

Dear Baptiste Vandecrux,

The additions to our data product and the revision of the manuscript took longer than expected. We would like to thank you for extending the deadline.

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).

In particular, the following four major changes were made to the dataset and manuscript.

- We have significantly increased the coverage of our data product from 19 to 42 glaciers. The updated data product is now available on PANGAEA (same doi).
- The introduction has been significantly revised. This is in line with Benjamin Davison's (RC1) comments and our responses.
- The accuracy assessment of our data has been extended. The biggest changes are the introduction of the Hausdorff distance estimate and the inter-model distance. The inter-model distance is now also included in the attribute table for each extracted calving front.
- Section 3 now includes a comparison with existing data products for three examples. This helps to contextualize our work.

Thank you for the work on our Manuscript. Best wishes,

Erik and all co-authors

Benjamin Davison, 28 Feb 2024

General comments:

Firstly, I am not an expert in machine or deep learning techniques, and I can see that the underlying method has already been described in detail in Loebel et al. (2022) and applied to 23 Greenland glaciers in Loebel et al. (2023, in review). I will therefore focus my review on the (1) the dataset itself and the aspects of the methodology relevant to producing a time-series of calving front positions for the community and, (2) the accuracy assessment of the method.

The paper presents an exciting application of an existing deep learning method for delineating glacier calving fronts to 19 glaciers on the Antarctic Peninsula. Compared to Greenland, relatively few terminus position datasets are available for the Antarctic Peninsula and the generation of new terminus positions has not kept pace with the generation of new velocity measurements. As a potential future user of this dataset, it is great to see this application and I am confident that new terminus position delineations on the Peninsula will benefit the community. As such, I am wholeheartedly in support of the generation and publication of these datasets. However, I do not yet think that the presented dataset or manuscript meets the quality and scope required for publication in ESSD, but I am hopeful that the authors will take on board my criticisms and suggestions so that this manuscript and dataset can meet the needs of the community and make best use of the deep learning tool that the author has developed.

Specific comments:

1a) The scope of the dataset

The authors present a total 2064 calving front delineations across 19 outlet glaciers from 2013 to 2023. One of the big questions I had after reading the manuscript was “why not more?”. Just to be clear, I don’t wish to belittle the efforts of the authors – I am sure it is a lot of work to do this and I know it is a lot of work to generate new datasets. However, there are 1,728 basins in the Cook et al. (2014) basin dataset, roughly half of which terminate in an ice shelf, so there are perhaps 800-odd glaciers on the Peninsula that could be targeted by this method. Since the deep learning method was already developed and the majority of the training dataset already existed, and because comparisons to regions outside of Greenland have already been presented in Loebel et al. (2023), it seems like a relatively small additional contribution to run the processing system for just 19 glaciers, especially given that ESSD does not demand any analysis seeking to develop new understanding from the presented dataset, which is typically the bulk of the work in other journals. Again, I am sure it was a lot of work to do this, which I don’t want to detract from, but one of the key benefits of the method used in the manuscript is that it is automatic and much faster than manual approaches, so it should be able to provide “additional and more comprehensive data products”. Therefore, I don’t think it is sufficient to present a terminus position dataset that is (for example) ~25% smaller than that in Wallis et al. (2023), given that the dataset in Wallis et al. (2023) was a relatively small component of their publication. It would be great to see a definitive dataset of terminus positions for the Antarctic Peninsula over the last decade – this and the lead author’s earlier papers demonstrate that we now have the tools and imagery available to achieve this, so I think that is something we should strive for. In order for this dataset to be suitable for publication in ESSD, and to really demonstrate the utility of the underlying deep learning method, I strongly suggest that it should be applied to many more glaciers on the Antarctic Peninsula.

If there is a good scientific or resource reason for limiting the analysis to a small subset of glaciers, then I would still argue for a larger subset including other major glaciers (e.g. Cadman Glacier, which seems like a major omission here), and I think that more justification for the choice of glaciers should be given. At present, the choice is justified twice in the paper, but only briefly and different reasons are given each time.

We have significantly increased the scope of our data product. It now includes 42 glaciers (4817 calving front traces) instead of the original 19 glaciers (2064 traces). Much of the text and figures have been modified or expanded.

Please find the updated data product here (same doi as the first submission): <https://doi.pangaea.de/10.1594/PANGAEA.963725>

For the selection of these 42 glaciers we defined solid criteria (see P3 L67).

