**General Comments:**

Presented in this manuscript is an automated data processing pipeline for extracting glacier termini positions, and the associated dataset that consists of data spanning 295 Greenlandic glaciers over period 1984-2021. The dataset consists of 278,239 glacier termini for 295 glaciers, and includes ice/ocean masks for the years 2018-2020. The pipeline consists of a Google Earth Engine based downloader , combined with a deep neural network to extract termini locations from the subsetted and preprocessed satellite imagery. The literature review covers most of the existing work in the field. The deep learning methodology also incorporates the greatest diversity of sensors (Landsat 5-8, Sentinel 1 & 2) and sensor types (both optical and SAR), which is a novel development. The methodology is quality controlled by assessing its performance on two uncertainty quantification metrics.

In summary, the study represents a significant contribution to the cryosphere and scientific community, by providing a new glacial termini dataset for Greenland, and an automated deep learning based pipeline for automated glacial feature extraction. However, there are certain comments to be addressed regarding the dataset and the manuscript before acceptance at the editor's discretion, as detailed below.

We greatly appreciate the detailed review and constructive comments by Reviewer 2. We have made our best effort to revise the manuscript based on the referee's comments and suggestions. In the following, we made an item-by-item response to the specific comments by the referee.

**Major Comments:**

- A primary concern to be noted is the lack of certain validation metrics that are commonly used in works such as this. Previous studies use the same established validation metrics (average area/distance between predicted and observed termini, or Mean distance error) to ensure ease of comparison. This measure is used in existing works such as Mohajerani et al. (2019), Baumhoer et al. (2019), Cheng et al. (2021), Heidler et al. (2021), Gourmelon et al. (2022), Loebel et al. (2022), and specifically Zhang et al. (2019, 2021). The average uncertainty of 37m, which is calculated using the average distance between duplicate picks from Landsat-8 and Sentinel-2, is somewhat misleading given this context, and the lack of such mean distance error calculation with respect to the ground truth should be addressed. Use of existing validation sets (Cheng et al. (2021), TermPicks/Goliber et al. (2022), and specifically Gourmelon et al. (2022)) would be advisable, as this would allow a fair comparison of this method with existing studies on established measures.

  We agree with the reviewer that using average uncertainty to compare with the measure of uncertainty defined in other related works is somewhat misleading. Following our response to comments from Reviewer 1, we have built a test dataset by randomly choosing 100 traces from TermPicks and used the rest of the TermPicks dataset to train the network from scratch. After training the network, we apply it to the test dataset and quantify the mean distance error between the network's predictions and the manual delineations. The test error of the network is 79 meters, which is now used to compare with others.

- A related concern to be noted is the biases inherent in the chosen validation metrics. One validation metric (average distance between duplicate picks from Landsat-8 and Sentinel-2) is biased towards lower/better values, since it is only calculated on higher resolution images, and doesn't measure the method's performance with respect to manual delineated observations that function as the ground truth. Furthermore, this uncertainty quantification cannot be calculated across the entire dataset, so its use as a metric to gauge the quality of the dataset is questionable.

  We only use duplicated Landsat-8 and Sentinel-2 since (i) duplicate Sentinel-1 traces are used for the georeferencing offset, and (ii) Landsat-5 or -7 lacks overlap with other datasets. We have clarified

this point in the text. We now build a test set to quantify the overall error by measuring the deviation between the network's predictions and manual delineations. Since this comment is similar to Major Concern 1 from Reviewer 1, we respond with the same comment as we did there: Quantifying error based on manual delineation involves a trade-off: the more representative the error is, the more manual effort it takes. Since we aim to produce as large a terminus dataset as possible (with a resulting 278,239 glacier termini), a highly representative error would require too much manual effort, which violates our primary objective to save manual effort. For this reason, we still keep the two automated ways to quantify the uncertainty of the terminus data. We agree that uncertainty and error are not the same.

