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.

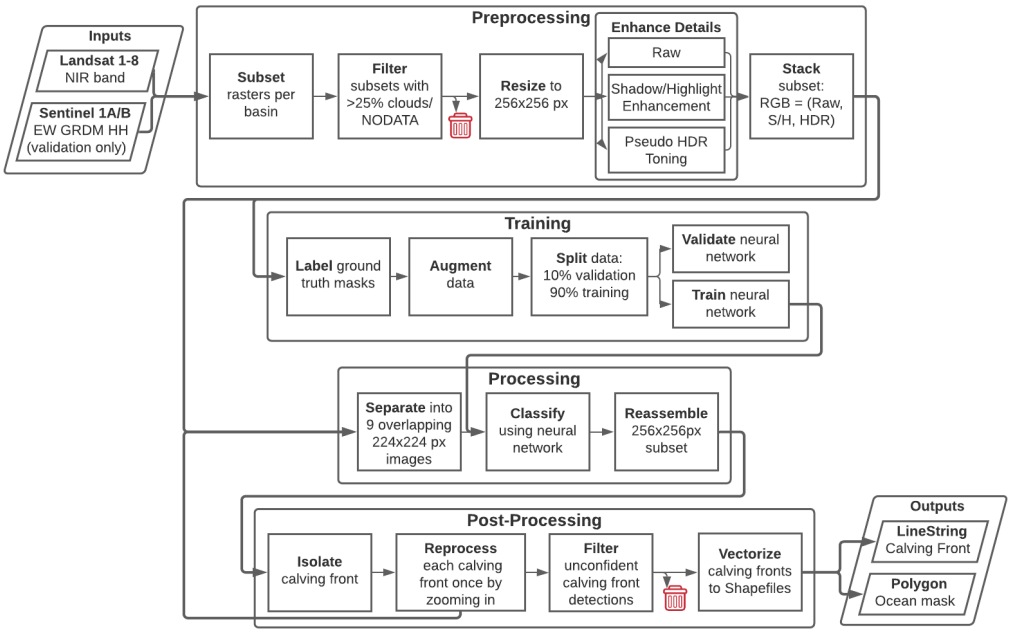

Figure R1. CALFIN Processing Flowchart

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.

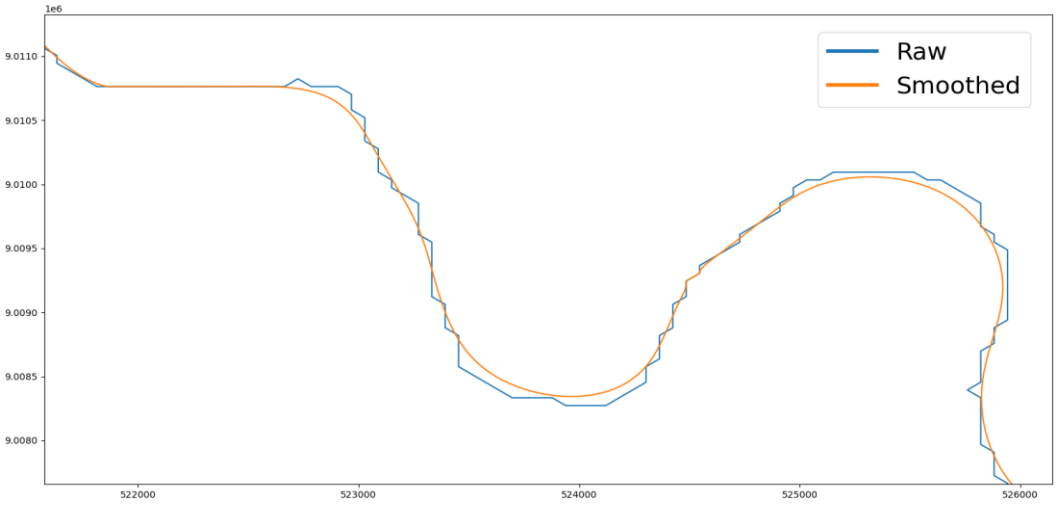

Figure R2. Smoothed (Orange) Versus Raw Coastline (Blue)

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).

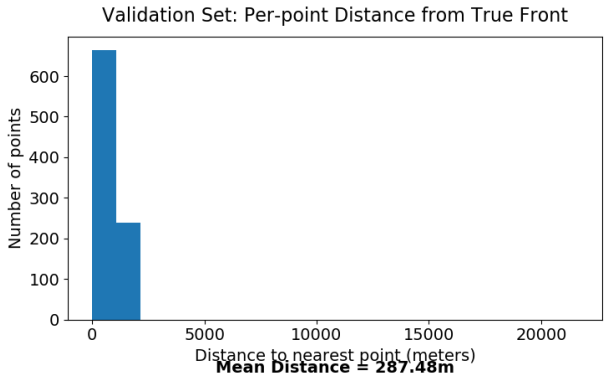

Figure R3. Shackleton Pixelwise Mean Distance Error Histogram

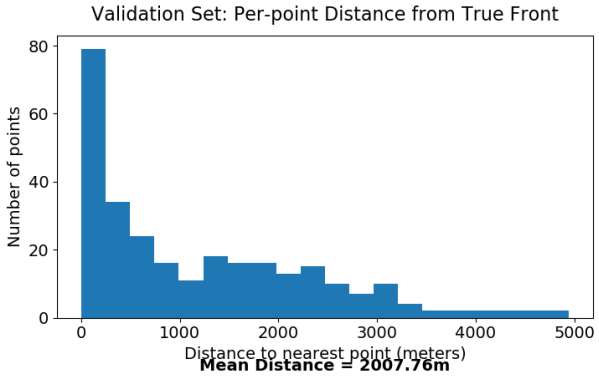

Figure R4. Voyeykov Pixelwise Mean Distance Error Histogram

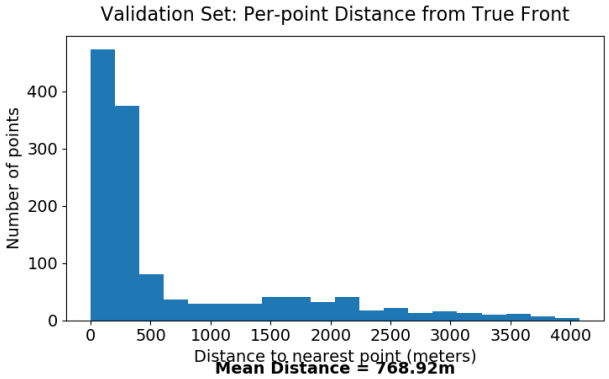

Figure R5. Land Pixelwise Mean Distance Error Histogram

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.

Anonymous Referee #2

General Comments:

General Comment Cheng et al. present an automated method for delineating glacier calving fronts – named Calving Front Machine (CALFIN) - based on a deep learning approach, accompanied by a new dataset of Greenland glacier termini. The principal input data are Landsat optical images acquired since 1972. The methodology builds on previous work by Mohajerani et al., Zhang et al., and Baumhoer et al. and uses computing systems, named neural networks, that learn patterns in training data, in order to identify similar patterns (such as glacier termini) in new data. The authors detail the various steps of the processing chain and produce a set of shapefiles, which are evaluated and intercompared with both internal and external (manually) retrieved calving front datasets using different quality metrics. The main outcome is an extensive dataset covering 66 outlet glaciers around Greenland with in total 22,679 individual calving fronts encompassing the period 1972-2019. The method and new data set reportedly exceeds the accuracy of previous work and approaches human levels of accuracy in delineating glacier termini, the key takeaway being the maturation of neural networks for automated calving front detection.

Automated calving front extraction is a long sought after goal, that recently gained new attention thanks to advances in modern computing technology and increasing availability of satellite EO data. The use of deep learning/neural networks – the subject of this paper - to achieve this is very promising indeed. This paper by Cheng et al. is a welcome addition to existing literature on this topic as is the associated dataset for the community, expanding on previous efforts. In particular, the extension to the early days of Landsat acquisitions, enabling the retrieval of a dense Greenland dataset covering nearly 50 years, is of great relevance for exploring factors that are controlling the varying response to climate change for the outlet glaciers in this region and for quantifying their contribution to future sea level rise.

That said, I do think there is some room for improvement of the manuscript, both in terms of presentation as well as substance. What is missing is a clear description of the objectives in the introduction, based on a literature review on the current standing, issues and knowledge gaps in calving front extraction based on machine learning. This gives the reader, not so familiar with the topic, as well as the presented methodological decisions and improvements a better context.

We thank the reviewer for their time, comments, and suggestions, which have been integrated into the manuscript. A clear description of the objective has been added to the introduction and abstract. This is based on issues and knowledge gaps covered in the added literature review, which repurposes existing sections to provide better methodological context. Additional references have been added throughout the introduction, and a new paragraph has been integrated 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."

Another weak point is that the 'data analysis' does not go any further than a figure showing a rather simple comparison with existing data sets along a flowline of one single glacier. Even though this is clearly written as a methodology paper this is a missed opportunity to showcase a nice data product in my opinion. Perhaps something can be said about general trends in advance/retreat in different regions. Also, I think some sections and descriptions are too brief and need further expansion. Further comments and suggestions for improvement are provided below:

Several sections have been expanded based on provided feedback. Additionally, the data analysis has been expanded, with a new figure showing the regional trends for NW, CW, CE, SW, and SE Greenland, along with 9 additional glacial flowline graphs:

"

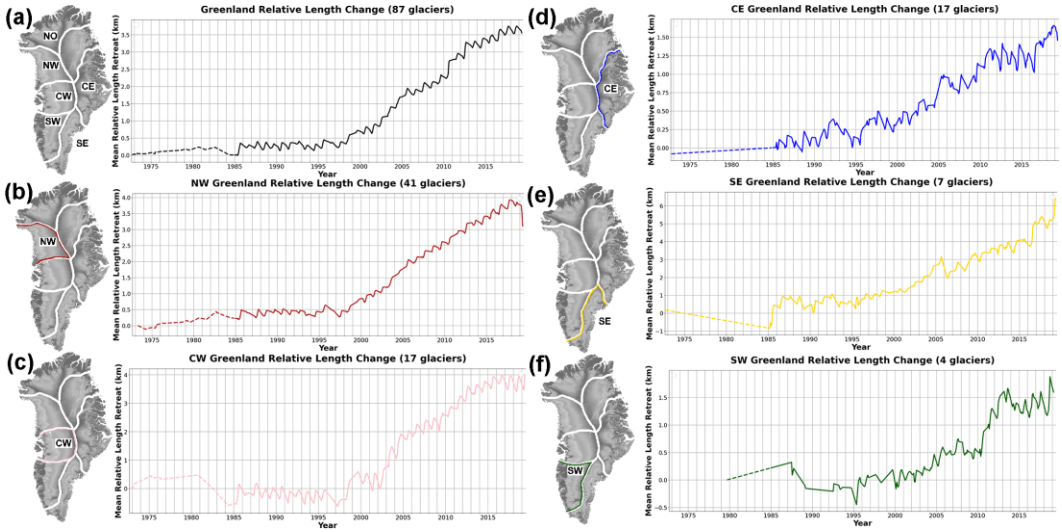

Figure 14. Regional Terminus Advance and Retreat Over Time. (a-f) Regional delineations (left) and terminus position graphs (right)for Greenland (a) and the northwestern (b), central western (c), central eastern (d), southeastern (e), and southwestern (f) regions. Note that the total Greenland mean advance and retreat is unadjusted, and dominated by the trend lines of numerous smaller glaciers in CW and NW Greenland. Note that branches in the 66 studied basins are independently counted, for a total of 87 glaciers.

Additionally, Fig. 14 shows the regional mean advance and retreat change, alongside the mean for the entirety of Greenland covered by the CALFIN dataset. Contributions from NW Greenland influence the overall trend the most, due to the presence of many small glaciers/branches in the regions. Note that the mean for Greenland also includes contributions from Petermann, which is visible in the summers of 2010 and 2012. Shared

regional trends are visible across NW and CW Greenland, which both show relative stability before 2000, followed by steady retreat up until 2017-2018. CE and SE Greenland also share similar but less pronounced retreat, showing accelerating retreat beginning around 1995. These regional trends are less visible in SW Greenland, which is dominated by Narsap Sermia's retreat from 2010-2013. Overall, these regional trends generally agree with studies such as Wood et al. (2021) and King et al. (2020), helping further validate the CALFIN method and data."

Specific Comments:

Pg 1 – Ln 2: The results uses -> the method uses

Done.

Pg 1 – Ln 6: CALFIN provides improvements: briefly describe these improvements

Among existing works, CALFIN improves on the spatial accuracy, is applied towards a large selection of glacial basins, and provides the outputs for scientific usage. "…improvements on the current state of the art." is now described as "…improves on the state of the art in terms of the spatio-temporal coverage and accuracy of its outputs."

Pg 1 – Ln 7: CALFIN's ability to generalize to SAR imagery is also evaluated: briefly describe the outcome.

CALFIN is able to process SAR imagery with similar levels of accuracy when compared to its performance on Landsat image, and is competitive with existing studies. "CALFIN's ability to generalize to SAR imagery" has been moved from the abstract and expanded upon in Sect 2. (see the response to Pg 2 – Ln 12).

Pg 1 – Ln 8: ..deviating by 2.25 px -> deviating by on average 2.25 px

Done.

Pg 2 – Ln 4: Previous techniques -> Previous automated techniques

Done.

Pg 2 – Ln 3: . . .is a a strong. . . -> is a strong

Fixed.

Pg 2 – Ln 7: Something seems to be missing after this sentence, what has been done already on this topic and what are you going to do/improve in this study? See also above issue raised above.

Thank you for raising these points - the section has been expanded upon, and now includes a literature review of existing work and a statement of goals. The added text is 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 for 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. Zhang10et al.

(2019) evaluates a modified UNet applied to TerraSAR-X data over Jakobshavn, and achieves a mean distance error of104 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, long time-series, high temporal resolution, and potentially multi-sensor extraction of glacial terminus positions. This study seeks to assess the feasibility of achieving robust automatic extraction for a15selection of Greenland's glaciers, and to provide the resulting dataset for use by the wider community. Additionally, this study seeks to assess improvements to the neural network design and post-processing methods."

Pg 2 – Ln 9: Sect 4.1 -> Sect 4

Fixed.

