Anonymous Referee #1

General Comments:

This manuscript introduces the novel developed Calving Front Machine “CALFIN” for the automated extraction of Greenlandic calving fronts. This is a major contribution to the field as it replaces time-consuming manual delineated fronts by automatically extracted dense glacier front time series. The CALFIN algorithm was validated extensively against test datasets and results from previous studies through a model intercomparison. The scientific community will definitely benefit from this development as an automatically derived calving front position data set of 66 Greenlandic glaciers will be released with this publication.

Despite the impressive results and technical details of this manuscript, I have some concerns about the structure of this paper and the (sometimes) very short explanations. However, after re-structuring some parts of the manuscript and adding additional information as indicated below, this paper will present an important contribution to the field. In my opinion, the abstract should be structured more clearly. For a better understanding, I would recommend to re-order the abstract by using the common schema: 1) Statement of the problem, 2) Research question, 3) Research design, 4) Central results, 5) Brief interpretation of the results, and 6) Outlook/ future use of the data set.

We thank the reviewer for this feedback and have integrated the suggestions into the manuscript. The abstract has now been rewritten according to the standardized schema as follows:

“Sea level contributions from the Greenland Ice Sheet are influenced by the rapid changes in glacial terminus positions. However, the manual delineation of these calving fronts is time consuming, which limits the availability of this data across a wide spatial and temporal range. Automated methods face challenges that include the handling of clouds, illumination differences, sea ice mélange, and Landsat-7 Scanline Corrector Errors. To address these needs, we develop the Calving Front Machine (CALFIN), an automated method for extracting calving fronts from satellite images of marine-terminating glaciers using neural networks. CALFIN's results are often indistinguishable from manually-curated fronts, deviating by on average 86.76 meters ± 1.43 m from the measured front. CALFIN's outputs use Landsat imagery from 1972 to 2019 to generate 22,678 calving front lines across 66 Greenlandic glaciers. This improves on the state of the art in terms of the spatio-temporal coverage and accuracy of its outputs. The current implementation offers a new opportunity to explore sub-seasonal trends on the extent of Greenland's margins, and supplies new constraints for simulations of the evolution of the mass balance of the Greenland Ice Sheet and its contributions to future sea level rise.”

P2L4: The paper introduces a new method and provides an inter-comparison with other studies. For readers not familiar with the studies of Zhang et al, Mohajerani et al. and Baumhoer et al. it would be helpful to have a brief state-of-the-art paragraph reviewing existing calving front extraction methods. For example, P2L4 could be extended and give more insights into the studies used in the inter-comparison as well as the studies of Seale et al. 2011 and similar approaches.

These suggestions are appreciated, and we focus on the shortcomings of studies like Seale et al. 2011 to handle Landsat 7 Scanline Corrector Errors, as well as expand upon the state of the art by integrating the Existing Works Sect. 6.2 into the introduction. The edited lines are as follows:

"Existing work by Mohajerani et al. (2019) pioneers the usage of these techniques by applying the Ronneberger et al. (2015) UNet deep neural network towards Jakobshavn, Helheim, Sverdrup, and Kangerlussuaq. It achieves a mean distance error of 96.3 m, but is restricted by the preprocessing requirement of aligning the flow direction to be vertical, and inability to handle branching/non-linear calving fronts. Zhang et al. (2019) evaluates a modified UNet applied to TerraSAR-X data over Jakobshavn, and achieves a mean distance error of 104 m, but is limited in scope. Baumhoer et al. (2019) expands the application of the UNet to Sentinel 1 imagery of Antarctica, extracting full coastline delineations and achieving a mean distance error of 108 m. Ultimately, these case studies provide the groundwork for the automatic, accurate, large scale, longtime-series, high temporal resolution, and potentially multi-sensor extraction of glacial terminus positions."

P2L11: In my opinion this section is incomplete. Please mention all potential data sources in Table 1 (add Sentinel-2, Envisat, ERS, Radarsat) and justify why they are not suitable. Another option would be to just focus on Landsat data and remove the incomplete Table 1. Figure 1 is really great so I would try to put the focus on it and highlight the incredible amount of processed data and outline the advantages, data amount, and characteristics of Landsat.

Thank you for these comments - Table 1 has been removed in favor of elaborating on the advantages/characteristics of the data sources evaluated in the study, which now covers Sentinel 1A/B as well.

