Response to Review 1

We greatly appreciate the detailed and thorough review by Reviewer 1 throughout the entire review progress.

Font Color in this response

The **yellow** color represents the second round of the response. The **black** represents the third round of the reviewer's comments. The **blue** color represents the third round of the response.

1. Thank you. Please add the information that the test set is randomly chosen from TermPicks to the manuscript. Line 225 *„We measure the test error by calculating the averaged width of the enclosed area bounded by the TermPicks traces and the network predictions"* – How is width defined here? What happens when the prediction crosses the trace? Will that negate the error? I'm still not sure how exactly the error is calculated. A formula or figure would be helpful.

   The information that the test set is randomly chosen is described in Line 221: "*From TermPicks traces, we randomly select 100 traces as the test set and take the rest into the training set.*"

   The averaged width means the enclosed area bounded by the TermPicks traces and the network predictions divided by the length of the TermPicks traces. If there is a cross, we calculated the area on both sides of the cross and add them up.

   We rephrase the sentence in Line 225 as: "*We measure the test error by calculating the enclosed area bounded by the TermPicks traces and the network predictions divided by the length of the TermPicks traces. In the case where there are crosses between a* TermPicks trace *and a network prediction, we calculate the area on both sides of the crosses and then add them together.*" We choose not to an equation in the manuscript as the revised text explains itself.

2. I think it is unclear what is meant by „the deep learning network" (Line 404 and same in line 79) – I was assuming you meant previous deep learning models. From what I understand now, you mean your own model, which was improved by incorporating TermPicks into the train set. Please rephrase these sections, as like this it sounds like you would improve over previous models (which is not proven – actually, you just have indices that it performs on par).

   We agree with the reviewer and changed "network" to "model" throughout the manuscript.

3. Additionally, please state in section lines 326-328 that the test sets are different and that the test errors are calculated slightly differently (please see this quote from Cheng et al. 2021: „The primary quality assessment method is the mean distance error (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 have rephrased the sentence as: "*Our averaged test error is 79 meters (Table S1), which is comparable to previous studies where errors range from 33 to 108 m (Mohajerani et al., 2019;*

Zhang et al., 2019; Baumhoer et al., 2019; Cheng et al., 2020), although the test set and the way of calculating test error are slightly different".

4. Actually, the confusion does not come solely from the method section. For example, section 2 – „input data" - input to what? The pipeline? I was under the impression that the model would just get remote sensing imagery as input and not additionally an ice/ocean mask?

   It is the input data of the pipeline. The model would just need remote sensing imagery to produce glacier termini. We need a reference ice/ocean mask to update the mask using the newly produced termini. The name of the section 2 is now changed to "Input Data of the Pipeline".

5. We add one sentence in Line 453: "Another limitation is that even though we include additional training data, the network might struggle with some challenging situations (Fig. S2)."

   Should this not be investigated to prove the correctness of your produced dataset? Please check and report it in the manuscript.

   We have calculated the uncertainty of the terminus traces, which reflect such a limitation. Our screening module can detect some of the wrong picks caused by this limitation. We also mentioned in Line 487 that: "*We depend on future community feedback about our products to assist in identifying issues not caught by our screening module. This is because the massive amount of data precludes the ability to guarantee the quality of each individual trace.*"

6. We use the cubic upsampling method.

   Please add this information to the manuscript.

   We rephrase the sentence as: "*Specifically, we upsample small images (image width less than 1000 pixels) by an integer value using cubic interpolation so that their widths are just over 1000 pixels.*"

7. This is Table S3 now. Please correct.

   Corrected in the manuscript.

8. Network „architecture" is actually a technical term. Hence, please do not call it „structure". Moreover, please also add the reference (Chen et al., 2018b) after this sentence directly, as the large learning capability has been shown by it and not Zhang et al. (2021).