Pg 2 – Ln 9/10: Sect. 5 and Sect. 6 shows as well as discusses the results -> Sect. 5 and Sect. 6 show and discuss the results.

Done.

Pg 2 – Ln 12: Sentinel: Sentinel-1 or 2? Not clear from table or text.

We use Sentinel 1 - this is now addressed by a new paragraph at the end of Sect. 2, describing the addition of Sentinel 1A/B Antarctic SAR data for the sole purposes of training and validating the CALFIN methodology. We have added the following new paragraph to this Section:

"For the training and validation of the CALFIN methodology, Sentinel 1A/B SAR images are added to enforce the applicability of the method to other sensor types and domains. The area of interest for the training and validation of the methodology thus includes Antarctic SAR data in addition to the Greenlandic Landsat optical data (see Sect. and Fig. S4). The product used is the Extra Wide Swath, Ground Range Multi-Look Detected, 40 meter resolution HH polarization band. The other data products and polarization bands are not used since the HH backscatter intensity provides sufficient information for the data processing methodology to succeed. A characteristic of Sentinel 1A/B - and SAR data in general - is the presence of speckle noise, which is addressed by the methodology described in the following section."

Pg 2 – Section 2: This section is too brief and there is no need to add the table if only Landsat data is used in the current work as stated. Aside, it is not clear which Sentinel is meant, e.g. the Sentinel-1 SAR satellite has a repeat cycle of 6/12, not 10/12, Sentinel-2 has 10 days but is optical. Why not use higher resolution 15 m panchromatic band Landsat data?

Thank you for raising these points – the first is addressed by the revisions to Sect. 2, which describes the use of Landsat data for dataset production, and both Landsat as well as Sentinel 1A/B data for training and validation.

The 15-meter resolution panchromatic band is not used due to resolution bottlenecks in the data processing methodology. In other words, the increase in resolution did not provide significant increases in accuracy, as it would be downscaled to the same

resolution as the 30 meter inputs to fit the small neural network input size. This clarification has been added to the end of the first paragraph in Sect. 2.

Pg 2 – Ln 15: The basin selection is based on high drainage volume, based on what source? Also, for robust methodological development it is better to base the selection of study sites on different (fjord/glacier) morphology, scale or front type (e.g. with melange, no melange).

The selection metric is based off the basin area/velocities from Nagler et al., 2015. The basins are indeed also selected for robust methodological development, and the 10 areas of interest as well as any nearby basins were selected to contain unique features like ice tongues, branches, and various mélange types. The line now states this explicitly as "The basins are selected for their high drainage volume, wide spatial distribution, and diverse morphological features."

Pg 2 – Ln 20: remove space at beginning.

Fixed.

Pg 3 – Ln 1: This produces -> This results in

Done.

Pg 4 – Ln 2: resized: Do you mean crop or actually resize, as the latter would involve changing the resolution?

The subsets are resized, and the resolution is indeed changed. This loss of resolution is addressed by the reprocessing step, where the subset is recropped at the original resolution and resized again, to allow for maximum resolution within the constraints of the neural network input size.

Pg 4 – Ln 1: ..cloud pixel.. -> how are the cloud pixels identified? Did you include a cloud detection?

The cloud pixels are identified using the Landsat QA band, which assigns each pixel a value based on its detected cloud coverage. The line has been clarified as "…cloud pixels detected in the Landsat QA band,". We rely on the provided cloud masks given by Landsat to do additional filtering per subset, as the scene cloud cover filtering only filters raster based on whole scene cloud coverage.

Pg 4 – ln 14/16: encoder/decoder: it would be nice to show this in the figure for clarity

Done.

Pg 4 – Ln 22: 224 px: wasn't it 256, can you clarify?

The 256px subsets are split into 9 224 px overlapping windows. The Sect. 3, Methodology flowchart (Fig. 3) and Sect. 3.2p4 now clarifies this apparent discrepancy.

Pg 4 – Ln 22: What is the effect of the reduction in input resolution?

This is a good question, as the reduction of input resolution allows for greater complexity, faster training, and higher practical accuracy of the model, but limits the maximum theoretical spatial accuracy of the network. We use other methods (such as

overlapping subsets) to extract higher accuracy predictions from the lower input resolution model.

These considerations have been clarified, and the line has been rephrased to state how reducing the input size results indirectly in increased accuracy, from “To facilitate faster training and performance, the input size is reduced from 512 px to 224 px” to “The input size is reduced from 512 px to 224 px to facilitate better computational performance, allowing for additional training and thus higher accuracy”.

Pg 6 – Ln 4: This section is too brief and needs more details on the confidence measure and applied filter criteria.

This is a fair point - the section has been expanded, and surrounding sections have been rearranged to better support the new narrative. The added material is as follows:

“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. After reprocessing, the nature of CALFIN-NN’s dual outputs as a confidence measure is exploited to filter and discard uncertain detections. 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 be used as a measure of uncertainty. The filtering method averages the deviation of the ice/ocean classification masking a 5 pixel wide buffer around the calving front, and discards any fronts whose mean deviation exceeds an empirically chosen threshold of 0.125.”

Pg 6 – Ln 12: Fjord boundary masks: how are these created and based on what source data? Can you expand on this? Also, are they static for the whole time series? I can imagine that ice thinning over several decades affects the ice/ocean/fjord boundary.

Thank you for these questions and comments - the masks are static and manually created using the image subsets and BedMachine V3 for reference. They are static and averaged across the whole time series – while there are indeed minor changes in the coastline over this time, they do not affect the accuracy of the calving front delineation within the fjord.

This has been clarified as “Static masks of the average fjord boundaries are first created for each basin using the image subsets and BedMachine V3 for reference”

Pg 6 – Ln 18: . . .verification each. . . -> verification of each

Fixed.

Pg 7 – Ln 2: error -> the error

Fixed.

Pg 7 – Ln 7: data that is -> data that are

Fixed.

Pg 8 – Ln 2: list tables that print -> show tables with

Done.

Pg 8 – Ln 8: CALFIN-VS-L7-only/none: explain what this means

A new sentence has been added to this section, which now defines CALFIN-VS-L7-only/ CALFIN-VS-L7-none: “To evaluate performance on Landsat 7 Scanline Corrector Errors, the validation subset CALFIN-VS-L7-only isolates images with L7SCEs, and the CALFIN-VS-L7-none excludes images with L7SCEs.”

Pg 8 – Ln 11: Antarctic basins: this contradicts Pg 2 - Ln 14 stating that the area of interest is restricted to Greenland

This observation is appreciated - the response to Pg 2 – Ln 12 addresses this by adding a new paragraph at the end of Sect. 2, Data Source and Scope, describing the addition of Antarctic SAR data for the sole purpose of training and validating the CALFIN methodology.

Pg 8 – section 4.3.1: The varying conversion of pixels to distance in this paragraph is confusing, can you clarify this, what is the pixel resolution, how is this calculated, why does it vary?

The pixel conversion varies due to 2 effects: images are reprocessed at lower sizes due to detection failures (see Fig. 5c), and pixel error increases as resolution decreases (see Sect. 4.1). Since the pixel-to-meter rate is depends on the scaling factor of each subset, the distribution of rates changes as Landsat 7 images are added/removed.

The methodology flowchart and the elaboration of the filtering/reprocessing step should make this interaction of effects more understandable.

Additionally, the addition of scales to the subsets should aid in communicating the different pixel to meter conversion ratios per subset.

Furthermore, the pixel error metrics have been removed from the paragraph to reduce confusion and to not detract from the more intuitive meter error metrics.

Pg 9 – Ln 2: generalization capability: please briefly explain what this means.

In this context, generalization capability is the ability of a neural network to accurately make new predictions on data it has not been trained on before.

The line “This demonstrates the generalization capability of CALFIN-NN” has been clarified as “This demonstrates CALFIN-NN’s ability to accurately process new data”.

Pg 9 – section 4.3.3 & 4.3.4: For both intercomparisons the mean pixel distance comparisons is skewed, in the caption of figure 11 it is also mentioned ‘undeservedly’. How then can we use this metric to decide which one is better?

This is a good question - the mean pixel distance metric can be used to decide which network is better only when comparing neural networks of the same input size. Indeed, the metric is not useful when comparing networks of different input sizes, since it favors smaller input sizes.

We still provide the metric for comparison to provide additional context when comparing CALFIN with existing studies, as these studies have done the same.

Pg 11 – Ln 14: make sure to make this an active link.

Fixed and verified.

Pg 12 – Ln 3-5: Too brief, more discussion needed to explain the loss function.

Thanks for this noting this shortcoming in the manuscript - a more detailed explanation and relevant equations have been added as follows:

"To increase accuracy, a custom loss function optimizes the binary cross entropy and Intersection-over-Union (see Eq. 1, Sect.4.1). This penalizes mismatches between calving front pixels in the predicted ($\mathbf{I_{cf}}$) and measured ($\hat{\mathbf{I}}_{\mathbf{cf}}$) image masks. Similarly mismatched ice/ocean pixels in the predicted ($\mathbf{I_{io}}$) and measured ($\hat{\mathbf{I}}_{\mathbf{io}}$) image masks are less heavily weighted by an empirically chosen factor of $\alpha = 1/25$, as seen in the final loss function $\mathbf{L}$ in Eq. 2."

$$BCE\_IoU(\mathbf{I},\hat{\mathbf{I}}) = -\mathbf{I}\cdot\log(\hat{\mathbf{I}}) - (1-\mathbf{I})\cdot\log(1-\hat{\mathbf{I}}) - \log\left(\frac{\mathbf{I}\cap\hat{\mathbf{I}}}{\mathbf{I}\cup\hat{\mathbf{I}}}\right) \quad (1)$$

$$\mathcal{L}(\mathbf{I_{cf}},\hat{\mathbf{I}}_{\mathbf{cf}},\mathbf{I_{io}},\hat{\mathbf{I}}_{\mathbf{io}}) = \alpha\cdot BCE\_IoU(\mathbf{I_{io}},\mathbf{I_{io}}) + (1-\alpha)\cdot BCE\_IoU(\mathbf{I_{cf}},\hat{\mathbf{I}}_{\mathbf{cf}}) \quad (2)$$

Pg 12 – Ln 5: Explain what is meant by "over-fitting"

In this context, "over-fitting" means that the model has been trained too heavily on a small dataset, and has only effectively memorized it instead of learning more general features of the observed data. This prevents it from accurately making predictions on new data, as it has "over-fit" the training data.

These lines have been rephrased to be clearer, from "To prevent over-fitting the neural network" to "In order to train the neural network", and from "Another measure to prevent over-fitting involves data augmentation" to "Data augmentation is used during training to increase the accuracy of the network when processing new data".

The only other instance of "over-fitting" on Pg 15, Sect 6.4. is elaborated as "over-fitting, or memorizing,".

Pg 12 – Ln 12-13: Once. . .processing: sentence incomplete.

Thanks for catching this error. The line has been fixed and rephrased from "Once trained, an NVIDIAGTX 1080 with 6GB VRAM for off-line data processing" to "Once trained, an NVIDIA GTX1060 with 6GB VRAM is used for the off-line data processing of the 20188 GeoTIFF subsets". The phrase "of the 20188 GeoTIFF subsets" has been moved from a subsequent line to clarify what data is being processed off-line.

Pg 12 – Ln 25: While the methodology is restricted by its preprocessing requirements and inability to handle branching/nonlinear calving fronts: How are the preprocessing requirements different?

The primary difference in preprocessing requirements is the necessary alignment of the flow direction to be vertical. This line has been elaborated as "the preprocessing requirement of aligning the flow direction to be vertical".

Pg 12 – Section 6.2: Some of this existing work description should go to the introduction to show where gaps/shortcomings are and as motivation for the improvements introduced in the current implementation.

Thank you for this suggestion - Sect. 6.2 has been integrated into the introduction, along with descriptions of the gaps/shortcomings of each approach that form the motivation for the study. The end of the first paragraph of the introduction now reads, “Existing work by Mohajerani et al. (2019) pioneers the usage of these techniques by applying the Ronneberger et al. (2015) UNet deep neural network for 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-10X 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. This study seeks to assess the feasibility of achieving robust automatic extraction for a selection of Greenland’s glaciers, and to provide the resulting dataset for use by the wider community. Additionally, this study seeks to assess improvements to the neural network design and post-processing methods.”

Pg 13 – Section 6.3: As mentioned in the general comment, this section is hardly a data analysis and very brief, even the description of the figure. A clear improvement, obvious from the figure, is the much denser and longer temporal coverage, this should be mentioned somewhere.

Thank you for pointing out this weakness in the original manuscript. The figure description has been expanded with the following details: “Note the seasonal variations shown by the solid lines, and the dotted lines from 1972-1985 that indicate a lack of such seasonal observations. Also note that the vertical axis scaling is applied differently for each graph to highlight seasonal trends.” Text that highlights the denser and longer temporal coverage has been added throughout the section. Furthermore, the original Fig. 12 (now Fig. 13) has been expanded to include additional flowlines:

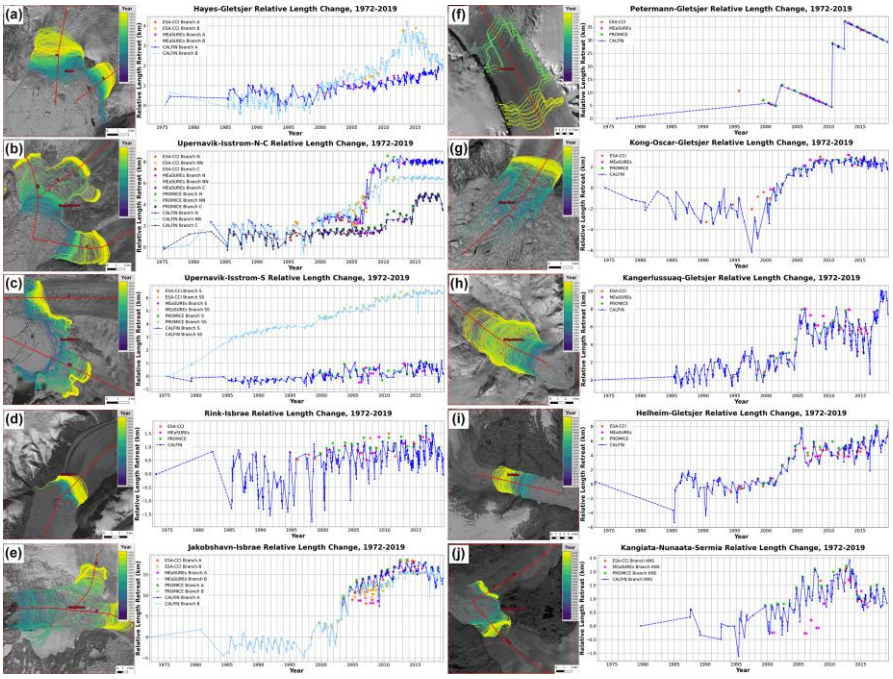

Figure R1. Updated Terminus Advance and Retreat Over Time

See also the response to the general comment for additional added content that adds to the data analysis.