1b) Filtering of ‘raw’ terminus positions

One of the main focuses of this paper is that it generates new time-series of terminus positions from an existing method. I was surprised therefore that the manuscript didn’t describe much post-processing of the terminus positions in order to make an analysis-ready time-series. The only filtering step I could see is that the authors “separate all entries that have an area difference of more than 1 km² from the previous and following entries”. I don’t think that is a sufficiently robust outlier

removal technique, especially if you choose to apply this to more glaciers. I suggest that the outlier removal routine should (1) account for the speed of the glacier and time separation between measurements; (2) account for the width of the glacier, because 1 km² changes might be realistic or not depending on their width, and; (3) account for changes along flowlines or similar, not just width-averaged metrics. The time-series presented in the manuscript look reasonably clean, but I had a quick look at the dataset which showed up some places where I think the outlier identification may not be working, for example Birley glacier on 2022/10/03 has what looks like an unrealistic advance along its southern branch and Sjogren-Boydell on 2022/03/04 has a ~4 km retreat across what I think is a large section of land. Whichever outlier removal approach is used, the authors should provide details of how many delineations are removed through using it and how that affects the results.

We have applied the suggested filtering technique (second suggestion) to all 42 glaciers. Section 2.1 has been expanded to include all relevant information (P5 L106). Figure 2 has been expanded to include more details on quality control.

Furthermore we provided more insights into the statistics of the filtering (P6 L113).

On a related note, I couldn't see any description in the manuscript of partial terminus positions. Does the deep learning approach always provide a full terminus trace? What if the glacier terminus is partially obscured by clouds? Please add detail to the manuscript as necessary.

This information has been included in the revised manuscript (P6 L119).

1c) Vectorization of the land/ice probability masks

I couldn't see much justification for the choice of a 0.5 threshold or the impact of that choice of the resulting terminus location. This might be described in the author's earlier papers, but I think ESSD is a suitable place to provide more detail and I think it is relevant to the terminus dataset. Firstly, I think the chosen threshold should be clearly stated in the text, rather than only in the figure (apologies if I just missed it). Secondly, I think there should be a clear quantification, and ideally visualisation, of the impact of that threshold on individual terminus locations and the resulting time-series of area change.

See our response (<https://doi.org/10.5194/essd-2023-535-AC1>).

We have included additional explanations of the vectorisation process (P5 L96). We have also added a new figure to the Supplement (Fig. S2).

1d) Error metric for predicted delineations

Could you provide an error metric for each individual delineation, perhaps by using the spread amongst the 5 models and/or the spread amongst different thresholds? I'm aware that such errors are not provided for manual delineations because it is impractical, but it seems achievable and useful for this method, especially given the apparent differences between each of the 5 models on challenging images shown in Figure 3.

We have included this new metric under the name "inter-model distance". Each calving front trace in our data product now has a corresponding estimate. The metric and its implementation are explained in the new section 2.2.2.

1e) Dataset format

I think that dataset would be much easier and quicker to use if there was just one shapefile for each glacier plus one shapefile containing all delineations for all glaciers. Some users are now also using the geopackage format, so the authors might want to consider providing the output in that format also.

All data is now available in geopackage format in addition to shapefile format.

1f) Dataset contents

It would be great if the training and test data were also released, along with the automatic delineations. Some glaciers and times seem to be missing a 'coastline' shapefile. Is that expected? Also, the justification in the manuscript for providing two outputs is not clear. Why is the coastline file better than the glacier file for merging with an ice mask? It's implied that the terminus file contains only the glacier edge, whereas the coastline file contains the glacier edge and the edge of the surrounding fjord walls, but this isn't demonstrated clearly nor how that distinction is made if part of the terminus is obscured by clouds, for example. Please add some clarification and justification in that respect to the manuscript.

All reference data used in this study (both Greenland and AP glaciers) are now publicly available (with persistent doi). See section 4 for details.

We have added additional information on processing and coastline files (P5 L100). Coastlines files are now only available for the new annual product. There are two reasons for this: Firstly, coastlines are almost static and do not benefit from sub-seasonal sampling. Secondly, as our quality control focuses only on the calving front, we cannot guarantee a high quality of all coastline extractions (all annual entries are quality controlled manually, see answers for RC2 below).

2a) Accuracy assessment

If I have understood it corrected, the accuracy assessment consists of a comparison between the 5-model mean delineation (with a single threshold) and three manual delineations per glacier outside of the training window. Only one of those images for 10 of the 19 glaciers is shown and otherwise we are provided with some simple metrics summarising the results of 19x3 comparisons. In my view, that is not sufficient to characterise the accuracy of the model in this region. Figure 3 is a useful illustration (though I have some suggestions below), but there is no evidence given that demonstrates that the examples given in Figure 3 are representative of the typical accuracy of the method under those conditions, or what the spread in performance is like in each of those conditions. I suggest that the accuracy assessment should include enough images to provide statistically significant accuracy measures of accuracy for glaciers and images with each of the different characteristics shown in Figure 3. Ideally, it would also show the effect of combinations of those conditions, such as times of low illumination with a scene border, with or without mélange and cloud cover.