   We agree with the reviewer and changed "network structure" to "network architecture" throughout the manuscript and add the reference (Chen et al., 2018) right after the sentence: "*This network architecture has been proven to have large learning capability (Chen et al., 2018), spatial transferability, and the capability of using multi-sensor remote sensing images (Zhang et al., 2021).*"

9. The question was if the screening can only be done on time series or whether it is possible also when I just want one single trace from one satellite image.

   The screening will not be able to work effectively if there is only one trace as the screening method is essentially an outlier-detection method. However, this limitation is not a concern in the context of glacier termini extraction because we will typically have a wealth of remote sensing images as well as terminus data for screening purposes.

10. We add one sentence in Line 413: “Despite the success of the screening module in Greenland, further validation will be needed as applying it globally.”

    Thank you, however please add this in the limitations section.

    We have moved this sentence to the limitation section, Line 454: “*The fourth limitation is that further validation will be needed as applying it globally despite the success of the screening module in Greenland.*”

11. The duplicate uncertainty is calculated from the Landsat-8 and Sentinel-2, but we use the value to represent the network’s uncertainty on a certainty glacier. Therefore, such an uncertainty is more representative of Landsat-8 and Sentinel-2, and would be biased towards a lower value when representing the uncertainty of results obtained from other satellites. To clarify, we added on sentence in Line 465: “Since Landsat-8 and Sentinel-2 images have the highest resolution among the five satellites, using the duplicate uncertainty to represent the error of results obtained from other satellites would be biased towards lower values.”

    This must also be part of the limitations section, not only the „Difference of the two types of uncertainties“ section.

    We added on sentence in Line 455: “*The fifth limitation comes from the biased value of duplicate uncertainty as Landsat-8 and Sentinel-2 images have the highest resolution among the five satellites.*”

12. We randomly choose ten images from each of the five sensors and make three inferences for each of the ten images. The related description can be found in Line 298. We combine For loop and “shuf -n 10” in Bash to automatically and randomly choose ten images for calculating MC dropout uncertainty.

    Sorry, I think I did not make myself clear. Let me rephrase: When I apply your pipeline to a new glacier in Antarctica for one date and one sensor, will it calculate the MC dropout uncertainty for it?

    Yes, it will.

13. Your understanding is mostly correct. It’s just that when all three variables are uniformly distributed, even if we have a lower T_L and higher T_U, we will filter out more predictions than normal cases as the results will have a concentrated distribution in the normal cases. I’m sorry, I do not understand what you mean by normal cases here and what your difference between results and predictions is. Which concentrated distribution? Please explain, as my argument still stands, and I think this is important for both your dataset and future use of the pipeline.

    In the normal case, the distribution will have a Gaussian-like distribution (concentrated distribution, Figure R1). And you are correct, when variables (length, smoothness, and enclosed area) are evenly distributed, the screening module will fail to detect the wrong picks. Sorry that we made a mistake in the previous response.

    The limitation of the screening module is described in Line 458 : “*The second limitation is caused by our assumption that the screening module provides high-quality results. This assumption rests on the choice of thresholds defined by the interquartile range in the screening module. Thus, when* *most results for a glacier are not credible, the screening module might not be able to clean the results because the random distribution of the terminus attributes leads to improper thresholds.*"

However, the low success rate can still be used to indicate poorly performing glaciers, although it may not serve as a warning for all of them. To compensate for this limitation, we will use both uncertainty and success rate to indicate poorly performing glaciers. We rephrased the sentence in Line 480 as: "*The network's failure will result in many termini not passing the screening and high uncertainty. The pipeline can use the low success rates and high uncertainty to alert us to prepare more training data for the corresponding glaciers.*"