- The data itself has a few issues that require reevaluation of the automated screening module. Within the provided dataset, there are fronts that are closed loops, make large spatio-temporal jumps, or are otherwise erroneous. Additionally, there is a non-negligible number of glaciers with termini that are cutoff by the boundaries of the ROI, which should be expanded and/or otherwise addressed.

  Without specific time/location identification of these issues, it is difficult to address this comment. Perhaps the reviewer is referring to Figure 3, which shows pre-screened results. Our aim with this figure is to demonstrate some of the issues that we built the screening module to detect. Regarding the ROIs, without a clear identification of the glaciers/times with these issues, it is hard to address this comment. Our ROIs are prepared manually at the beginning of the entire process, and we carefully choose the ROIs to make them cover the glacier termini over their entire image acquisition period.

- While the primary contributions of this study are the data processing pipeline and dataset, there is value in providing some analysis of the results, such as commenting on the general/regional area change trends (as shown for individual glaciers in the supplement, and to a degree in Figure 6), volume loss (when integrated with velocity datasets, though this may be out of scope), or correlations with temperatures/other measurements.

  The main objective of this study is to build a fully automated pipelined that can continuously produce terminus traces and generate a huge terminus trace dataset. We agree with the reviewer that scientific investigation is important and interesting. However, it is out of the scope of this study and will be accomplished in future works that leverage our data compilation.

- The integration of figures in the manuscript could be better handled. Specifically, few figures are referenced within the manuscript (6, 8, 9, and 10 being the exceptions).

  We merge Figure 3 and Figure 5 together as they both show the screening module. We believe that the rest of the figures in the manuscript have distinct and relevant purposes, and we have double-checked that they are all referenced in the manuscript.

- It would be in the best interests of the community for the TermPicks derived training data to be released for ease of use for future projects.

  Thank you for this suggestion. The reference polygons and label polygons converted from TermPicks traces have been included in the AutoTerm dataset now (10.5281/zenodo.7527485).

- The training & pre/postprocessing of the network can be elaborated upon. The learning rate/regularization factors are less important/useful than information such as the optimizer used, number of epochs trained on, the total number of images trained on, loss function used, vectorization algorithm, and data augmentations used (i.e., if no data augmentations were used, why not, and if so, what were they).

We add the missing information in Line 208: "*To train the network, we use binary cross entropy as the loss function and stochastic gradient descent method as the optimizer with an L2 regularization factor of $5\times10^{-4}$,as recommended by Zhang et al. (2021). Based on the learning rate in Chen et al. (2018b) and Zhang et al. (2021), we train the network with learning rates of $5\times10^{-3}$, $2\times10^{-3}$, and $1\times10^{-3}$, and choose $2\times10^{-3}$owing to its lowest validation loss.*"

We adopt similar post-processing procedures with Zhang et al. (2019) that vectorize deep learning output to generate terminus traces. It is now added in Line 139.

The information about data augmentation can be found in Line 199.

**Specific Comments:**

**P2 L58:** I would recommend adding Gourmelon et al. (2022) and Loebel et al. (2022) to this list.

We thank the reviewer for pointing this out and have added these references to the reference list.

**P3 L70-71, P7 L210:** There are automated verification steps in Cheng et al. (2021), which includes filtering out unconfident predictions from the DL classifier.

We thank the reviewer for pointing that out. Cheng et al. (2021) has an automated data screening based on the deviations of two classifications of the network.

We have changed the related text as: "*Many previous DL methods applied to terminus delineation do not have quality control (Mohajerani et al., 2019; Zhang et al., 2019). Where it does exist, data screening has been simplistic and not automatically applied. For example, Zhang et al. (2021) only considers the complexity of the terminus shape and removes traces with abnormal complexity (which, in turn, requires a threshold to be established for each glacier), Baumhoer et al. (2019) only considers outliers that arise in a time series of terminus position change over time, and Gourmelon et al. (2022) remove the outliers based on terminus length. Cheng et al. (2021) however did design an automated data screening based on the deviation of two classifications from the network. Our screening module is based on using the physical properties of glacier termini.*" Line 224.