Pg 13 – Ln 2: validate -> compare

Done.

Pg 13 – Ln 7: length change -> I would rather call it “advance and retreat”

Done.

Pg 13 – Ln 18/19: To perform . . . the results: this sentence seems incomplete.

Thank you for noticing this – the sentence has been rephrased for clarity, from “To perform this task, the M-NN is retrained using CALFIN training data, process validation data, and compare the results” to “This task involves retraining the M-NN on CALFIN training data, and comparing its performance against CALFIN-NN using a shared validation set”.

Pg 14 – Ln 18: ground truth fronts: None of these fronts are actual ground truth fronts, even when manually delineated (also elsewhere in manuscript).

This is a good point, and has been corrected from “ground truth” to “manually delineated” throughout the manuscript.

Pg 15 – Ln 2: Overall, the goal of . . .: this goal was nowhere clearly stated

This observation is appreciated, and the introduction has been edited to include this goal, which is stated as, “This study seeks to assess the feasibility of achieving robust automatic extraction for a selection of Greenland’s glaciers, and to provide the resulting

dataset for use by the wider community. Additionally, this study seeks to assess improvements to the neural network design and post-processing methods."

Figures/Tables:

Most figures lack a proper scale bar, this would be very helpful to evaluate the different results. Also, individual lines are sometimes very difficult to distinguish (for example in fig 10). Not sure if this can be improved.

Thank you for this suggestion - scale bars have been added in Figs. 9-13, and high contrast colorblind-friendly line colors have been added for Figs. 6-12.

Table 1: As no data other than Landsat is used in the study, I don't see much need for this table. See issue raised previously.

Table 1 has been removed.

Figure 1: For a nicer figure, updated maps, without gaps, are available at the Greenland Ice Sheet CCI website (see: http://esa-icesheets-greenland-cci.org/)

Thanks for this suggestion, Fig. 1 has been updated to utilize an updated gapless velocity map.

Figure 2: The legend should provide a range

Fig. 2 key has been updated to show the full range of the data.

Figure 3 & 5: No need to add c) in my opinion

This is a fair suggestion that highlights the lack of importance placed on the filtering step in the manuscript. To address this concern, Fig. 3 & 5 (now 4 & 6) have added a visualization of the filtering under (c), as shown in the new flowchart (now Fig. 3).

Figure 6: It appears that several 'difficult' sections/gaps are connected with a straight line, how does this work (e.g. what gap tresholds are used)?

This is a valuable question that highlights the manuscript's insufficient explanation of this algorithm. Gaps are given negative exponential distance-based weights, so that they add a penalty to the maximum path, but can be used if they connect two long paths in the final Minimum Spanning Tree. An explanation of this behavior has been added to the end of Sect. 3.3.1: "Such gaps are given weights based on the negative exponential distances between nodes, which allows for connections if the paths connected are significantly longer than the gap itself."

Figure 6a: I don't see a red coastline mask

Fig. 6 (now Fig. 7) has been updated to use a high contrast colorblind-friendly color scheme, and the red coastline mask has been enhanced to make it more visible.

Figure 8-12: There seem to be no references in the text to these figures, please add.

Thank you for noting this, references to these figures (now Fig. 9-13) have been added in the text.

Figure 12: caption “Sample” -> Examples

Done.

Figure 13: caption “1995-2016 (ESA-CCI), 2005-2017 (MEaSUREs)”: check years vs line in image, ESA CCI starts in 1990, MEaSUREs in 2000

Fixed.

Calving Front Machine (CALFIN): Glacial Termini Dataset and Automated Deep Learning Extraction Method for Greenland, 1972-2019

Daniel Cheng[1], Wayne Hayes[1], Eric Larour[2], Yara Mohajerani[1,3], Michael Wood[2], Isabella Velicogna[1,2], and Eric Rignot[1,2]

[1]University of California at Irvine, Irvine CA, USA
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA, USA
[3]University of Washington, eScience Institute and Department of Civil and Environmental Engineering, Seattle, WA, 98195, USA

Correspondence: Daniel Cheng (dlcheng@uci.edu)

Abstract. ~~We present~~ 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 ~~. Our results use~~ using neural networks. The results are often indistinguishable from manually-curated fronts, deviating by on average 86.76 meters ± 1.43 m from the measured front. Landsat imagery from 1972 to 2019 ~~, and generate 20,004~~ is used to generate 22,678 calving front lines across 66 Greenlandic glaciers. ~~Our method uses deep learning, and builds on existing work by Mohajerani et al., Zhang et al., and Baumhoer et al. Additional post-processing techniques allow our method to achieve accurate segmentation of imagery into Shapefile outputs. This method is uniquely robust to the impact of clouds, illumination differences, ice mélange, and Landsat-7 Scan Line Corrector errors. CALFIN provides improvements on the current~~ This improves on the state of the art ~~. We show this by performing a model inter-comparison and evaluate performance against existing methodologies. We also evaluate CALFIN's ability to generalize to SAR imagery. CALFIN's fronts are often indistinguishable from manually-curated fronts, deviating by 2.25 pixels (86.76 meters) from the true front on a diverse set of 162 testing images~~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.

1 Introduction

The evolution of Greenland's tidewater glaciers is an important constraint on the evolution of the Greenland Ice Sheet (Nick et al., 2013) . Likewise, changes in Greenland are important in tracking and predicting future sea level rise over the next century (Andersen et al., 2015; F

. Constraining Greenland's glacial evolution is thus an important part of improving ~~our~~ the understanding of the earth system as a whole. One constraint on glacial evolution is the position of glacial calving fronts and ice margins over time (King et al., 2018) . Currently, most calving front delineation is done with time-consuming manual labor (Carr et al., 2017; Bunce et al., 2018; Catania et al., 2 . This results in the ~~severe~~ under-utilization of available satellite imagery~~. As a result, many smaller glaciers have no calving front data, while others have annual or seasonal coverage at best~~, and causes gaps in seasonal records that introduce uncertainty when modeling past and projected climate change (Catania et al., 2020). Significant efforts have been made to improve this situation, which include the ESA-CCI dataset of 26 Greenlandic glaciers from 1990-2016, the PROMICE dataset of 47 glaciers from 1990-2018, and the MEaSUREs dataset of 200+ glaciers from 2000-2017 (ENVEO, 2017; Andersen et al., 2019; Joughin et al., 2015). Yet the increasing availability of new datasets through missions like Landsat 8 and the release of old datasets through improved reprocessing call for new automated ways of detecting calving front delineations. In particular there is a ~~a~~ strong need for these automated ways to be robust, specifically against cloud cover, ice mélange, ~~and shadows. Previous~~ shadows, and Landsat 7 Scanline Corrector Errors. Traditional automated techniques such as ~~edge detection and texture analysis~~ the edge detection utilized by (Seale et al., 2011) and Paravolidakis et al. (2016) have significant challenges with respect to these issues~~(Paravolidakis et al., 2016; Malik et al., 2001)~~. Modern machine learning techniques and deep neural networks provide a robust, scalable, and accurate solution to these processing challenges. Existing work by Mohajerani et al. (2019) pioneers the usage of these techniques by applying the Ronneberger et al. (2015) UNet deep neural network for 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, long time-series, high temporal resolution, and potentially multi-sensor extraction of glacial terminus positions. This study seeks to assess the feasibility of achieving robust automatic extraction for a selection of Greenland's glaciers, and to provide the resulting dataset for use by the wider community. Additionally, this study seeks to assess improvements to the neural network design and post-processing methods.

In this study, ~~we present in~~ Sect. 2 ~~our data~~ covers the data source along with the spatial and temporal coverage. ~~In~~ Sect. 3 ~~we present our~~ examines the CALFIN algorithm and method for processing the data. ~~In Sect. 4.1 we validate our~~ Sect. 4 validates the algorithm through error analysis. ~~In~~ Sect. 5 and Sect. ~~?? we~~ 6 show and discuss ~~our results (~~the results - the calving front dataset and algorithm~~)~~.

2 Data Source and Scope

~~We begin by evaluating several potential data sources, including Terra/MODIS, TerraSAR-X, Landsat , and Sentinel (see Table ??). Landsat is selected for its~~ For the production of the CALFIN dataset, Landsat optical images are used for their long time-series availability and reasonable spatial distribution/resolution.

~~**Potential Data Sources**: A comparison of the data sources available for use. Name Resolution(s) Time Series Repeat Cycle Sensor Seasonal Coverage Landsat 30 m , 60 m 1972-present 16 day Optical Spring-Fall Terra (MODIS) 250 m , 500 m , 1000 m 1999-present 1, 8, 16 day Optical Spring-Fall Sentinel 10 m , 20 m , 60 m 2014-present 10, 12 day SAR Spring-WinterTerraSAR-X 1 m , 3 m , 6 m 2007-present 3-11 day SAR Spring-Winter~~

~~Here we restrict our~~ The area of interest for the dataset production is restricted to Greenland, in particular the calving fronts for 66 Greenlandic basins shown in Fig. 1, spanning the 1972 to 2019 time period shown in Fig. 2. The basins are selected for their high ~~drainage volume and~~ discharge volumes, wide spatial distribution~~.~~, and diverse morphological features. The product used is the 60/30 meter resolution Near Infrared band. The 15 meter resolution panchromatic band was not used, due to computational and logistical limitations. A unique characteristic of this data source is the presence of Landsat 7 Scanline Corrector Errors from 2003-2013, which manifests as black stripes that interfere with automated calving front extraction methods.

For the training and validation of the CALFIN methodology, Sentinel 1A/B SAR images are added to enforce the applicability of the method to other sensor types and domains. The area of interest for the training and validation of the methodology thus includes Antarctic SAR data in addition to the Greenlandic Landsat optical data (see Sect. 3.2 and Fig. S4). The product used is the Extra Wide Swath, Ground Range Multi-Look Detected, 40 meter resolution HH polarization band. The other data products and polarization bands are not used since the HH backscatter intensity provides sufficient information for the data processing methodology to succeed. A characteristic of Sentinel 1A/B - and SAR data in general - is the presence of speckle noise, which is addressed by the methodology described in the following section.

3 ~~Methodology~~Methods

The automated data processing methodology uses innovative techniques and state-of-the-art neural networks to process raw Landsat and Sentinel 1A/B data into useful calving front Shapefiles. The following section explores this methodology, as outlined by the flowchart below (Fig. 3).

3.1 Preprocessing

~~We develop a pipeline (see Fig. 4)that automates much of the data preprocessing that prepares raw data for input into~~ The first stage involves preprocessing the input data for use with the neural network~~.~~

, as illustrated in Fig. 4. The proceeding steps cover the details of handling Landsat data, but can be applied to Sentinel 1 data for validation purposes.

~~The first step is to collect all the input raster images~~ To begin, raster images are selected from areas centered around one of 9 primary glacial basins. These basins include Kong Oscar, Hayes, Rink Isbrae, Upernavik, Jakobshavn, Kangiata Nunaata, Helheim, Kangerlussuaq, and Petermann. Next, ~~we select all the~~ all L1TP (precision and terrain corrected) rasters from Landsats 1-8 with low cloud coverage (<20%) are collected. A few L1GS/L1GT (non-corrected) products are also selected, which ~~we manually georeference, and use~~ are manually georeferenced, and used to fill in Landsat 1-2 time series gaps (1972-1985).

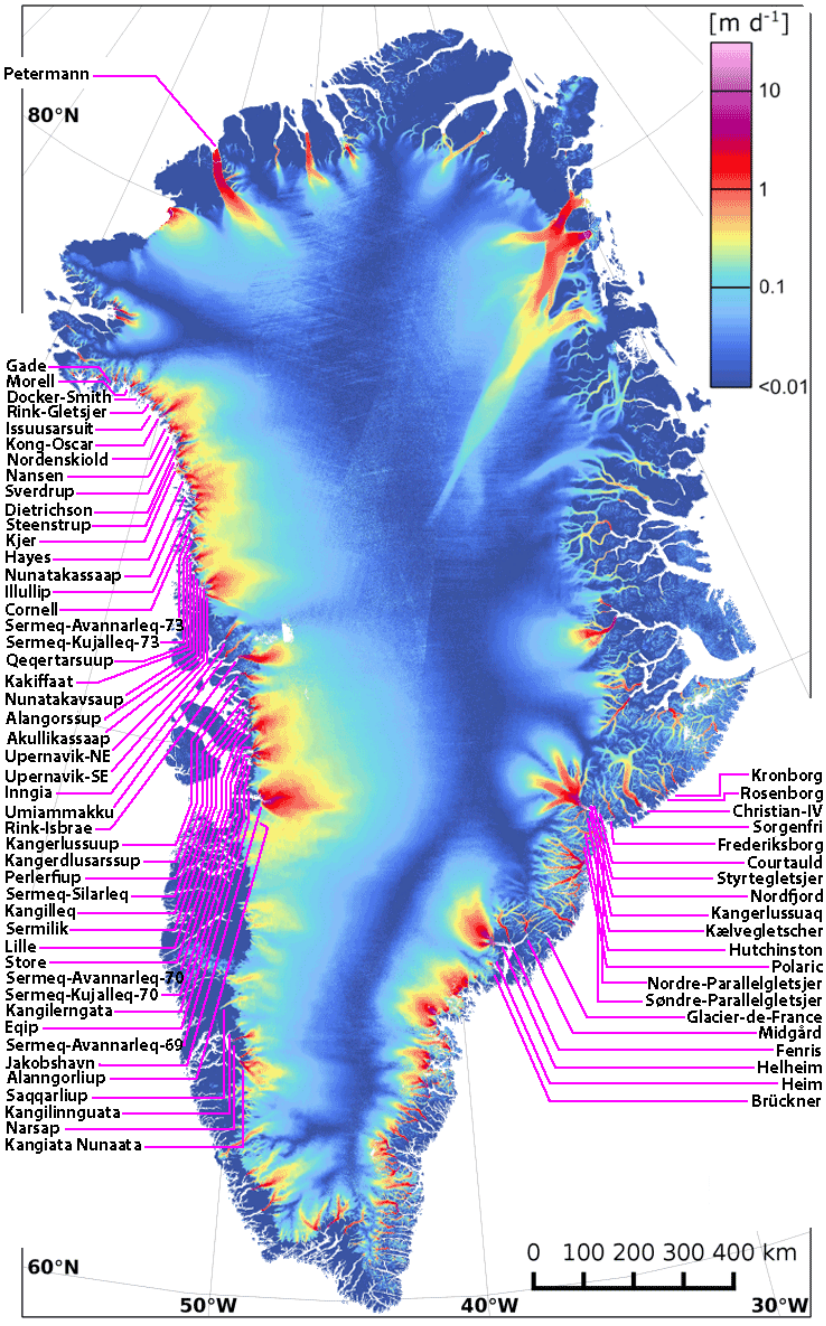

Figure 1. Spatial Coverage Map: Spatial distribution of 66 selected Greenlandic glaciers. The velocity map is taken from Nagler et al. (2015).