P2L17: The methodology section could give a short overview of the entire workflow from pre-processing to the final extracted calving front by showing a flow chart. This would guide the reader through the methodology part and link the numerous subchapters of section 3. Besides, in my opinion, the training of the network explained in P12L2 should be part of the methodology and not subject to the discussion.

These are good points, and a methodology flowchart has been added to the beginning of Sect. 3 (see Fig. R1 below). Additionally, the network training discussion subsection Sect. 6.1 has been integrated into the methodology as Sect 3.2p4.

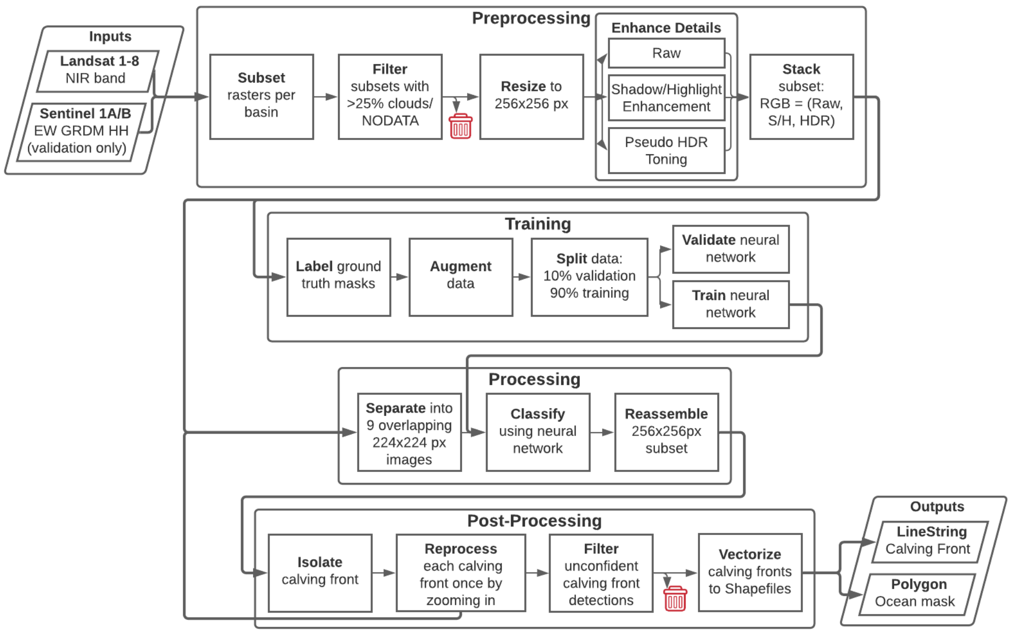

Figure R1. CALFIN Processing Flowchart

Specific Comments:

P5 Figure 5c: How does the filtering of unconfident predictions work? Please describe this in the methodology section.

The filtering of unconfident predictions is performed by measuring the certainty of each pixel's classification in a 5 pixel wide buffer around the calving front. Predictions with a mean certainty exceeding an empirically chosen threshold will be filtered from the results. The following explanation of the method is now given at the end of Sect 3.3p4:

"Since the neural network assigns each pixel a value between 0 and 1 based on its perceived class, any deviation from these two values can used as a measure of uncertainty. The filtering method averages the deviation of the ice/ocean classification mask in a 5 pixel wide buffer around the calving front, and discards any fronts whose mean deviation exceeds an empirically chosen threshold of 0.125."

P6L1: Please outline the calving front re-processing in more detail. Does the reprocessing allow a higher spatial accuracy when re-processing a part of the image?

Yes, the reprocessing allows for higher spatial accuracy when re-processing the image. The re-processing step is now more clearly shown in the Fig. R1 flowchart and described at the beginning of Sect 3.3p4: "Once each front is located, its bounding box is used to extract a higher resolution subset from the original image, and reprocessed. This innovation allows for increased spatial accuracy when processing multiple fronts in large basins."

P6L16: How much smoothing of the extracted coastline is allowed and can this also decrease accuracy?

The smoothed coastline is allowed to vary by no more than 1 pixel from the raw extracted coastline, as seen in Fig. R2. Since the variations are on the sub-pixel scale, the error introduced is no more than the uncertainty of the base resolution, and well within the neural network uncertainty. The following clarification has been added to the end of the line: "“, deviating no more than 1 pixel from the raw extracted coastline.”. Fig. R2 has also been added to the Supplement as Fig. S2.