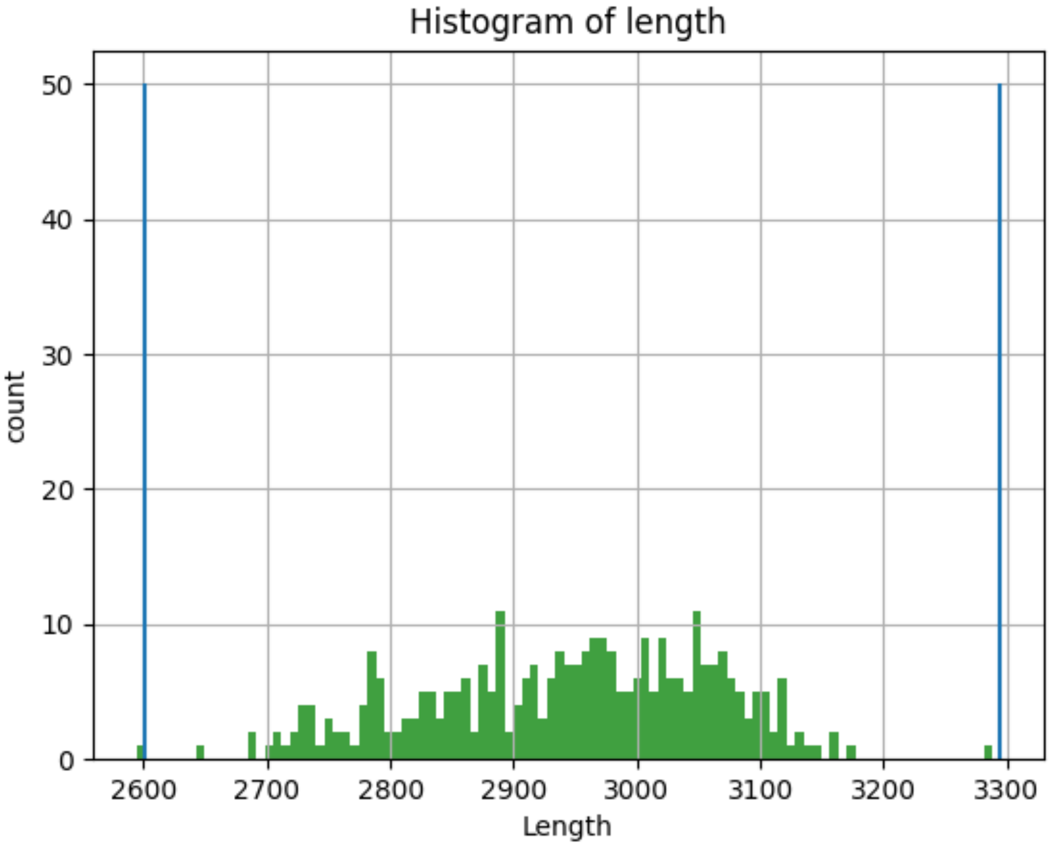

Figure R1. One example of the terminus length distribution in the normal case. The two bars indicate the lower T_L and higher T_U, respectively.

14. Please revise all figures, as there are still several that show green and red in one plot.

    We have changed the green color to dark green and checked the figure through https://www.color-blindness.com/coblis- color-blindnesssimulator/.

15. We thank the reviewer for the comment and added one sentence in Line 465: "Since Landsat-8 and Sentinel-2 images have the highest resolution among the five satellites, using the duplicate uncertainty to represent the error of results obtained from other satellites would be biased towards lower values."

Sorry, I should have been clearer. This must also be part of the limitations section, not only the „Difference of the two types of uncertainties“ section.

We added on sentence in Line 455: “*The fifth limitation comes from the biased value of duplicate uncertainty as Landsat-8 and Sentinel-2 images have the highest resolution among the five satellites.*”

Response to Review 2

General Comments:

The authors' response addresses the questions and comments raised by all reviewers, and integrates the feedback into the revised manuscript.

Specific revisions include detailed responses and technical corrections to issues in the dataset, and integration of suggested quality control measures into the methodology. The authors acceptably address reviewer concerns, and the added text is free from syntactical errors.

After reviewal of the author's responses, as well as the changes to the revised manuscript & associated dataset, I can recommend that this submission should be accepted, at the editor's discretion.

We greatly appreciate the careful inspection of the data and the constructive suggestions by Reviewer 2 throughout the entire review progress.