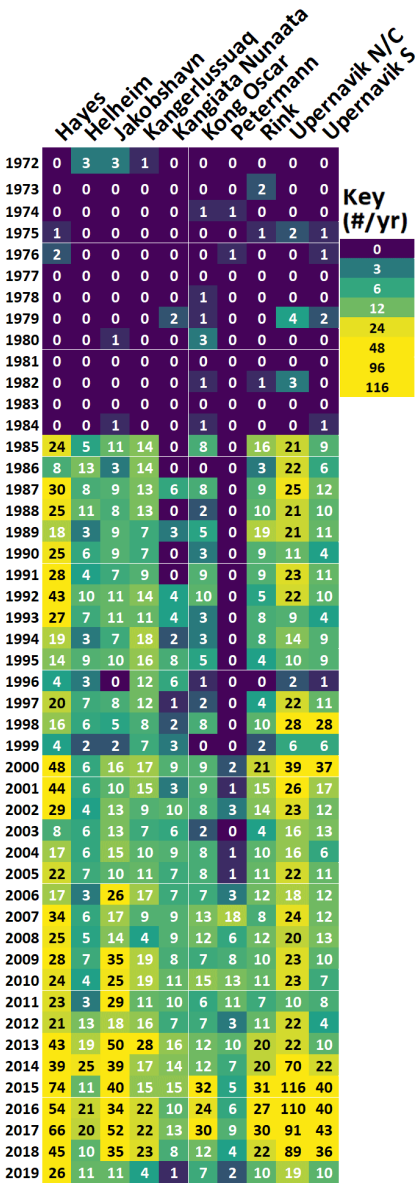

Figure 2. Temporal Coverage Map: Number of fronts per year from 1972-2019 for ~~9 high-drainage glaciers~~10 high discharge volume basins. For the full temporal coverage map, see attached Supplement, Fig. S1.

This ~~produces~~ results in a total of 4956 Landsat rasters. Next, predefined basin domain Shapefiles that enclose the terminus are used to clip the Landsat raster subsets. Additional filtering removes subsets that still contain ≥30% NODATA pixels or ≥20%

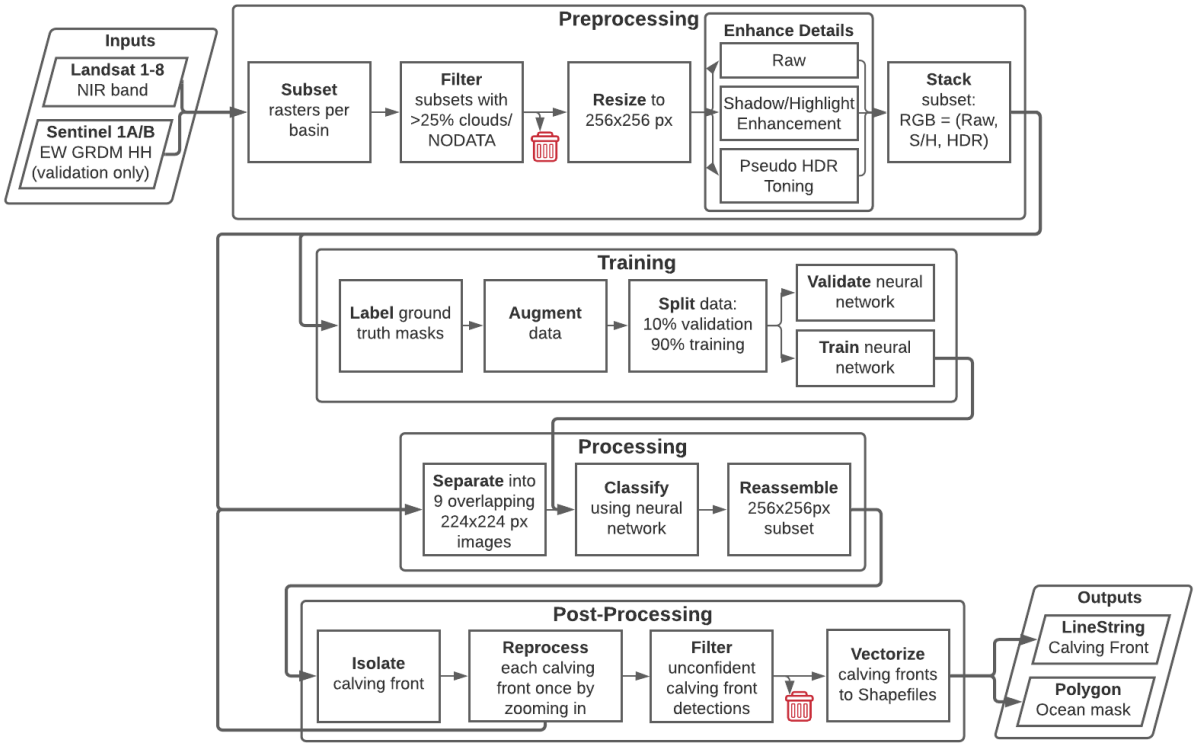

Figure 3. Methodology Flowchart: The CALFIN workflow, which processes single band raster imagery into calving front and ocean mask Shapefiles. Note that Sentinel 1A/B imagery is only used for validation, as it is not corrected and thus not qualified for geolocation/extraction.

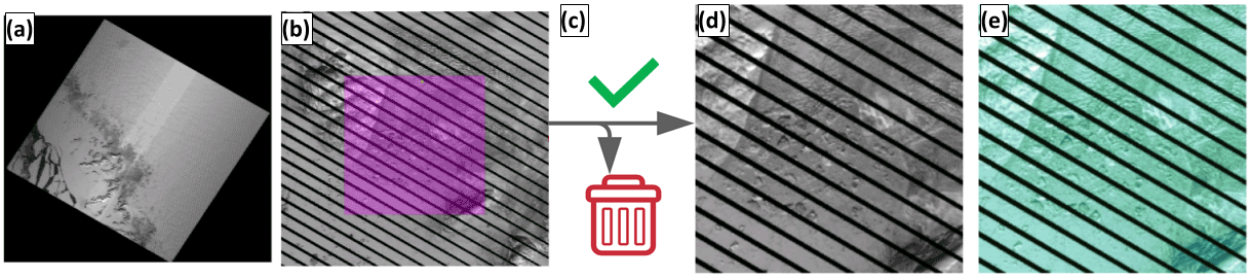

Figure 4. Preprocessing Pipeline: (a) First, input the raw Landsat GeoTIFF rasters with <20% clouds. (b) Next, subset using QGIS/GDAL and the domain Shapefile to clip each raster. (c) Then, filter the clouded/NODATA subsets. (d) Now, resize the subsets to 256x256 px. (e) Finally, enhance contrast and stack with the raw subset.

cloud pixels detected in the Landsat QA band, as subsets that exceed these thresholds are not likely to contain detectable fronts. At this stage, ~~we accumulate~~ 20188 GeoTIFF subsets are accumulated. Each subset is then resized to 256x256 px, and lastly enhanced using Pseudo-HDR Toning (HDR) and Shadows/Highlights (S/H) through Adobe Photoshop. The raw, HDR, and S/H enhanced subsets are then stacked into a single RGB image. At this point, the images are ready for processing into calving front masks.

3.2 Neural Network Processing

Images are processed using the Calving Front Machine Neural Network (CALFIN-NN), as illustrated in Fig. 5. Neural networks like CALFIN-NN work by learning patterns in training data, and finding them in new data. ~~We train~~ CALFIN-NN is trained using manually delineated calving front masks~~, discussed in Sect. ??.~~ . Once trained, CALFIN-NN outputs a probability mask that shows each pixel's likelihood of lying on the coastline/calving front. CALFIN-NN also generates a ~~land~~ice/~~ice-ocean~~ ocean probability mask as a secondary output. ~~Once each image is processed~~Following this, the calving front is ~~ready to be~~ extracted during post-processing~~.~~ , discussed in Sect. 3.3.

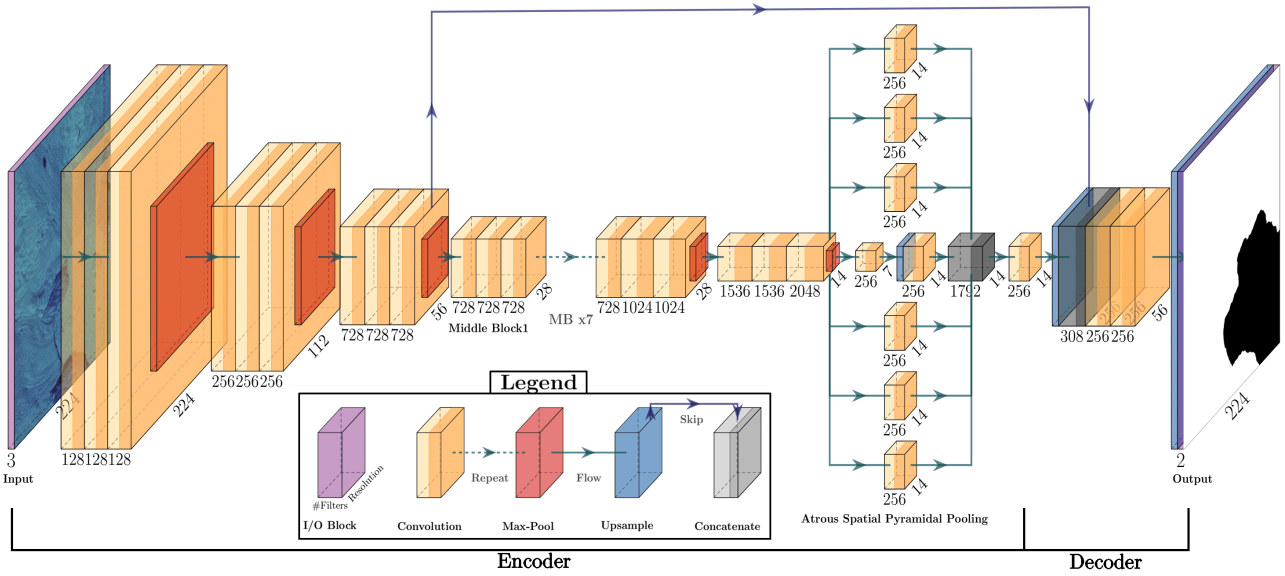

Figure 5. The CALFIN-NN Processing Architecture: Each orange "Xception" block consists of convolution kernels that detect features in the previous block. Blocks are reduced in size periodically to pool increasingly complex and numerous feature maps. "U" shaped connections help refine the probability masks during up-sampling. Note that the 7 repeated "Xception" blocks in the middle section are omitted for brevity.

3.2.1 ~~Network Architecture & Modifications~~

Neural networks are the foundation of several automated delineation methods, including ~~Zhang et al. (2019), Mohajerani et al. (2019) ,~~ Mohajerani et al. (2019), Zhang et al. (2019), and Baumhoer et al. (2019). ~~We build~~ This method builds upon this work, and ~~use~~ uses a modification of the DeepLabV3+ Xception neural network from Chen et al. (2018), as shown in Fig. 5. The first half, the encoder, uses the Xception-65 network to extract image features (Chollet, 2017). It does this by assembling basic features, like edges and corners, into more abstract features, such as glacier/land textures. The second half of the network, the decoder, takes the output of the encoder and up-samples the features to predict the final probability mask outputs.

~~We make several modifications to the~~ Several architectural modifications are made to the original DeepLabV3+ Xception ~~network~~model to enhance its performance. To accurately recognize line-like features such as calving fronts, additional Atrous Spatial Pyramidal Pooling (ASPP) blocks are added in between the encoder and decoder, with the dilation scales 0, 1, 2, 3, 4,

and 5. The number of Middle Blocks (MB in Fig. 5) is reduced from 16 to 8, as the extra discriminative power from those blocks is not needed. ~~To facilitate faster training and performance, the~~ The input size is reduced from 512 px to 224 px to facilitate better computational performance, allowing for additional training and thus higher accuracy. Since the input resolution is reduced, the encoder is also modified to remove several down-sampling "max-pool" layers. ~~Our~~ The last contribution adds a 2-channel output to the decoder, allowing for both calving front ~~mask~~ masking and ice/ocean masking. Together, these changes reduce ~~the size of the network~~ number of model parameters from 40M ~~parameters~~ to 29M~~parameters~~, while also increasing the overall accuracy.