Figure 3: this is one of the main pieces of evidence presented in the manuscript to convince the reader that the automatic delineations performs as well as a manual delineation. However, showing ~15x15 km images to illustrate differences in position of less than 100 m is not a very clear way to illustrate

those differences – the figures would need to be produced at an impractical resolution and I would need a much better monitor to see anything meaningful, and even then I wouldn't be able to measure the differences. I suggest that an additional, more quantitative figure should be provided, to show differences between the automatic and manual delineations for the full test dataset. Perhaps some simple graphs with 'distance along terminus' on the x-axis and 'difference from manual delineation' on the y-axis would allow the authors to plot the differences through the full test dataset every 30 m along each delineation? That kind of plot would also clearly show how those differences are affected by your choice of model and threshold for vectorization. You could have one graph per glacier and perhaps show a histogram for each glacier.

The accuracy assessment section (section 2.2) has been revised. There is an updated Figure 3 showing the distance error along the calving front trajectories and a new Figure 4 showing the histogram of the average minimal distance over the entire test data. Figure 5 shows the mean inter-model distance histogram over all data product entries (Figs. S3 and S4 separately for each glacier).

I am unsure that the accuracy metrics of mean and median difference from manual delineations is representative of the differences between the automatic and manual delineations with regard to evaluating the use of the automatic delineations for scientific purposes. As far as I can tell, both of those metrics would be insensitive to large differences between automatic and manual delineations if those differences occur over a short section of the terminus. For example, Figure 3 shows automatic delineations on Prospect are in several places over 1 km from the manual delineation, but the mean difference is small because there are comparatively long sections where the two sets of delineations are in close agreement. This shows up a bit in the Prospect timeseries in Figure 4g, where there is a ~10 km² difference between the automatic and manual delineation in late-2022. For glaciological and modelling applications, it might be that those areas of large difference are the bits that matter, so the mean error across the terminus wouldn't be a useful error metric. The other problem with this is that the mean or median difference between the delineations will be highly dependent the length of glacier terminus compared to the length of non-glacier digitised coastline.

We included the Hausdorff distance estimate in our accuracy assessment in Section 2.2.1.

As presented, there isn't a compelling demonstration that the dataset is as applicable to science cases than manually-derived datasets, which I think is important given the proposed justification for making the dataset. Another more holistic approach the authors should take to demonstrate the quality of their time-series product, which would go a long way to addressing that concern, would be to compare time series of area change from these new delineations to area change time-series derived from other terminus position datasets, where both/multiple datasets have sampled the same glacier during overlapping time periods.

We compared our dataset with existing time series for three examples. See Section 3 (from P11 L222) and Figure 7.

3) Introduction

I do not think the introduction adequately justifies the need for improved monitoring of outlet glacier terminus position change. As written, it states that (1) ice shelves have reduced in thickness and extent, which has led to glacier speed-up, (2) calving fronts can be used to study ice-ocean interaction and that (3) they can be used to improve model simulations. Those points are all true, but I think they

need more detail and specifics in order to make a convincing argument for this new dataset. Consider including more detail of ice shelf and glacier area changes (citing the various papers by Cook et al on the subject) and how much the Peninsula has contributed to Antarctica's total sea level contribution. Are there specific examples of where model performance has been limited or improved by the availability of calving front positions, or has it been quantified in a more general sense? For such models, I think they would they need a continual coastline across the whole domain, not small subsets as provided here. In addition, consider drawing on the literature from Greenland, where measurements of terminus position change have led improvements in our understanding of glacier response to environmental conditions over a range of spatial and temporal scales (e.g. Cowton et al., 2018) or have at least aided the interpretation of changes in ice speed, and how they have been used in combination with estimates of submarine melt rates to develop new parameterisations for the impact of submarine melting on calving and terminus position (Slater et al., 2019). Terminus positions are really useful, but I don't think that comes across in the introduction as currently written.

The introduction has been significantly expanded to include more details on AP ice mass loss, changes in ice shelf area, and the use of calving fronts in ice dynamics modeling.

