Dear Referee #3,

We thank you for the constructive comments and the careful assessment on our manuscript. All comments have been taken into account and a list of responses and actions is given below. Many of your points were also raised by the other referees and we thus refer 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 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.

We thank you for raising this important point. In the current version of our manuscript, this aspect is somewhat neglected. We understand that the comparison with the existing data products in particular is very important for precisely the concerns you mentioned. This was also suggested by the editor and referee #2.

The revised version of the manuscript will include a new section presenting and discussing the results of this comparison. In addition to the CALFIN dataset (*Cheng et al. 2021)*, we also analyzed the recently published AutoTerm dataset (*Zhang et al. 2023)* as well as the manually delineated TermPicks repository (*Goliber et al. 2022)*. The results of this comparison are already at hand and are briefly introduced at our response to referee #2. Please have a look at our *Reply of RC2* (page 2).

Based on the results of the comparison, we will also highlight differences in processing. A comparison of our ANN architecture with that of CALFIN and many others has already been published in *Heidler et al. (2023)*. Our focus will therefore be on the input data, pre-processing, post-processing and quality control.

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.

Thank you for this comment. We agree that the discussion (section 4) needs to be improved. Especially in the points you highlighted. The revised version of our manuscript will include a literature review and contextualisation of our results. This was also pointed out by the other referees. Please see our *Reply of RC2* (page 4).

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.

That's very well put. As with your first point, we agree. For the revised version we demonstrated that our processing system has a higher sampling rate than CALFIN and than the TermPicks repository. Within this reference we also show that 13% of our extracted calving front traces were not extracted by either CALFIN, TermPicks or the AutoTerm Product.

For a more direct comparison we also created Figure R1 (see *Reply of RC2* (page 3)). We directly compare our results with CALFIN, TermPicks and AutoTerm for four example glaciers. This figure highlights differences in sampling rate, but also the different approaches in input data and filtering.

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.

We agree that this sentence has to be rephrased completely. Manual digitization will always be capable of creating data sets with comparable or better temporal resolution compared to automated delineation, at least in theory. Our argument is (and that's why this sentence has to be changed) about the time-consuming process of manual digitization. We apologize for not making this clear.

Many manual delineated data sets are side products of glaciological studies where the authors most likely spend a lot of time on manual digitization. For example, to delineate the calving front data used in *Catania et al. (2018)*, it took the authors approximately 48 hours per glacier (with an average of 367 traces per glacier) (*Goliber et al., 2022)*. That is something we want to change. And there are still many large Greenland glaciers where we don't have enough data to perform seasonal analyzes. Manual delineation did not, and in our opinion never will, keep up with the amount of satellite imagery. The spatially uneven sampling of the

TermPicks repository (*Goliber et al., 2022)* makes this very clear. Although having much higher satellite revisit times it has significantly less calving front data in north and north-east Greenland than in west Greenland. Our statements were intended to refer to these glaciers (and for L173 explicitly naming Humboldt Glacier, Zachariæ Isstøm and Nioghavfjerdsbrae).

That said, we are very thankful for raising and discussing this point. This issue has also been brought up by the other referees. The formulations (L31-35, L52-53, L166-168 and L173) are not accurate in their current form and we will reframe the language. The phrase "*unprecedented temporal resolution*" will be 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"?

When assembling the reference data we also made sure that glaciers with heavily crevassed terminus are included. Some examples are shown in Figure 3.

Probably one of most crevassed calving fronts we processed is at Zachariæ Isstrøm. Figure R3 (will not be included in the main manuscript) shows such a predicted calving front. Some of these crevasses get so large that the open water or ice mélange between them becomes visible. A reliable and consistent ANN delineation under these conditions is essential for ensuring an accurate calving parameterization and jump-free time series (see Figure 7 (f) and Figure R1 (in *Reply of RC2)*).