Several techniques are used during the training of CALFIN-NN to improve its performance. First, a large set of training data is manually delineated (see Fig. S4), totalling 1541 Landsat and 232 Antarctic Sentinel 1A/B image/mask pairs, with the Antarctic data taken from the same training scenes used by Baumhoer et al. (2019). Data augmentation is used to increase the accuracy of the network by expanding the training set, which entails adding random amounts of flips, Gaussian noise, sharpening filters, rotations of up to 12°, crops, and scaling to the pre-processed training images. Through empirical testing, it is determined that excessive image padding, rotation, warping, and cropping of calving fronts to close to the image bounds result in sub-optimal performance. Another helpful technique is the use of test-time augmentations, wherein each image subset is cut into 9 overlapping 224x224 image windows and processed individually, before being reassembled into the final 256x256 output mask. This allows for multiple independent classifications of the central pixels, ensuring agreement and confidence in detected calving fronts. To increase accuracy, a custom loss function optimizes the binary cross entropy and Intersection-over-Union (see Eq. 1, Sect. 4.1) (Mannor et al., 2005). This penalizes mismatches between calving front pixels in the predicted ($\mathbf{I_{cf}}$) and measured ($\hat{\mathbf{I}}_{\mathbf{cf}}$) image masks. Mismatched ice/ocean pixels in the predicted ($\mathbf{I_{io}}$) and measured ($\hat{\mathbf{I}}_{\mathbf{io}}$) image masks are less heavily weighted by an empirically chosen factor of $\alpha = 1/25$, as seen in the final loss function $\mathcal{L}$ in Eq. 2.

$$BCE\_IoU(\mathbf{I}, \hat{\mathbf{I}}) = -\mathbf{I} \cdot \log(\hat{\mathbf{I}}) - (1 - \mathbf{I}) \cdot \log(1 - \hat{\mathbf{I}}) - \log\left(\frac{\mathbf{I} \cap \hat{\mathbf{I}}}{\mathbf{I} \cup \hat{\mathbf{I}}}\right) \tag{1}$$

$$\mathcal{L}(\mathbf{I_{cf}}, \hat{\mathbf{I}}_{\mathbf{cf}}, \mathbf{I_{io}}, \hat{\mathbf{I}}_{\mathbf{io}}) = \alpha \cdot BCE\_IoU(\mathbf{I_{io}}, \mathbf{I_{io}}) + (1 - \alpha) \cdot BCE\_IoU(\mathbf{I_{cf}}, \hat{\mathbf{I}}_{\mathbf{cf}}) \tag{2}$$

After integrating these improvements, CALFIN-NN is trained for a total of 80 epochs, with 4000 batches per epoch, and 8 images per batch. Training is carried out on a K40 Nvidia Tesla GPU with 12GB of VRAM, with each epoch taking about 126 minutes to complete, and almost 1 week in total to obtain the optimal weights at epoch 65. Once trained, an NVIDIA GTX1060 with 6GB VRAM is used for the off-line data processing of the 20188 GeoTIFF subsets. The CALFIN algorithm takes about 3.5 days to process all of the subsets into calving fronts, excluding preprocessing, but including post-processing, as discussed in the following section.

3.3 Post-Processing

At this stage, the 2-channel pixel mask output of CALFIN-NN is post-processed to extract ~~useful~~ the Shapefile data products ~~as shown in~~ (Fig. 6).

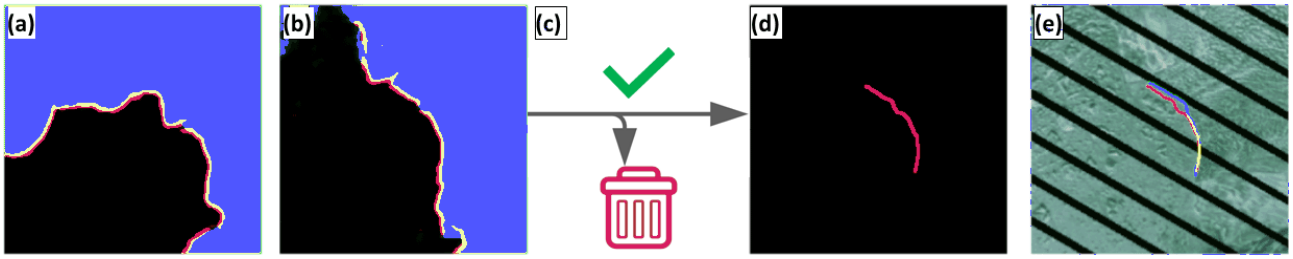

Figure 6. Postprocessing Pipeline: (a) First, get the processed image from CALFIN-NN. (b) Then, isolate and re-process each front. (c) Next, filter unconfident predictions. (d) Now, fit line and mask static coastline (see also Fig. 7). (e) Lastly, export and validate the Shapefile.

3.3.1 ~~Calving Front Reprocessing~~

~~We first isolate individual fronts from the processed image and reprocess subsets of the input image wherever they are detected. The front detection method is described in Sect. ??. We also exploit the nature of CALFIN-NN's output as a confidence measure, so that generated fronts can be filtered out based on classification confidence thresholds.~~

3.3.1 ~~Pixel Mask to Coastal Polyline~~

~~Next, we fit a polyline~~ First, a polyline is fit to the pixel mask to retrieve the correct coastline boundary. This is performed by converting each pixel in the mask to nodes in a graph, connecting the nearest neighboring nodes, then finding the single longest path in the graph's minimum spanning tree (MST) (Kruskal, 1956). This polyline not only corresponds with the coastline edge, but also out-performs other contour finding algorithms by eliminating noise, errors, and gaps inherited from previous steps. Such gaps are given weights based on the negative exponential distances between nodes, which allows for connections if the joined paths are significantly longer than the gap itself. A visual example is given in Fig. 7a-d. ~~For additional context, see Supplement Fig. S2.~~

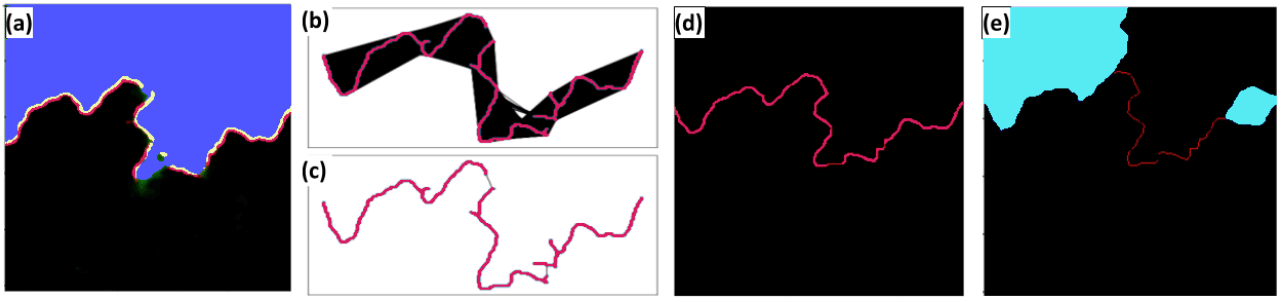

Figure 7. Mask to Polyline Algorithm: (a) First, extract the ~~red~~ coastline mask (red/yellow) from the CALFIN-NN output. (b) Then create a graph, connecting each pixel (~~blue~~red) to 15% of its nearest neighbors with an edge (black). (c) Next, create an MST from the graph. (d) Now, extract the longest path from the MST. (e) Finally, mask the static coastline using the fjord boundaries (~~blue~~cyan) to extract the calving front.

3.3.1 ~~Coastline to Calving Front~~

~~Next, we isolate~~ Next, the calving front is isolated from the coastline polyline. ~~We use fjord boundary masks~~ Static masks of the average fjord boundaries are manually created for each basin using the image subsets and BedMachine v3 for reference (Morlighem et al., 2017). By calculating the distance from each point in the coastline to the nearest fjord boundary pixel, then selecting the contiguous pixels which are the farthest from the fjord boundaries, the calving front can be isolated. The result of this is shown in Fig 7e.

3.3.1 ~~Calving Front to Shapefile~~

~~The~~ 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. After reprocessing, the nature of CALFIN-NN's 2-channel output as a confidence measure is exploited to filter out uncertain detections. 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.

The last step is to export the ~~polyline and~~ polylines and the corresponding polygon as geo-referenced Shapefiles. ~~We first smooth the polyline~~ First, the polylines are smoothed to eliminate noise artifacts inherited from previous steps~~. Next, we combine~~ , deviating no more than 1 pixel from the raw extracted coastline (see Supplement Fig. S2). Next, the smoothed polylines, fjord boundary mask, and land-ice/ocean masks are combined to create a polygonal ocean mask. Optionally, ~~we can manually verify~~ manual verification of each output with the original GeoTIFF subset can be performed. This was done for all cases in this study to ensure the validity of ~~our~~ the automated pipeline. This constrains the mean distance error to be <100 m, as covered in the following section.

4 ~~Error Analysis and Quality Assessment~~Validation

~~We use two methods~~ Two methods are used to evaluate CALFIN. For ~~our~~ the primary method, ~~we estimate the error~~ the error is estimated by calculating the Mean/Median Distance between predicted and manually delineated fronts (see Fig. 8a and Sect. 4.1). For ~~our~~ the secondary method, ~~we calculate~~ the classification accuracy is calculated with the Intersection over Union metric (see Fig. 8b and Sect. 4.2). Additionally, ~~we can evaluate~~ the detection accuracy ~~and provide~~ is evaluated, and the associated confusion matrix ~~(see~~ is provided (see Table 1 and Sect. 4.4). ~~We evaluate these metrics~~ These metrics are evaluated on several validation sets, taken from existing studies as discussed in Sect. ~~??~~1. These validation sets contain data that ~~is~~ are excluded during model training. This prevents the models from memorizing data and skewing the accuracy assessment.

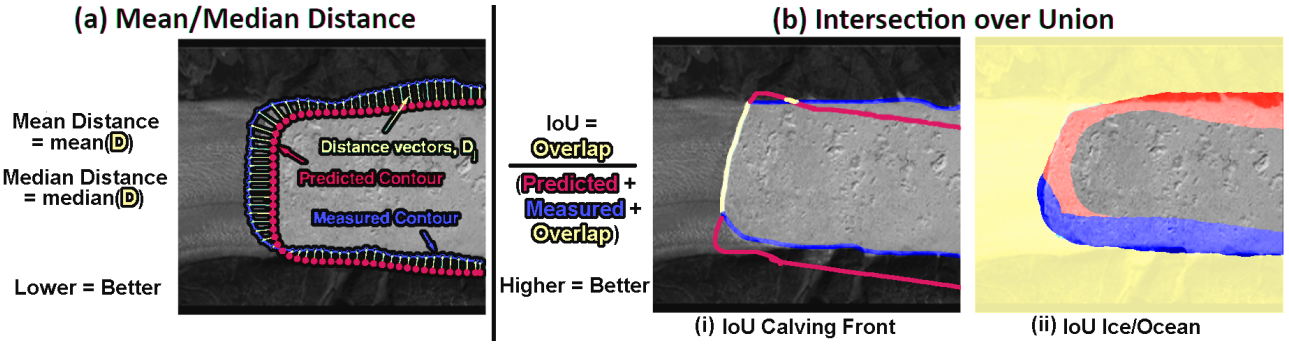

Figure 8. Error Measures: (a) A visual outline of Mean/Median Distance Error Estimation and (b) Classification Accuracy using Intersection over Union (IoU) for (i) the primary calving front, and (ii) the secondary ice/ocean mask, respectively.

4.1 Error Estimation

The primary quality assessment method is the Mean Distance Error ~~(Mohajerani et al., 2019; Zhang et al., 2019; Baumhoer et al., 2019)~~ (Mohajerani et al., 2019; Zhang et al., 2019; Baumhoer et al., 2019). Conceptually, this method resembles the numerical integration of the area between two curves, normalized by the average length of the curves (see Fig. 8a). Also referred to as the Area over Front (A/F) in literature, this method can also be seen as a generalization of the method of transects along arbitrarily oriented fronts (Mohajerani et al., 2019; Baumhoer et al., 2019). This metric is implemented by taking the mean/median of the distances between closest pixels in the predicted and manually delineated fronts. ~~We note~~ Note that pixel distance is biased to be inversely proportional to a network's input size, so ~~we also provide~~ the error in meters is also provided in the following analysis.

4.2 Classification Accuracy

The secondary quality assessment method calculates the Intersection over Union (IoU) (Baumhoer et al., 2019). This metric evaluates the degree of overlap between the predicted and ~~ground truth~~ manually delineated masks of the calving front. It is calculated by dividing the number of pixels in the intersection of two masks over the number of pixels in the union of the two masks (see Fig. 8b). When calculating the IoU of 3 pixel wide edges, this measure is very strict: 1 pixel of difference results in a score of 0.5000, and scores in that range or above are indicative of human levels of accuracy. When calculating the IoU of land-ice/~~ocean-mélange~~ ocean masks, this measure is less strict, and scores in the range of 0.9000 and above ~~are indicative of~~ indicate human levels of accuracy.

4.3 Validation Results

The following subsections ~~list tables that print~~ show tables with the above metrics for the associated validation sets, the values from the original studies, and a subset of the outputs of CALFIN-NN on each. ~~Our~~ The primary validation set, the CALFIN validation set (CALFIN-VS), consists of 162 images with clouds, illumination differences, ice mélange, and Landsat 7 Scan-

line Corrector Errors (L7SCEs). The CALFIN-VS contains data from 62 Greenlandic basins, including Helheim, which was specifically excluded from CALFIN's training set for validation purposes - as done by Mohajerani et al. (2019). The CALFIN-VS ensures CALFIN-NN produces consistent results on new data, addressing concerns raised by Zhang et al. (2019) Sect. 7.3. ~~We also evaluate the two validation subsets,~~ To evaluate performance on Landsat 7 Scanline Corrector Errors, the validation subset CALFIN-VS-L7-only ~~/none, which isolate and exclude~~ isolates images with L7SCEs, and the CALFIN-VS-L7-none excludes images with L7SCEs~~, respectively~~. To allow for comparisons between studies, ~~we also output~~ CALFIN-NN's performance metrics on previous studies' validation sets are also shown, where appropriate. The sets include the 10 Landsat Helheim subsets used in Mohajerani et al. (2019) (M-VS), the 6 TerraSAR-X Jakobshavn subsets used in Zhang et al. (2019) (Z-VS), and 62 Sentinel-1 Antarctic basins taken from the 11 validation scenes used in Baumhoer et al. (2019) (B-VS). Note that the error metrics are still sensitive to how each study implements them, which ~~we nevertheless reproduce and document to the best of our ability~~are nevertheless reproduced and documented for comparison's sake. These concerns are also addressed in the comprehensive inter-model comparison, discussed in Sect. 6.