Cowton, T.R., Sole, A.J., Nienow, P.W., Slater, D.A. and Christoffersen, P., 2018. Linear response of east Greenland's tidewater glaciers to ocean/atmosphere warming. *Proceedings of the National Academy of Sciences*, 115(31), pp.7907-7912.

Slater, D. A., Straneo, F., Felikson, D., Little, C. M., Goelzer, H., Fettweis, X., and Holte, J.: Estimating Greenland tidewater glacier retreat driven by submarine melting, *The Cryosphere*, 13, 2489–2509, <https://doi.org/10.5194/tc-13-2489-2019>, 2019.

Minor comments

Line 4: suggest “rely on manual delineation, **which is** time-consuming”

Fixed

Line 8: suggest “The data product presented here”

Done

Line 16: ice **shelf**

Fixed

Line 17: “forcing from ocean and atmosphere has led to reduced ice shelf thickness and extent. And this, in turn, has reduced buttressing strength and thereby increased outlet glacier dynamics”. I don't think this is a fair summary of our current understanding of ice shelf and glacier changes on the Peninsula. Can you add more detail on what is meant by “forcing”. I don't really know what is meant by “increased outlet glacier dynamics” because “dynamics” is a general term for changes in glacier speed, thickness and extent. Consider rewording this sentence to clarify your meaning.

Introduction has been overhauled

Line 20: **utmost**

Fixed

Line 33/34: I think you can make this point more strongly. It's quite possible a reader could look at Table 1 and this paragraph and think "wow, there are loads of terminus position measurements on the AP", because thousands looks like a lot, then they would be confused when they read this statement on line 33/34. So I think it would help to provide some context along the lines of: "there are approximately 800 tidewater glaciers on the AP [you could count them?], so we are currently missing 800 glaciers x 8 illuminated months x 10 years = 64,000 terminus delineations since 2013 (minus five thousand or so from existing studies), even if we only mapped them once per month, but weekly measurements are now possible with the abundance of satellite imagery. Plus many glaciers have only ever been measured a handful of times since 1940 (Cook et al)", or something to that effect.

Introduction has been overhauled

Line 34: "we need to use automatic annotation methods". We don't really need to, as demonstrated by the numerous manual delineations on Greenland, but it is much much faster to do it automatically. So consider rephrasing and combining with the following paragraph to emphasise that we now have the tools available to map them automatically.

Rephrased

Line 44: "new reference data": later this is called "training data" are those different or have you switched terminology?

Clarified in section 2.1

Line 45: This glacier justification is quite weak, but see my major comment above.

We now have solid criteria (see P3 L67).

Line 49: Sjøgren and Boydell were tributaries of Prince Gustav Ice Shelf, not Larsen-A, weren't they?

Fixed

Line 63: need a comma after "pre-processing"

Fixed

Line 82: need a comma after "receptive field"

Fixed

Line 83/84: please specify the threshold for vectorization here and include justification for the choice, and if you add a new figure/section quantifying that impact, it would be good to signpost it here too.

Clarified in Section 2.1 (P5 L96)

Line 85: I don't quite follow this step because the mask hasn't been described. What is the static mask

and how was it derived? What do you do if the glacier retreats or advances beyond the extent of the mask?

Clarified in Section 2.1 (P5 L101)

Line 89: “separated entries are checked manually” and then put back in if you disagree with the algorithm? Or something else?

Clarified in Section 2.1 (P6 L110)

Line 105: “more accurate predictions” than what?

Clarified (P7 L147)

Line 123: for glacier modelling, I think the preference would normally be for a raster mask rather than a vector. Consider including masks in addition to the vector dataset, to facilitate use by the modelling community.

Polygonal masks are included for all coastline predictions.

Line 130: Without expanding this study to other glaciers, I think a combined analysis of circum-Antarctic calving front change would not be possible, so I’m not sure that this statement is warranted with the current dataset.

More glaciers are now included.

Line 134: “such high temporal resolution” is carrying a lot of weight here. Given that the terminus of those glaciers have already been delineated regularly in recent studies (Ochwat et al., 2022; Surawy-Stepney et al., 2023), I don’t think this statement is justified.

Removed

Line 141: I’m not sure what the purpose of this statement is given that this doesn’t appear to be an operational product. Consider removing or rephrasing.

Removed

Figure 1: “Larsen Ice Shelf” should be “Larsen-C Ice Shelf”, if it even needs to be labelled at all.

Label has been removed from Fig. 1

Table 2: I’m not a machine learning user, so this may be a stupid question. Should any of the binary classification metrics have units?

Description added

Figure 3: I think this would be clearer without the manually digitised terminus on. Or at least it would be nice to see a version like that in the supplementary information.