**P8 L225:** Could a detail/edge preserving speckle filter be applied? Or other types of Sentinel-1 processing steps to reduce speckle noise?

Considering the coarse resolution of the Sentinel-1 images, we did not apply the speckle filter to avoid blurry images. Also, glacier termini are still observable from the original Sentinel-1 images, even with speckle noise.

**P11 L341:** Is there a limitation (such as spatial coverage gaps) restricting ice mask generation to 2018-2020, or could they be made for other years?

They could be made for other years. For certain years, some glaciers might lack terminus traces, but there are no significant spatial coverage gaps in general. We only create updated masks annually beginning in 2018 to serve the ICESat-2 community needs for improved accuracy of laser returns during periods of extensive glacier terminus retreat. We have clarified this in Line 295.

**P21 Figure 1:** The flowchart is a not straight forward to follow. Perhaps consider separating the training/inference flowcharts, or organizing it in a more linear fashion.

The figure is composed of three parts: network training (black arrow), terminus inference (blue arrow), and longevity maintenance (red arrow). We chose a figure design to separate these procedures. Moreover, our figure

was designed to make good use of the figure space. Based on this comment, we have revised the figure to make the figure clearer by highlighting the training and inference.

**P26 Figure 6:** The color of the uncertainty bars and your results are the same (both are black). This makes the figure hard to interpret. Additionally, consider using colorblind friendly color schemes.

We have changed the color schemes and removed the uncertainty bars in the figure, as this figure is mainly for demonstrating the improved temporal resolution of our results. We also modified the color scheme of Figure 8 and Figure 11 and checked the figure through https://www.color-blindness.com/coblis-color-blindness-simulator/ following comments from Reviewer 1.

**P31 Figure 11:** Are the uncertainty bars for all of GID164's picks the same size?

Yes. The uncertainty bar is measured by using duplicate traces. Termini of the same glacier have the same uncertainty valued from duplicate traces. We added the description of the uncertainty bar in the figure caption. We have added two sentences to make the description clear:

Line 274: "*For each glacier, we average the uncertainties from all duplicated traces and use the mean to represent the uncertainty of that glacier.*"

Line 291: "*Thus, in total, each glacier will have six measures of uncertainty: one from duplicate traces and the other five estimated by MC dropout for each sensor."*

**Reference**


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 – 22, https://doi.org/10.3390/rs11212529, 2019.

Cheng, D., Hayes, W., Larour, E. Y., Mohajerani, Y., Wood, M. H., Velicogna, I., and Rignot, E. J.: Calving Front Machine (CALFIN): Glacial Termini Dataset and Automated Deep Learning Extraction Method for Greenland, 1972-2019, The Cryosphere, 2020, 1 – 17, https://doi.org/10.5194/tc-2020-231, 2020.

Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, Proceedings of the 33rd International Conference on Machine Learning, 48, 2016.

Gourmelon, N., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery, Earth System Science Data, 14, 4287–4313, https://doi.org/10.5194/essd-14-4287-2022, 2022.

Loebel, E., Scheinert, M., Horwath, M., Heidler, K., Christmann, J., Phan, L. D., Humbert, A., and Zhu, X. X.: Extracting Glacier Calving Fronts by Deep Learning: The Benefit of Multispectral, Topographic, and Textural Input Features, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12, https://doi.org/10.1109/TGRS.2022.3208454, 2022.

Mohajerani, Y., Wood, M. H., Velicogna, I., and Rignot, E. J.: Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study, Remote Sensing, 11, 74 – 13, https://doi.org/10.3390/rs11010074, 2019.

Zhang, E., Liu, L., and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13, 1729 – 1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

Zhang, E., Liu, L., Huang, L., and Ng, K. S.: An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery, Remote Sensing of Environment, 254, 112 265, https://doi.org/10.1016/j.rse.2020.112265, 2021.