4.3.1 ~~CALFIN Validation Set~~

CALFIN-NN performs well on the CALFIN-VS ~~. We calculate the~~ (Fig. 9). The true mean distance error of the CALFIN dataset ~~to be within~~ is calculated to be 86.76 ± 1.43 m with 95% confidence. When including only images with L7SCEs (CALFIN-VS-L7-only), the error ~~changes to 2.22 px (~~is 91.93 m~~)~~, showcasing CALFIN-NN's unique robustness to L7SCEs. ~~When excluding outliers such as Kong Oscar, the~~ Intuitively, excluding "difficult" images with L7SCEs in the validation set (CALFIN-VS-L7-none) decreases the error to 81.65 m. The median distance error is only ~~1.21 px (~~44.59 m~~)~~, showing that only a few outliers contribute considerably to the mean. For full outputs, see Supplement Fig. ~~S4-S7~~S5-S8.

Validation Set	Model	Mean Distance	Median Distance	IoU Calving Front	IoU Ice/Ocean
CALFIN-VS	CALFIN-NN	2.25 px, 86.76 m	1.21 px, 44.59 m	0.4884	0.9793
CALFIN-VS-L7-none	CALFIN-NN	2.27 px, 81.65 m	1.16 px, 44.01 m	0.4880	0.9819
CALFIN-VS-L7-only	CALFIN-NN	2.22 px, 91.93 m	1.33 px, 49.24 m	0.4888	0.9766

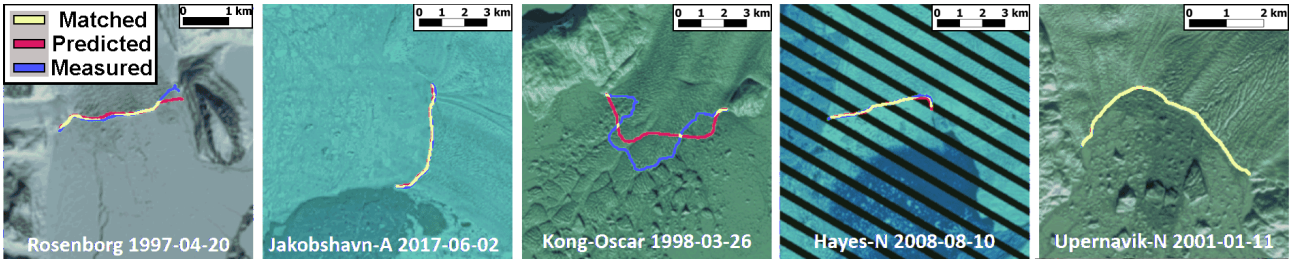

Figure 9. CALFIN-VS Validation Output Results: Yellow represents human (green) and machine (red) agreement on the front location. Note that the drop in mean pixel distance despite the increase in mean meter distance (and vice versa) comes from L7SCE images being reprocessed at lower sizes due to detection failures (see Fig. 6c), and pixel error bias being inversely related to input size (see Sect. 4.1).

4.3.1 ~~Mohajerani et al. (2019) Validation Set~~

CALFIN-NN performs well on the M-VS ~~. This demonstrates the generalization capability of~~ (Fig. 10). This demonstrates CALFIN-NN~~, which improves~~ 's ability to accurately process new data, which builds upon the Mohajerani et al. (2019) neural network (M-NN). ~~Also note~~ Note that M-NN implements distances errors differently, and omits ice/ocean masks from the evaluation. This ~~is~~ differences are further explored in the ~~model inter-comparison discussed in~~ Sect. 6 model inter-comparison.

Validation Set	Model	Mean Distance	Median Distance	IoU Calving Front	IoU Ice/Ocean
M-VS	CALFIN-NN	2.56 px, 97.72 m	2.55 px, 97.44 m	0.3332	N/A
M-VS	M-NN	1.97 px, 96.31 m	N/A	N/A	N/A

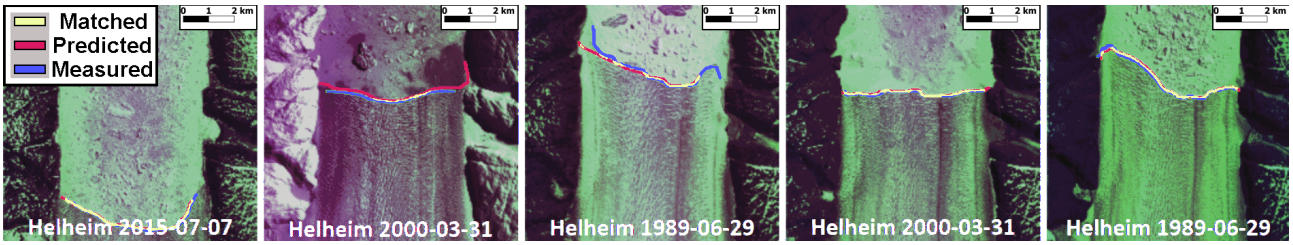

Figure 10. M-VS Validation Output Results: Note that CALFIN-NN has never trained on Helheim, but can still predict the front ~~at multiple scales and~~ under different conditions and preprocessing methods. See Fig. ~~S8~~S9. for full outputs.

4.3.1 ~~Zhang et al. (2019) Validation Set~~

~~CALFIN~~ CALFIN-NN performs competitively on the Z-VS ~~. We achieve~~ (Fig. 11). It achieves a similar mean meter distance (115.24 m vs. 104 m) despite being constrained to using lower resolution TerraSAR-X data. Note though that the Zhang et al. (2019) neural network (Z-NN) uses higher resolution input data (960×720) compared to CALFIN-NN (224x224), which skews the mean pixel distance comparison, where CALFIN-NN performs better (2.11 px vs. 17.3 px). Another source of skew comes from CALFIN-NN confidence filtering, as only 8 of 12 fronts in the set are confidently detected (see Sect. 4.4). ~~We suspect that increasing~~ Increasing CALFIN-NN's input resolution and training on higher resolution SAR data ~~will~~ may enable CALFIN-NN to detect more fronts with greater accuracy.

Validation Set	Model	Mean Distance	Median Distance	IoU Calving Front	IoU Ice/Ocean
Z-VS	CALFIN-NN	2.11 px, 115.24 m	1.65 px, 77.29 m	0.3832	0.9761
Z-VS	Z-NN	17.3 px, 104 m	N/A	N/A	N/A

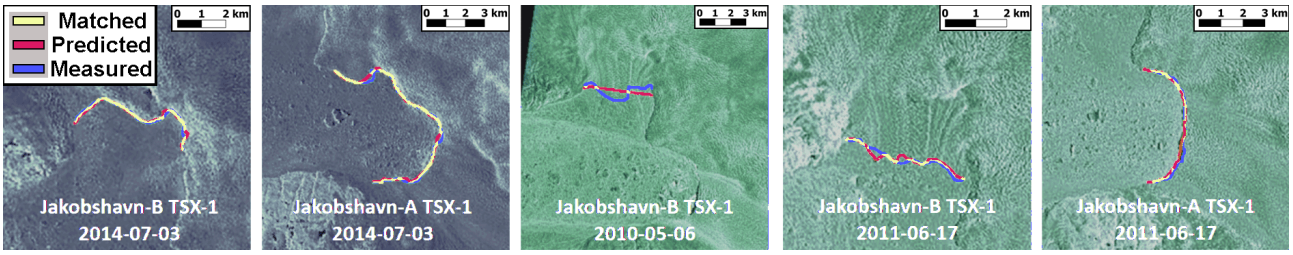

Figure 11. Z-VS Validation Output Results: CALFIN-NN works well on SAR data in addition to optical data. See Fig. ~~S9~~S10. for full outputs.

4.3.1 ~~Baumhoer et al. (2019) Validation Set~~

~~CALFIN~~ CALFIN-NN performs sub-par on the B-VS ~~. However, when we compare our~~ (Fig. 12). When comparing the mean distance error ~~estimate~~ with the Baumhoer et al. (2019) equivalent Area over Front (A/F) error, ~~we mask out~~ 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 ~~that could raise our relative error~~are masked out, raising the relative error, and negatively impacting CALFIN-NN's performance on this metric. When comparing metrics that isolate the calving front, ~~we calculate~~ the absolute median distance error is calculated (achieving 112.75 m) whereas ~~the~~ Baumhoer et al. (2019) uses signed median distance error (achieving 0 m), which is not ~~applicable~~ directly comparable in this context, and thus omitted. Currently, ~~our~~ the error is affected by ~~kilometer range~~ kilometer-range deviations in very large domains like Voyeykov Ice Shelf, and differences in sea-ice mélange as seen along the Gillet and Wordie Ice Shelves, which would be consistent with findings in Baumhoer et al. (2019) Sect. 5.2. After excluding such outliers, ~~we detect fronts~~ fronts are detected in 55 out of 62 domains (88.71%), ~~and achieve~~ achieving median distance errors of 0.95 px (127.87 m). Intensive retraining on ice shelves may be required for CALFIN-NN to improve.

Validation Set	Model	Mean Distance	Median Distance	IoU Calving Front	IoU Ice/Ocean
B-VS	CALFIN-NN	2.35 px, 330.63 m	0.74 px, 112.75 m	0.6451	0.9879
B-VS	B-NN	2.69 px, 108 m	N/A	N/A	0.905

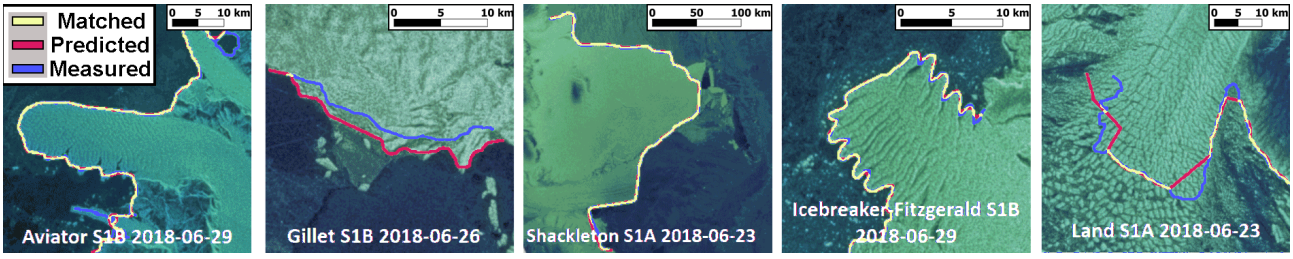

Figure 12. B-VS Validation Output Results: Similar to Z-NN, B-NN uses a high resolution input (768×768) relative to CALFIN-NN (224x224), which skews the mean pixel distance comparison ~~undeservedly~~ in CALFIN-NN's favor. See Fig. ~~S10-S11.~~ S11-S12 for full outputs.

4.4 Detection Accuracy

Lastly, ~~we show that~~ CALFIN-NN ~~has the ability~~ is shown to automatically filter images that do not have detectable calving fronts. To verify this, ~~we include~~ 13 images are included in the CALFIN-VS which do not contain calving fronts discernible to the human eye. ~~We calculate the~~ The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates are computed for the entire 162 image CALFIN-VS, and ~~output the confusion matrix below~~the associated confusion matrix is shown in Table 1. Note that CALFIN-NN does not output any false positives on the CALFIN-VS. While this ensures ~~we output accurate fronts rather than output~~ accurate fronts are output rather than incorrect fronts, this filtering behavior removes potentially large errors, and must be accounted for when comparing errors across other sets.

Table 1. Confusion Matrix: CALFIN-NN misses fronts in 8 of 149 valid CALFIN-VS images, but ~~we deem~~ this is deemed as an acceptable ~~tradeoff~~trade-off.

		Front Detected?	
		Yes	No
Front Detectable?	Yes	TP = 141/149 = 94.63%	FN = 8/149 = 5.76%
	No	FP = 0/13 = 0.00%	TN = 13/13 = 100.00%

5 ~~Data Product~~ Results and Discussion

~~We release the~~ The code implementation of the CALFIN method is released, along with its associated calving front data products as described in the following ~~subsections~~section, for use within the scientific community.

5.1 ~~CALFIN Dataset~~

~~**CALFIN Dataset Samples**: Data products for Upernavik (left), Jakobshavn (center), and Helheim (right), from 1972-2019.~~

~~We release the CALFIN dataset , which~~ The CALFIN dataset spans 66 Greenlandic ~~glaciers~~basins, over the period Sept. 1972 - June 2019. ~~This~~ It consists of over 1500 manual delineations and ~~20,004~~ 22,678 total calving fronts. ~~We provide 2~~ Two levels of CALFIN data products ~~.~~ are provided. The Level 0 products include the ~~raw inputs and basic outputs used for CALFIN-NN. These products consist of the raw GeoTIFF domain subsets, the domain Shapefiles used for subsetting, neural network pixel mask outputs, and a quality assurance image~~ Shapefile domains used for subsetting, the neural network training image/mask pairs, the fjord boundary masks, the full Landsat scene ID list, and the quality assurance images for validation purposes. ~~Use~~ The use cases of Level 0 products may include studies of reproducibility, validation, or training new neural networks. The Level 1 products ~~includes~~ include the calving front polyline and polygon Shapefiles. The polyline product consists of the isolated, refined, geo-referenced, and verified calving fronts for each domain. The polygon product consists of an ocean mask bounded by the domain subset, the fjord boundaries, and the calving front(s), for each domain. ~~Shapefiles are projected to EPSG:3413. These data products can currently be found at .~~

5.1 CALFIN-NN Implementation

We release an implementation of CALFIN-NN, available at , which includes the parameters and architecture we develop throughout this study. It is our intention that any innovations as described in Sect. 3.2 can be applied to other networks and investigations. The implementation is written in Python 3 using the Keras & Tensorflow libraries. Note that access to the network parameters are also hosted as part of the associated DataDryad dataset linked above (Cheng et al., 2020). For additional insight into the network training and processing requirements, see the following discussion in Sect. **??**.

6 Discussion

5.1 Training Insights

Throughout the course of the study, we develop several innovations to improve the performance of CALFIN-NN. To increase accuracy, we utilize a special loss function that heavily favors correct calving front predictions. To prevent over-fitting our neural network, a large set of training data was manually delineated (see Fig. S3) , totalling 1541 Landsat and 232 Antarctic SAR image/mask pairs, with the SAR data taken from the same training scenes used by Baumhoer et al. (2019). Another measure to prevent over-fitting involves data augmentation, which entails performing random flips/transpositions, random Gaussian noise, random sharpen filters, random rotations of up to 12°, random crops, and random scaling on the pre-processed images during CALFIN-NN training. We determine through empirical testing that excessive image padding, rotation, warping, and cropping calving fronts to close to the image bounds result in sub-optimal performance. Yet another helpful technique is the use of test-time augmentations. More specifically, each image subset is cut into 9 overlapping 224x224 image windows and processed individually, before being reassembled into the final 256x256 output mask. This allows for multiple independent classifications of the central pixels, ensuring agreement and confidence in detected calving fronts.