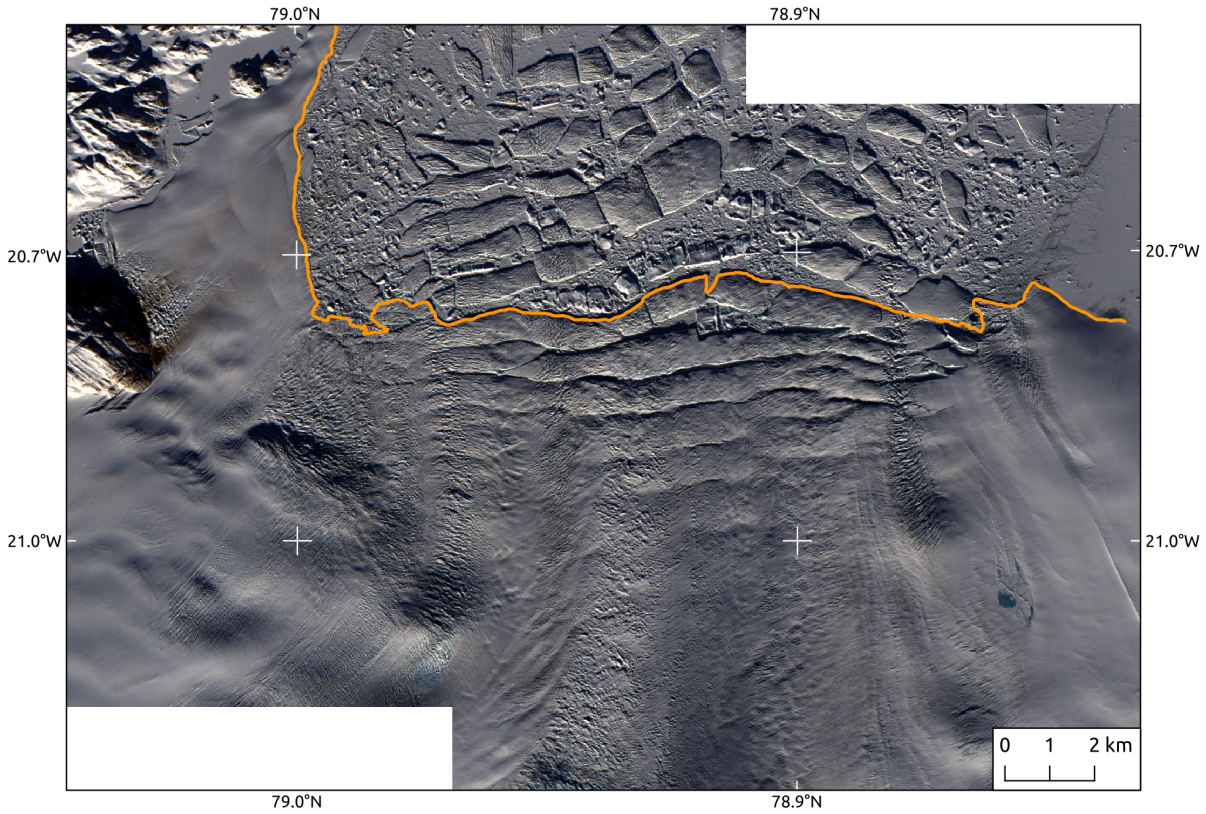

Figure R3: *Calving front of Zachariæ Isstøm on 29.04.2016. ANN prediction is shown in orange. The image is not included in the training data set and is therefore unseen by the ANN.*

We will incorporate and explicitly mention the crevassed conditions in the revised text. Thank you for this question.

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

This point has also been raised by referee #1 (see *Reply of RC1* (page 2)). Although some relevant results are already presented in the figures (e.g. Figure 3 (b) and Figure 7 (j)), we agree that model generalization and spatial transferability has not been addressed sufficiently. Especially, since we put a lot of emphasis on this topic when developing our method.

For the revised manuscript we will discuss spatial transferability in a new subsection in section 2. We will present and discuss model accuracy separately for (1) glaciers outside the training data set, (2) glaciers outside Greenland and (3) glaciers within the training data set. Also we have created a figure showing example validation results specifically for glaciers from Antarctica, Svalbard and Patagonia (similar in style to figure 3).

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).

Thank you for raising this very important point. The distance accuracy estimates are different in almost every study which makes it challenging to directly compare the results. The distance estimate that we use (here: *average minimal distance*) is comparable to the estimates used in CALFIN (*Cheng et al, 2021*) and IceLines (*Baumhoer et al, 2023*).

Incorporating the Hausdorff distance (*Huttenlocher et al., 1993*) will increase comparability to the estimates of *Goliber et al. (2022)* and also improve categorizing our results in general. This is a great idea. Table R2 shows the results for the three testing datasets and will be included in section 2 of the revised manuscript.

Table R2: *Results of the accuracy assessment for the TUD, ESA-CCI and CALFIN test set. Both mean and median of the average minimal distance and the Hausdorff distance is given as a mean over 50 model runs.*

Test datast	Average minimal distance		Hausdorff distance	
	Mean (m)	Median (m)	Mean (m)	Median (m)
TUD	61.2 ± 7.5	28.3 ± 1.4	283.9 ± 28.1	156.4 ± 7.2
ESA-CCI	73.7 ± 2.9	45.9 ± 1.4	352.4 ± 14.1	205.4 ± 10.3
CALFIN	73.5 ± 3.3	43.6 ± 1.6	233.9 ± 5.7	162.9 ± 4.8

As you suspected, the Hausdorff distance is significantly larger than the average minimal distance. Since the Hausdorff distance only considers the greatest distance of all minimum distances along the two trajectories, it is very sensitive towards even small misclassified parts along the predicted calving front. Thus, there is a

bias regarding the length of the calving front. Results will be discussed in the revised version. Incorporating this second distance estimate is a very welcome addition.

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

Thank you for bringing this up. Please see the answer below and our *Reply of RC1* (page 3).

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?

We agree 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 #1. We will rework and expand Section 3 to include all the corresponding information. Please have a look at our *Reply of RC1* (page 3).

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.

We agree with this assessment. In the revised version we will link our results to the results of existing studies. As this point was also raised by the other referees, please refer to our *Response to RC2* (page 2).

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.

This point has also been raised by Referee #1. We welcome this recommendation and will incorporate appendix A and B into the main body of the manuscript, specifically in Section 2.

Technical corrections

L72: change "build" to "built"

Will be fixed, thank you.

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).

The version number will be added in the main text.

L172: remove "are" after Nioghalvfjerdsbræ

Will be removed, thanks.

References

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.

Heidler, K., Mou, L., Loebel, E., Scheinert, M., Lefèvre, S., & Zhu, X. X. (2023). A Deep Active Contour Model for Delineating Glacier Calving Fronts. *IEEE Transactions on Geoscience and Remote Sensing*.

Baumhoer, C. A., Dietz, A. J., Heidler, K., & Kuenzer, C. (2023). IceLines–A new data set of Antarctic ice shelf front positions. *Scientific Data*, *10*(1), 138.

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 (2022). TermPicks: a century of Greenland glacier terminus data for use in scientific and machine learning applications. *The Cryosphere*, *16*(8), 3215-3233.

Zhang, E., Catania, G., & Trugman, D. T. (2023). AutoTerm: an automated pipeline for glacier terminus extraction using machine learning and a “big data” repository of Greenland glacier termini. *The Cryosphere*, 17(8), 3485-3503.

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.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, *15*(9), 850-863.