manuscript, is also a “big data” calving front dataset with circum-Greenland coverage, which are the criteria listed for the comparative analysis presented here, and that work is specifically about seasonal changes. Given the emphasis on temporal resolution and seasonal changes presented here, that dataset would be more suitable for the in-depth comparison presented in this manuscript, in addition to the other three datasets already used.

We agree and are thankful for this recommendation. The inclusion of the *Black and Joughin (2023)* data set is a very welcome addition to our comparison.

We included the data set of *Black and Joughin (2023)* into our comparison on section 4.3. Please find the additions in Table 2, Figure 9 and the whole text. We also want to note that we had to change the reference time from 2013-2019 to 2015-2019 which led to changes in the statistics for the other data sets as well.

With regards to evidence for distinguishing this method from other automated methods: the authors highlight this method’s success in “challenging conditions" (L276 and in the conclusion). This seems to be the main strength of this method over previous automated delineation methods. However, the only evidence given to support this claim is the specific case of the seasonal ice tongue on Kangiata Nunaata Sermia (L264-269). Can a more general observation of the relative success of this method in challenging conditions be proven? Evidence of this point needs to be strengthened in the results in order to justify the conclusion.

We see why these statements are problematic. Although our method extracted a significant amount of calving front that other methods did not (Section 4.3, Table 2), especially compared to the CALFIN product which is using the same (Landsat) data basis, we agree that this does not necessarily mean that our method is more capable for difficult conditions. However, as our dataset is the only one benefiting from multi-spectral inputs, the integration of which leads to more accurate calving front predictions for challenging ice mélange and illumination conditions (*Loebel et al. 2022*), we believe this is a possible explanation.

The statement in L276 is not substantiated and is changed accordingly. It now reads:

[...] “Importantly, these 13% are likely to include extractions under challenging ice mélange and illumination conditions.“ [...].

The statement in the conclusion is removed entirely.

Specific comments

L25-27: Greene (2024) is cited later but would be a good reference to include here as well, to further motivate the importance of calving front retreat for quantifying sea level rise.

Citation is now included.

L32-34: Regarding calving front products lacking temporal resolution: I commented on this paragraph in my previous review, and the authors agreed in the response that it should be changed or clarified, along with several other related comments throughout the manuscript. It seems the other areas were addressed but this paragraph was perhaps overlooked, as the tracked changes show no edits to this paragraph. To recap, I agree with the statement in the response that manual delineation has not kept up with the volume of satellite imagery, but that is not the sentiment conveyed by this paragraph. Some manual studies have achieved temporal resolution comparable to automated methods, and the comment about the difficulty of performing seasonal analyses is contradicted by multiple papers cited in this paragraph. Please see my previous comment as well and revise.

L53: “the glaciology community requirement” – be specific about what this requirement is that you are meeting.

We have specified this statement. The sentence now reads:

[...] “By achieving this robust and scalable calving front extraction, we meet the glaciology community requirement for a comprehensive parameterization of glacier calving in Greenland and make important steps towards establishing artificially intelligent processing strategies for glacier monitoring tasks.” [...].

L54 and L346: “intelligent processing strategies” – perhaps say “artificially intelligent” or something similar to highlight the autonomous aspect of this work (previous approaches are not necessarily unintelligent!).

We follow this suggestion and change the phrasing to “artificially intelligent”.

L196: When looking at area differences >1km2 between two entries, is this area change scaled by the length of the calving front at all? An area change of 1km2 at Humboldt is relatively small compared to an area change of 1km2 at Hayes, for example.

The area difference is not scaled by the length of the calving front. The intentionally low value of 1 km^2 was found to be a good compromise for calving fronts captured in a single ANN input tile. For larger glaciers that are delineated by merging several tiles, namely Humboldt Glacier, Zachariæ Isstrøm and Nioghalvfjerdsbrae, this low value resulted in an increased amount of separated extractions resulting in more time spend manually checking these separated extractions. The 1 km^2 threshold did not negatively affect the quality of the data product or the results of this study.

However, we are very thankful for this suggestion. A scaled threshold, maybe even by calving front length as well as ice velocity, is indeed a good way to further improve our processing in the future. Especially when integrating more large calving fronts, for example in Antarctica.

L281-282: I agree that combining calving front products is helpful to increase spatiotemporal coverage; Goliber et al. (2022) has been cited several times in this manuscript and should be included here, as that is a key reference for combining different calving front products for further analyses.

Citation is now included.

L288: What qualifies a glacier as having “clear seasonality”? How did you determine that 19 glaciers are clearly seasonal? In the results (L218-219) it says that only two glaciers do not have clear seasonality, which would suggest that 21 of the 23 do have clear seasonality, rather than 19.

The "clear seasonality" referred to in this statement has been determined by a visual inspection of the time series. We appreciate this question and we also think that a "clear" seasonality needs a more quantifiable definition. We changed the phrasing of the sentence (L288) to:

[...] "A visual inspection of the time series shows that 19 of the 23 glaciers analyzed in this study show a seasonal pattern between the years 2013 and 2021." [...].

Figures of all time series can be found in the supplement.

The statement in L216-219 describes the time series in Figure 8, where there are two glaciers (out of 12) without seasonal calving front variation. In our entire dataset, 19 out of 23 glaciers show seasonal calving front variation.

Technical corrections

Generally: correct the inconsistent usage of é in mélange

Done

L20: "…imbalance is driven by changes…" --> "…imbalance is driven in part by changes…" (because the ice-ocean boundary is not the only control on dynamics)

The wording in this paragraph was somewhat vague and sub-optimal, so we have changed it. It now reads:

[...] "While about half of the ice mass loss is due to increased meltwater runoff, the other half is due to changes of ice discharge to the ocean related to changes of the ice flow dynamics of outlet glaciers (Otosaka et al., 2023). Several mechanisms act as controls and indicators for dynamic glacier changes. In particular, calving and calving front variations have been identified as crucial parameters for investigating the physical mechanisms of Greenland outlet glaciers (Joughin et al., 2008a; Moon and Joughin, 2008; Benn et al., 2017; Trevers et al., 2019; Cook et al., 2021; Melton et al., 2022)." [...]

Figure 5c caption: Patagonia

Fixed

L193: "Depending on" (not "of")

Done

L338: Humboldt

Fixed

References

Black, T. E. and Joughin, I.: Weekly to monthly terminus variability of Greenland's marine-terminating outlet glaciers, The Cryosphere, 17, 1–13, https://doi.org/10.5194/tc-17-1-2023, 2023.

Goliber, S., Black, T., Catania, G., Lea, J. M., Olsen, H., Cheng, D., Bevan, S., Bjørk, A., Bunce, C., Brough, S., Carr, J. R., Cowton, T., Gardner, A., Fahrner, D., Hill, E., Joughin, I., Korsgaard, N. J., Luckman, A., Moon, T., Murray, T., Sole, A., Wood, M., and Zhang, E.: TermPicks: a century of Greenland glacier terminus data for use in scientific and machine learning applications, The Cryosphere, 16, 3215–3233, https://doi.org/10.5194/tc-16-3215-2022, 2022.

Greene, C. A., Gardner, A. S., Wood, M., and Cuzzone, J. K.: Ubiquitous acceleration in Greenland Ice Sheet calving from 1985 to 2022, Nature, 625, 523–528, https://doi.org/10.1038/s41586-023-06863-2, 2024.

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.