After integrating these improvements, CALFIN-NN is trained for a total of 80 epochs, with 4000 batches per epoch, and 8 images per batch. Training is carried out on a K40 Nvidia Tesla GPU with 12GB of VRAM, with each epoch taking about 126 minutes to complete, and almost 1 week in total to obtain the optimal weights at epoch 65. Once trained, we used an NVIDIA GTX1080 with 6GB VRAM for off-line data processing. Our algorithm (excluding preprocessing, but including post-processing) is capable of handling about 4 subsets per minute, taking about 3.5 days to process all 20188 GeoTIFF subsets into calving fronts. Future investigations should thus consider the trade-offs between the processing time of large networks, their accuracy, and the required computational resources, as existing works may prove suitable for their needs.

5.1 Existing Works

Mohajerani et al. (2019) is an example of an existing study that uses deep neural networks to detect calving fronts. The study pioneers the UNet-style network for application towards the Greenlandic glacial basins Jakobshavn, Helheim, Sverdrup, and Kangerlussuaq. While the methodology is restricted by its preprocessing requirements and inability to handle branching/non-linear calving fronts, it nonetheless supports the viability of the neural network method. Zhang et al. (2019) and Baumhoer et al. (2019)

evaluate modified UNet architectures, as applied to SAR data in Jakobshavn and Antarctica, respectively. Their studies incorporate large spatial context in order to capture high resolution and potentially whole-coastline delineations. These larger networks, on the order of (960x720 Both of the Shapefiles share a common metadata feature schema (see Table S2) derived from the MEaSUREs Glacial Termini Dataset (Moon and Joughin, 2008; Joughin et al., 2015), and 768x768 pixels respectively), support the viability of both training and applying large networks to new data. The CALFIN-NN method builds on these studies by improving on the network design, capability, and post-processing methods. The following section shows a data analysis example as performed in Zhang et al. (2019), and similarly showcases a possible application of a calving front dataset in advancing our understanding of the Greenland Ice Sheet. names are derived from Bjørk et al. (2015). These products can be found via these links to Github and DataDryad (Cheng et al., 2020).

5.1 Data Analysis and Usage Example

With the new data available to use in the CALFIN dataset, we explore a subset and validate the evolution of Helheim Glacier it is possible to explore seasonal trends across the Greenland Ice Sheet, and validate a subset of 10 high discharge basins of interest against existing ESA-CCI, MEaSUREs, and PROMICE data products (ENVEO, 2017; Joughin et al., 2015; Andersen et al., 2019). Similar to Zhang et al. (2019), we graph the relative change in position of the calving front along the fjord centerline from 1972 to June 2019. Fig 13 shows the high temporal resolution and spatial accuracy of the CALFIN data product alongside corresponding available data products from 1972-2019. For Joughin et al. (2015), if a date range is given, we plot the same relative change at both start and end dates (Moon and Joughin, 2008) is plotted. For Andersen et al. (2019), we use August 15th is used as the "end-of-melt-season" date of delineation, as the date is otherwise not specified in the provided data. Fig. ?? shows the length change The advance and retreat of the calving front along the basin centerline, relative to its Sept. 6, 1972 position. centerlines is relative to their earliest positions. Note the large improvement in temporal/seasonal coverage and the general agreement of CALFIN with existing data products. Note also that the discrepancies such as that during 2005-2009 in Jakobshavn (Fig. 13e) mostly stem from a lack of winter coverage during Landsat's optical blackout period. Additional outliers in Kong Oscar (Fig. 13g) stem from the somewhat arbitrary delineation of the ice tongue front position. Kangiata Nunaata (Fig. 13j) suffers from both of the aforementioned effects, but otherwise shows the same general agreement with existing datasets from 2000 onwards.

Overall, there is high agreement between CALFIN and existing data products on the evolution of Helheim over the available time series. Note that while Helheim is relatively easy to accurately and automatically delineate, we still produce all of the above data without manual input outside of visual verification. Thus with this context in mind, we find that this comparison with existing data products help validate the applicability of this study's outputsAdditionally, Fig. 14 shows the regional mean advance and retreat change, alongside the mean for the entirety of Greenland covered by the CALFIN dataset. Contributions from NW Greenland influence the overall trend the most, due to the presence of many small glaciers/branches in the region. Note that the mean for Greenland also includes contributions from Petermann, which is visible in the summers of 2010 and 2012. Shared regional trends are visible across NW and CW Greenland, which both show relative stability before 2000, followed by steady retreat up until 2017-2018. CE and SE Greenland also share a similar but less pronounced retreat, showing

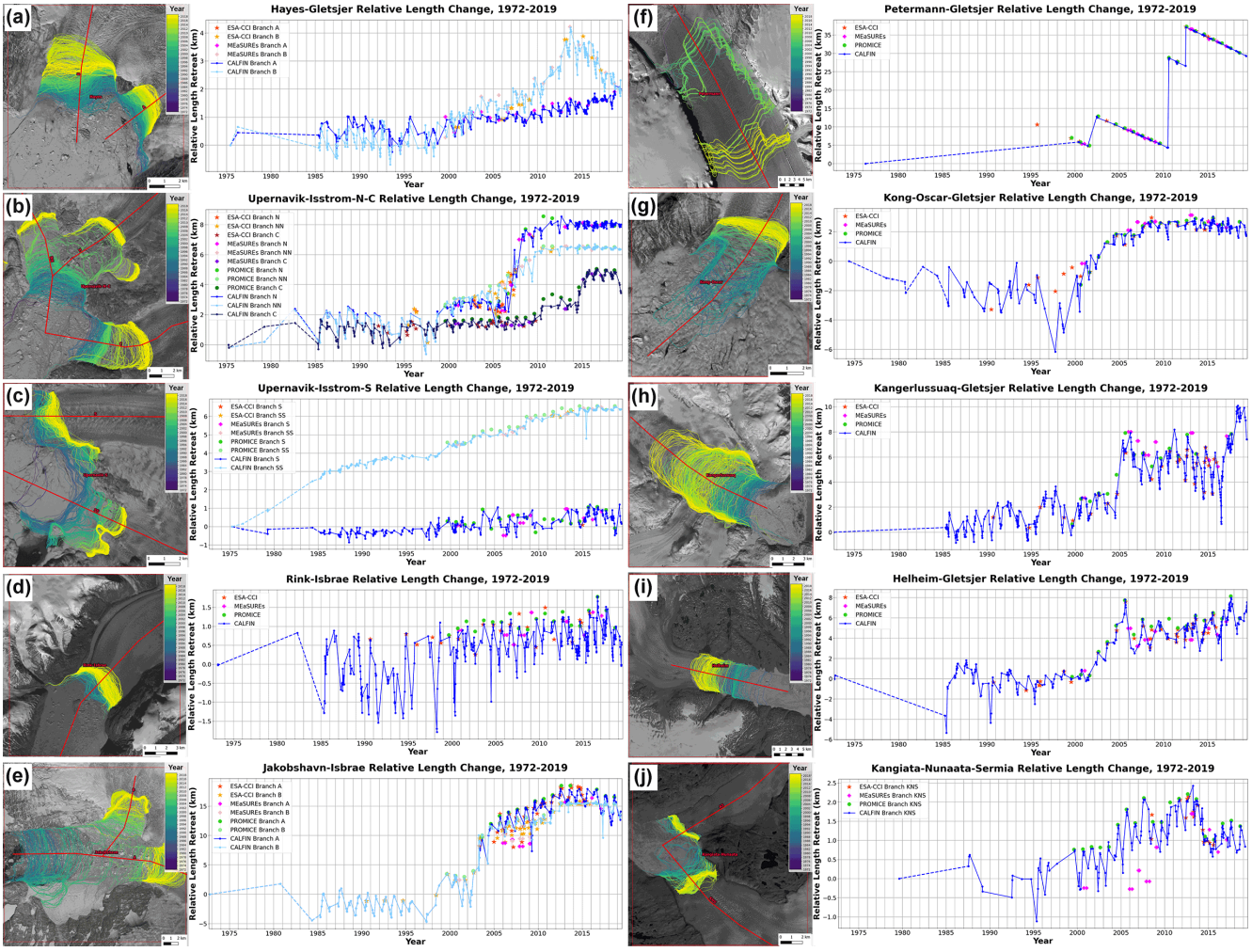

Figure 13. Terminus Advance and Retreat Over Time. (a-j) Basin setup (left) and graph (right) for 10 high discharge basins. Positive length change represents retreat relative to the earliest position along the centerlines in red. Note the seasonal variations captured by CALFIN, in blue. Time series for other studies span 1990-2016 (ESA-CCI), 2000-2017 (MEaSUREs), and 1999-2019 (PROMICE). Note the seasonal variations shown by the solid lines, and the dotted lines from 1972-1985 that indicate a lack of such seasonal observations. Also note that the vertical axis scaling is applied differently for each graph to highlight seasonal trends.

an accelerating retreat beginning around 1995. These regional trends are less visible in SW Greenland, which is dominated by Narsap Sermia's retreat from 2010-2013. Overall, these regional trends generally agree with studies such as Wood et al. (2021) and King et al. (2020), helping further validate the CALFIN method and data.

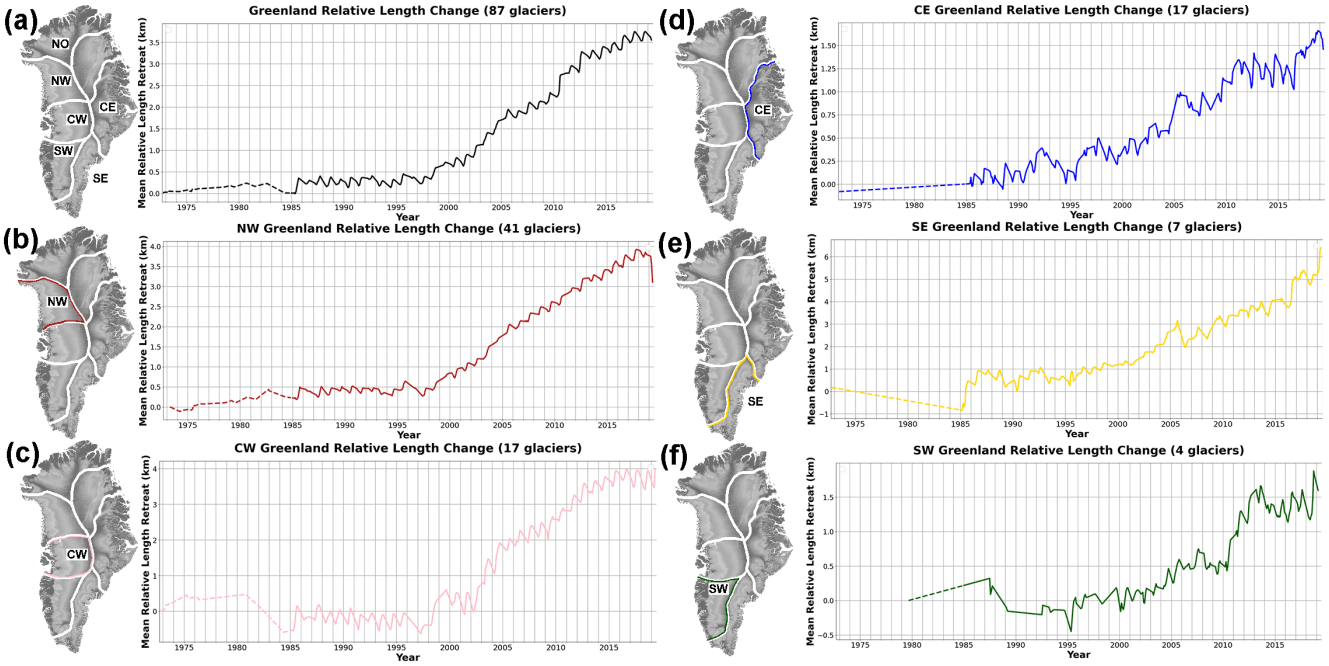

Figure 14. ~~Helheim Terminus Length Change Over Time~~**Regional Terminus Advance and Retreat Over Time**. ~~Positive length change represents retreat relative to the 1972~~ (a-f) Regional delineations (left) and terminus position ~~. Seasonal variations are captured by CALFIN~~ graphs (~~blue~~right) ~~. Time series~~ for ~~other studies span 1995-2016~~ Greenland (~~ESA-CCI~~a) and the northwestern (b), ~~2005-2017~~ central western (~~MEaSUREs~~c), central eastern (d), southeastern (e), and ~~1999-2019~~ southwestern (~~PROMICE~~f) regions. ~~See Fig~~Note that the total Greenland mean advance and retreat is unadjusted, and dominated by the trend lines of numerous smaller glaciers in CW and NW Greenland. ~~S12~~ Note that branches in the 66 studied basins are independently counted, for ~~an enlarged CALFIN/ESA-CCI 1995-2016 comparison~~a total of 87 glaciers.

5.1 ~~Inter-model Comparison~~

6 Inter-model Comparison

To ~~similarly~~ further reinforce the validity of ~~our~~ the study, and address the shortcomings of different error metric comparisons (as discussed in Sect. 4.3), ~~we conduct~~ a comprehensive inter-model comparison is conducted between CALFIN-NN and the model developed by Mohajerani et al. (2019) (M-NN). This experiment seeks to understand how both models perform, holding all other variables constant. In particular, ~~we want to understand~~ this experiment seeks to determine if the M-NN model, and by extension other UNet models, perform on par with the CALFIN-NN model, given the same training data. ~~To perform this task , we retrain~~ This task involves retraining the M-NN ~~using~~ on CALFIN training data, ~~process validation data, and compare the results~~and comparing its performance against CALFIN-NN using a shared validation set. For the fairest results, ~~we evaluate only images with~~ only images without L7SCEs ~~, which~~ are evaluated in this validation set - CALFIN-VS-L7-none - which is

within the known capabilities of the ~~M-NN is already known to be capable of handling~~M-NN. Furthermore, the same pre- and post-processing is applied to both models.