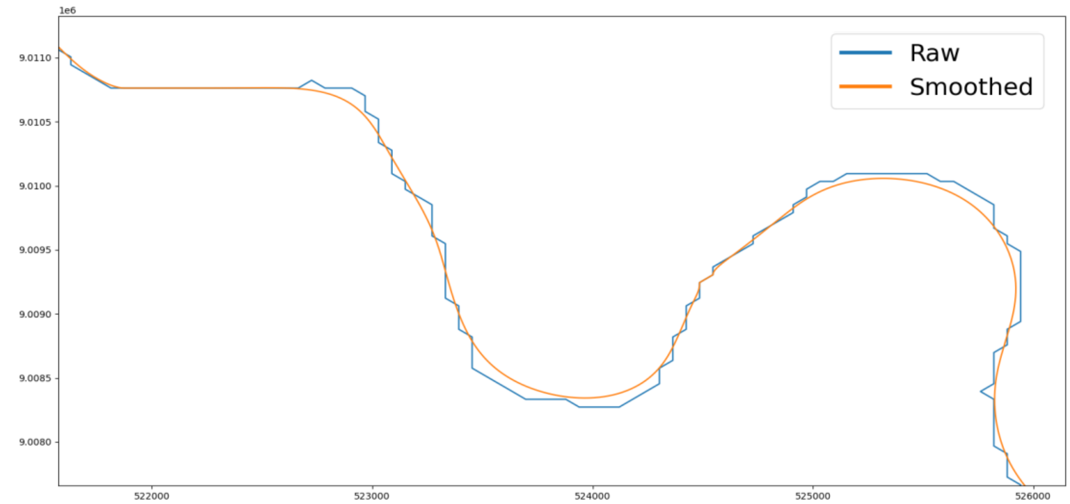

Figure R2. Smoothed (Orange) Versus Raw Coastline (Blue)

P8L1: How did you handle the issue that your network was trained for 3-channel RGB imagery but tested on 1-channel SAR data?

This is question is appreciated, as it highlights the manuscript’s shortcomings in describing the SAR preprocessing pipeline. A paragraph has been added in Sect. 2, Data Source and Scope, describing the usage of the Sentinel 1A/B Antarctic SAR HH band to measure backscatter intensity, which is then treated the same as a Landsat 1-channel NIR band and preprocessed into the final 3-channel false color RGB imagery.

The flowchart in Fig. R1 also helps clarify the input preprocessing steps needed to derive a 3-channel false color RGB image from 1-channel input rasters (now Fig. 3 in the manuscript).

P8L18: What are the characteristics of those outlier glaciers and how many glaciers are defined as “outlier”?

Glaciers with ice tongues such as Kong Oscar can result in large disagreements between the predicted front and the manually delineated fronts. Kong Oscar is the only glacier in the CALFIN Validation Set that contains such extensive ice tongues.

Since the “outlier” in this line refers only to the statistical outlying measurements, and no glaciers are excluded from the error metric calculations, the clause “When excluding outliers such as Kong Oscar, ” has been removed to reduce confusion.

P11L15: The information of this section could also be shifted to methodology. Then rename Chapter 5 to “CALFIN Dataset”.

Thank you for this suggestion - this change has been integrated, and Sect. 5.2 has been removed.

P10L4: But also mention the mean distance which is comparable here.

These lines have been rewritten to include the mean distance error as follows:

"When comparing the mean distance error with the Baumhoer et al. (2019) equivalent Area over Front (A/F) error, the Baumhoer et al. (2019) neural network (B-NN) outperforms CALFIN-NN (330.63 m vs 108 m). Note that the easily detected static coastlines are masked out, raising the relative error, and negatively impacting CALFIN-NN's performance on this metric."

P10 Figure 11: How did you consider the fact that ice shelves are much bigger than glaciers? For example, in Figure 11 you show the Shackleton ice shelf. It is approx. 200 km wide and if you resample that to 224x224 pixels, one pixel for your validation would be 892 m compared to 40 m pixels in the original study by Baumhoer et al. 2019. How did this influence the validation accuracy? For Zhang et al. you show that the use of higher resolution of TerraSAR-X data does not improve the mean distance accuracy (Figure 10).