We have made the figure easier to read by adding transparency to the markers and changing the order in which they are plotted.

Benjamin Davis

Anonymous Referee #2, 11 Mar 2024

General Comments:

This work consists of a automatically generated glacial termini data product for 19 key outlet glaciers along the Antarctica Peninsula, and includes 2064 calving front locations from 2013 to 2023 at sub-seasonal temporal resolutions.

The manuscript covers the current state of the art in the field of machine learning/deep learning based cryosphere data extraction methods, as well as the need for such methods to be applied towards glacial data extraction. It describes the importance of calving front data for understanding dynamic glacier changes of marine terminating glaciers, and improving ice modeling. To address the labor-intensive obstacles required by manual delineation, an automatic deep learning-based processing system is developed to extract glacier fronts from satellite imagery. By leveraging the generalization capability of machine learning techniques to provide new observational constraints, this study contributes to groundwork that will enhance the cryosphere community's understanding of glacial dynamics and ice-ocean interactions.

The method uses the deep neural network trained on existing datasets to process Landsat 8 & 9 multiband imagery at spatial resolutions of 30m, and output Shapefile polylines at a spatial accuracy of 59.3 ± 5.9 m (average distance between the measured and predicted fronts). This which falls within human levels of accuracy (<107 m, Goliber et al. 2022).

The dataset itself is composed of zip files of the 19 basins, which are further organized into folders for each observed date, which then contain 2 set of Shapefiles (1 for the coastline, and 1 for the extracted glacial front). Metadata provides the name and date of the processed front. While the scope is small, the dataset still provides valuable new observational constraints.

We have significantly increased the size of our data product. It now includes 42 glaciers (4817 calving front traces) instead of the original 19 glaciers (2064 traces).

The publication is well done, and is largely free of grammatical errors and typographical issues. There are minor remarks to be addressed by the authors, after which I can recommend acceptance at the editor's discretion.

Specific Comments:

Dataset Coastline Quality

While the majority of the dataset is well curated, there are some coastlines in the dataset (i.e., drygalski_20210301_coastline, murphy_wilkinson_20191114_coastline, cayley_20141015_coastline, cayley_20200227_coastline) that seem to have erroneous delineations, particularly along the domain boundaries. More validation or pruning of these data is needed, i.e. by manual pruning through visual

GIS software, or some automated pruning by checking inter-annual differences between fronts to detect outliers.

The entire dataset has been reprocessed (see RC1) using the new width-dependent filtering.

Dataset Coastlines Polygons

In conjunction with the above comment, it would be useful to have the glacial termini data in the form of land/ocean polygonal masks in addition to just a polyline, though this may be outside the scope of this work. This would also resolve the errors along the domain boundaries. Alternatively, provision of the domain boundaries would be helpful, as this would make it easier for modelers/community members to judge where errors are, and/or where the coastlines can be stitched to existing land/ocean masks.

Polygon masks are included for all coastline predictions.

Dataset format

The organization of the dataset could be streamlined, such that the user can load an entire time series of a single domain without having to enter/navigate individual folders for each date, and/or make it more manageable for GIS software on less capable machines to load in all at once. Alternatively, such shapefiles could be consolidated, and the ability of Shapefiles to hold multiple features/delineations within a single file would be of use. Provision of monthly, quarterly, annual, or full time series files (similar to IceLines, Baumhoer et al., 2023) should be within scope.

The updated date product now includes consolidated files for each glacier with all calving front traces. We have also included a new annual product for each glacier. All calving fronts in the annual product have been manually validated. We have not included a time series with monthly and quarterly resolution, as this is not possible due to the data gap during the polar night.

Full time series results (area changes) are available for all 42 glaciers (<https://doi.org/10.25532/OPARA-557>). Details are given in section 4.

Accuracy Comparison w.r.t. Other Datasets

The mean/median distance and binary classification metrics are established accuracy measures in the calving front delineation field, and this study performs well on the evaluated test set. Considering L113P6: (“Although completely different test data sets are involved...Loebel et al. (2023c).”), it may be within scope to see a comparison with existing test sets/studies, to ensure the chosen test set is not biased, and the accuracy metrics are comparable. That being said, the generalization of the network is recognizable, so this can be done at the author’s discretion.

Validation results using the CALFIN and ESA-CCI test sets (used in *Loebel et al. 2024*) are included in the revised version. This includes the new Hausdorff distance estimate.

See also the changes to the accuracy assessment in Section 2.

Minor comments:

- It would be helpful to provide the spatial accuracy of the data to readers in the abstract.

Accuracy is now included in the abstract.