Table 2. Model Inter-comparison Error Table: Metrics for the CALFIN-NN and M-NN models on all non-Landsat 7 test images in the CALFIN validation set.

Validation Set	Training Set	Model	Mean Distance	Median Distance	IoU Front	IoU Ice/Ocean
CALFIN-VS-L7-none	CALFIN	CALFIN-NN	2.27 px, 81.65 m	1.16 px, 44.01 m	0.4880	0.9819
CALFIN-VS-L7-none	CALFIN	M-NN	4.45 px, 201.35 m	1.25 px, 50.52 m	0.4935	0.9699

Across all non-Landsat 7 test images in the CALFIN validation set, CALFIN-NN attains a 2.27 pixel (81.65 meter) mean distance between the predicted and the ~~ground truth~~ manually delineated fronts. This exceeds the level of accuracy achieved by the model from Mohajerani et al. (2019), which after retraining on CALFIN training data, is 4.45 pixels (201.35 meters). Note again that Landsat 7 images were excluded during reevaluation for the M-NN. This supports ~~our~~ the findings that the CALFIN-NN architecture is an improvement over existing UNet models.

With this added context, ~~we reproduce~~ the validation table is reproduced from Sect. ~~??~~4.3, Fig. 10, and ~~continue~~ the error analysis is continued below. To reemphasize the differences in mean distance error calculation between different studies, Mohajerani et al. (2019) begins by breaking each ~~delineated~~ predicted front to 1000 smaller segments within a small buffer from the fjord walls and calculating the mean deviation between the segments of the ~~true and~~ predicted and manually delineated fronts. ~~Our~~ The method begins by averaging the mean distance between each pixel of the ~~delineated~~ predicted front and the closest pixel of the ~~true~~ manually delineated front as detailed in Sect 4.1. While the line-segment methodology of Mohajerani et al. (2019) provides a stricter estimate by enforcing close agreement between corresponding front segments, ~~our~~ the CALFIN method allows for non-aligned evaluation of the mean distance error. Although both implementations quantify the differences between the lines, the differences in implementation should still be considered when evaluating the comparison below.

Table 3. ~~M-VS Validation Output Results~~**M-VS Validation Output Results**: Accuracy and error metrics for the CALFIN-NN and the M-NN models on the M-VS. Again, some metrics are not provided by Mohajerani et al. (2019), so they are omitted from this table.

Validation Set	Training Set	Model	Mean Distance	Median Distance	IoU Front	IoU Ice/Ocean
M-VS	CALFIN	CALFIN-NN	2.56 px, 97.72 m	2.55 px, 97.44 m	0.3332	N/A
M-VS	Mohajerani	M-NN	1.97 px, 96.31 m	N/A	N/A	N/A

Across all 10 test images in the M-VS, CALFIN-NN attains a 2.56 pixel (97.72 meter) mean distance between the predicted and the ~~ground truth~~ manually delineated fronts. This approaches the level of accuracy achieved in the original study, which is 1.97 pixels (96.31 meters). This supports ~~our~~ the findings that the CALFIN-NN architecture generalizes to new data well. Note that CALFIN-NN's larger network size requires additional training data to avoid ~~overfitting,~~ over-fitting, or memorizing, the training data, which could explain the slightly lesser accuracy when compared to the M-NN. In summary, this comprehensive

model inter-comparison supports the hypothesis that the CALFIN-NN model improves on existing studies and is generalizing well.

7 Conclusion

Overall, ~~we accomplish our~~ the goal of automatically delineating calving fronts from satellite imagery ~~. Our~~ is accomplished. The CALFIN method uses the cutting-edge in deep learning architectures, allowing for robustness to minor cloud cover, Landsat 7 Scanline Corrector Errors, and illumination changes. Future work may entail accuracy improvements, expansion of included domains, usage of SAR data sources, and near-real time data products. Within the community, ~~we anticipate the benefit~~ the benefits of standardized training, validation sets, and outputs/metadata ~~. We also anticipate the~~ are anticipated. The community's development of new automated extraction studies, such as grounding line delineation, iceberg tracking, and sea ice mélange measurements~~. Our~~ , is also anticipated. A key takeaway is the maturation of neural networks for automated calving front detection. Specifically, a well trained network now approaches human levels of accuracy in picking arbitrary glacial calving fronts. This reinforces existing studies on the viability of the methodology, and paves the way for applications on other data processing tasks. Ultimately, this work showcases the state-of-the-art in automated calving front detection, and provides a new database of glacial termini positions for the cryosphere community.

Code and data availability. The code used to automate the implement the CALFIN pipeline is freely available at github.com/daniel-cheng/CALFIN. It is written in Python 3, using the Keras & Tensorflow libraries. The data generated by CALFIN is currently available at datadryad.org/stash/share/Q9guqsr

The Supplement related to this article is available online at: https://doi.org/10.5194/tc-0-1-2021-supplement

Author contributions. DC developed the code/model, created the training data, carried out the data processing/error analysis, and wrote the majority of the manuscript. WH provided input on the processing methodology, post-processing algorithms, error analysis, discussion topics, and writing the manuscript. EL provided key direction for the overall study, error analysis, outputs, and writing the manuscript. YM performed the model inter-comparison and assisted with the writing of the manuscript. MW performed the data preprocessing for the model inter-comparison. IV assisted in organizing collaborators and the model inter-comparison. ER contributed suggestions regarding the error analysis and inter-comparison. WH, EL, MW, and YM revised the manuscript and results.

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was conducted as a collaboration between NASA's Jet Propulsion Laboratory and the University of California, Irvine. ~~Our Python implementation of Deeplabv3+ Xception is based on~~ The CALFIN neural network architecture implementation is derived from Emil Zakirov's ~~code base (,~~ Deeplabv3+ Xception codebase at github.com/bonlime/keras-deeplab-v3-plus (last access: 13 August 2020). We acknowledge the USGS for providing Landsat-1-8 images, the ESA for their Sentinel-1 images, as well as the ESA-CCI, PROMICE, and MEaSUREs programs for providing calving front data used in this study. Additionally, we thank the editors and reviewers for their contributions to the improvement of this manuscript.

References

Andersen, J. K., Fausto, R. S., Hansen, K., Box, J. E., Andersen, S. B., Ahlstrøm, A. P., Dirk, Citterio, M., Colgan, W., Karlsson, N. B., and et al.: Update of annual calving front lines for 47 marine terminating outlet glaciers in Greenland (1999–2018), GEUS Bulletin, 43, https://doi.org/10.34194/GEUSB-201943-02-02, 2019.

Andersen, M., Stenseng, L., Skourup, H., Colgan, W., Khan, S., Kristensen, S., Andersen, S., Box, J., Ahlstrøm, A., Fettweis, X., and Forsberg, R.: Basin-scale partitioning of Greenland ice sheet mass balance components (2007–2011), Earth and Planetary Science Letters, 409, 89 – 95, https://doi.org/10.1016/j.epsl.2014.10.015, 2015.

Baumhoer, C. A., Dietz, A. J., Kneisel, C., and Kuenzer, C.: Automated Extraction of Antarctic Glacier and Ice Shelf Fronts from Sentinel-1 Imagery Using Deep Learning, Remote Sensing, 11, 2529, https://doi.org/10.3390/rs11212529, 2019.

Bjørk, A. A., Kruse, L. M., and Michaelsen, P. B.: Brief communication: Getting Greenland's glaciers right – a new data set of all official Greenlandic glacier names, The Cryosphere, 9, 2215–2218, https://doi.org/10.5194/tc-9-2215-2015, 2015.

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, Journal of Glaciology, 64, 523–535, https://doi.org/10.1017/jog.2018.44, 2018.

Carr, J. R., Stokes, C. R., and Vieli, A.: Threefold increase in marine-terminating outlet glacier retreat rates across the Atlantic Arctic: 1992–2010, Annals of Glaciology, 58, 72–91, https://doi.org/10.1017/aog.2017.3, 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, Journal of Geophysical Research: Earth Surface, 123, 2024–2038, https://doi.org/10.1029/2017JF004499, 2018.

Catania, G. A., Stearns, L. A., Moon, T. A., Enderlin, E. M., and Jackson, R. H.: Future Evolution of Greenland's Marine-Terminating Outlet Glaciers, Journal of Geophysical Research: Earth Surface, 125, e2018JF004 873, https://doi.org/10.1029/2018JF004873, 2020.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, European Conference on Computer Vision, pp. 801–818, http://arxiv.org/abs/1802.02611, 2018.

Cheng, D., Hayes, W., and Larour, E.: CALFIN: Calving Front Dataset for East/West Greenland, 1972-2019, https://doi.org/10.7280/D1FH5D, 2020.

Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions, Computer Vision and Pattern Recognition, pp. 1800–1807, https://doi.org/10.1109/cvpr.2017.195, 2017.

ENVEO: Greenland Calving Front Dataset, 1990 - 2016, v3.0, http://products.esa-icesheets-cci.org/products/downloadlist/CFL/, 2017.

Fürst, J. J., Goelzer, H., and Huybrechts, P.: Ice-dynamic projections of the Greenland ice sheet in response to atmospheric and oceanic warming, The Cryosphere, 9, 1039–1062, https://doi.org/10.5194/tc-9-1039-2015, 2015.

Joughin, I., Moon, T., Joughin, J., and Black, T.: MEaSUREs Annual Greenland Outlet Glacier Terminus Positions from SAR Mosaics, Version 1, https://doi.org/10.5067/DC0MLBOCL3EL, 2015.

King, M. D., Howat, I. M., Jeong, S., Noh, M. J., Wouters, B., Noël, B., and van den Broeke, M. R.: Seasonal to decadal variability in ice discharge from the Greenland Ice Sheet, The Cryosphere, 12, 3813–3825, https://doi.org/10.5194/tc-12-3813-2018, 2018.

King, M. D., Howat, I. M., Candela, S. G., Noh, M. J., Jeong, S., Noël, B. P. Y., Broeke, M. R. v. d., Wouters, B., and Negrete, A.: Dynamic ice loss from the Greenland Ice Sheet driven by sustained glacier retreat, Nature News, https://doi.org/10.1038/s43247-020-0001-2, 2020.

Kruskal, J. B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, Proceedings of the American Mathematical Society, 7, 48–50, https://doi.org/10.2307/2033241, 1956.

Malik, J., Belongie, S., Leung, T., and Shi, J.: Contour and Texture Analysis for Image Segmentation, Int. J. Comput. Vision, 43, 7–27, https://doi.org/10.1023/A:1011174803800, 2001.

Mannor, S., Peleg, D., and Rubinstein, R.: The Cross Entropy Method for Classification, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, p. 561–568, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/1102351.1102422, 2005.

Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study, Remote Sensing, 11, https://doi.org/10.3390/rs11010074, 2019.

Moon, T. and Joughin, I.: Changes in ice front position on Greenland's outlet glaciers from 1992 to 2007, Journal of Geophysical Research: Earth Surface, 113, https://doi.org/10.1029/2007JF000927, 2008.

Morlighem, M., Williams, C. N., Rignot, E., An, L., Arndt, J. E., Bamber, J. L., Catania, G., Chauché, N., Dowdeswell, J. A., Dorschel, B., Fenty, I., Hogan, K., Howat, I., Hubbard, A., Jakobsson, M., Jordan, T. M., Kjeldsen, K. K., Millan, R., Mayer, L., Mouginot, J., Noël, B. P. Y., O'Cofaigh, C., Palmer, S., Rysgaard, S., Seroussi, H., Siegert, M. J., Slabon, P., Straneo, F., van den Broeke, M. R., Weinrebe, W., Wood, M., and Zinglersen, K. B.: BedMachine v3: Complete Bed Topography and Ocean Bathymetry Mapping of Greenland From Multibeam Echo Sounding Combined With Mass Conservation, Geophysical Research Letters, 44, 11,051–11,061, https://doi.org/10.1002/2017GL074954, 2017.

Nagler, T., Rott, H., Hetzenecker, M., Wuite, J., and Potin, P.: The Sentinel-1 Mission: New Opportunities for Ice Sheet Observations, Remote Sensing, 7, 9371–9389, https://doi.org/10.3390/rs70709371, 2015.

Nick, F. M., Vieli, A., Andersen, M. L., Joughin, I., Payne, A., Edwards, T. L., Pattyn, F., and van de Wal, R. S. W.: Future sea-level rise from Greenland's main outlet glaciers in a warming climate, Nature, 497, 235–238, https://doi.org/10.1038/nature12068, 2013.

Paravolidakis, V., Moirogiorgou, K., Ragia, L., Zervakis, M., and Synolakis, C.: COASTLINE EXTRACTION FROM AERIAL IMAGES BASED ON EDGE DETECTION, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, III-8, 153–158, https://doi.org/10.5194/isprsannals-III-8-153-2016, 2016.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, CoRR, abs/1505.04597, http://arxiv.org/abs/1505.04597, 2015.

Seale, A., Christoffersen, P., Mugford, R. I., and O'Leary, M.: Ocean forcing of the Greenland Ice Sheet: Calving fronts and patterns of retreat identified by automatic satellite monitoring of eastern outlet glaciers, Journal of Geophysical Research: Earth Surface, 116, https://doi.org/10.1029/2010JF001847, 2011.

van den Broeke, M. R., Enderlin, E. M., Howat, I. M., Kuipers Munneke, P., Noël, B. P. Y., van de Berg, W. J., van Meijgaard, E., and Wouters, B.: On the recent contribution of the Greenland ice sheet to sea level change, The Cryosphere, 10, 1933–1946, https://doi.org/10.5194/tc-10-1933-2016, 2016.

Wood, M., Rignot, E., Fenty, I., An, L., Bjørk, A., van den Broeke, M., Cai, C., Kane, E., Menemenlis, D., Millan, R., Morlighem, M., Mouginot, J., Noël, B., Scheuchl, B., Velicogna, I., Willis, J. K., and Zhang, H.: Ocean forcing drives glacier retreat in Greenland, Science Advances, 7, https://doi.org/10.1126/sciadv.aba7282, 2021.

Zhang, E., Liu, L., and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multi-temporal TerraSAR-X images: a deep learning approach, The Cryosphere, 2019, 1–20, https://doi.org/10.5194/tc-2019-14, 2019.