Errors in large ice shelves are the primary contributor to CALFIN's large mean distance error values. For Shackleton ice shelf, the highly accurate detection prevents it from contributing excessive amounts of error, though indeed variations of even 1 pixel would cause significant error. The following graphs (Fig. R3-R5) shows a histogram that plots the distance between closest pixels in the predicted and manually delineated 3-pixel wide calving front masks. Shackleton's mean distance of 287.48 meters (Fig. R3) for a single validation image is better than the overall average (330 meters) when compared to other large domains like Voyeykov (Fig. R4) and Land (Fig. R5).

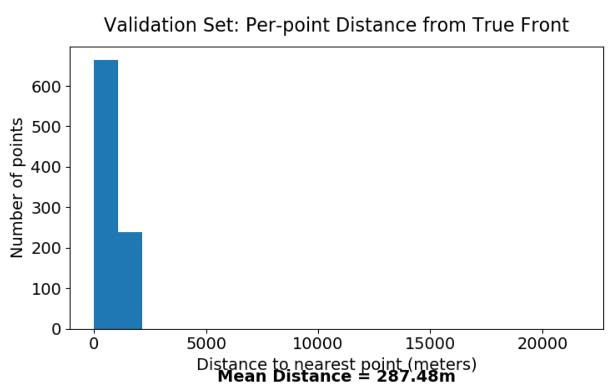

Figure R3. Shackleton Pixelwise Mean Distance Error Histogram

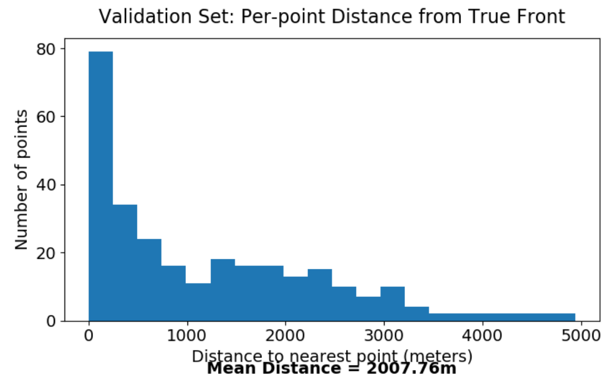

Figure R4. Voyeykov Pixelwise Mean Distance Error Histogram

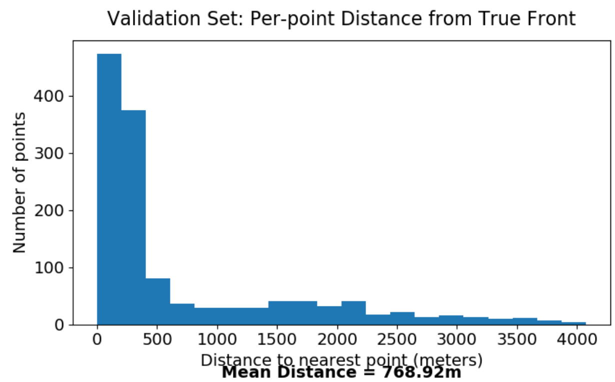

Figure R5. Land Pixelwise Mean Distance Error Histogram

For Zhang et al., the higher resolution inputs are resized to a lower resolution to fit into the 224x224 neural network input shape, and thus provides no improvements. A neural network with a larger input size would benefit from higher resolution imagery.

P13 Figure 13: Can you explain why the PROMICE data set (2008/2009 and 2010/2011) shows twice a very different front position compared to the CALFIN data set?

PROMICE (Anderson et al., 2019) does not provide dates for its delineations, instead stating that they are observed at the "end-of-melt season". August 15th was chosen as the apparent date of these measurements, and it generally corresponds to the other measurements, but it is not a reliable indicator of the calving front at sub-annual timescales, and is only provided for context.

P13L13: The model inter-comparison is only discussed for the study of Mohajerani et al. but validations were also done against the data sets of Zhang et al. and Baumhoer et al., hence those results should also be discussed.

This is a valuable suggestion, and should be investigated in a follow up study, but is unfortunately out of the scope of this study due to the computational and logistical challenges of retraining the original networks used in Zhang et al. and Baumhoer et al. with the CALFIN training set, and the necessary involvement of the original authors in such an in-depth intercomparison.