Font Color in this response
The **black** color represents the first round of the reviewer's comments, the **blue** color represents the first round of the response, the **green** color represents the second round of the reviewer's comments and the **yellow** color represents the second round of the response. **Purple** represents the third round of the reviewer's comments.

Major Concern 1: Evaluation Protocol
The pipeline has not been properly tested, and hence, we can not yet rely on its output. In my understanding, the authors seem to confuse uncertainty estimation with error assessment. In line 245, they call the calculation of the difference between prediction and ground truth „uncertainty quantification". The authors then claim that comparing to manually picked traces „requires significant manual effort" because it would have to be redone, as „network accuracy likely varies over time as glaciers experience different conditions". Instead, the authors use two different uncertainty quantifications that do not rely on ground truth data. Calculating uncertainties is definitely useful, and the two used ways of calculating the effect of different sources of uncertainty (model inherent and input inherent) look very promising. However, calculating the uncertainty is no substitute for an error assessment. The authors themselves state in line 395: „if both duplicated traces are deviated from reality but are close to each other, the uncertainty would not represent the reality." It is, therefore, indispensable to calculate the deviation of the network's predictions to manually delineated ground truth traces on a test set that is independent of the train set. First, we need to know how well the network is performing at the moment before we apply it to new unseen data and afterward assess whether the network's performance degrades when new sensors are used or other conditions change (called domain shift in machine learning).
We agree with the reviewer that the difference between predictions and ground truth should be called "error", while the difference between duplicate traces should be taken as "uncertainty". We have identified places in the manuscript where this terminology may have been confused and have updated the text. In addition, we have performed a test of the network as follows. We randomly choose 100 traces from TermPicks as a test dataset and use the rest of the TermPicks data to train the network from scratch. After training the network, we apply it to a test dataset and quantify the deviation of the network's prediction to manual delineations in the test dataset. This reveals a test error of 79 meters, which is similar to previous authors (Mohajerani et al., 2019; Zhang et al., 2019; Baumhoer et al., 2019; Cheng et al., 2020). The description of this test is added to the manuscript in Line 214 and Line 316.

Thank you for performing this test.
Please add more information in section 3.3 about the evaluation on the test set (train-test split – which images were picked for the test set exactly – this information ensures reproducibility; how exactly the error metric is calculated; etc.). Moreover, a split of the error on the test set between sensors would give additional valuable insights.

We appreciate the reviewer's comments. The list of the test set is provided in Table S1, which can be found in Line 326. The method of quantifying test error is described in Line 225. We added a new table (Table S2) to show the test error among the five sensors.

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.

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.

I'm not sure that I am understanding the authors correctly, but in my understanding, this trade-off between the representativeness of the error and the manual effort is not correct. If the test set is small, it needs to be chosen with greater care such that the error will be representative, i.e., the test set should cover the possible variability of the data the network will see.

What I understand from the author's second sentence is: They trade off quality assessment for quantity. As the authors do provide not only the dataset but also advertise their pipeline for future use, the quality assessment needs to be thorough. Still, keeping the uncertainty quantification is a great bonus.

We agree with the reviewer. We now have a test set that contains 100 images to quantify the test error. To further assess the quality of the data, we keep the original two ways of uncertainty quantification.

Although the reviewer states that we cannot rely on our model output, even without the model test we have now performed, we believe our data to be reliable for the following reasons. First, our terminus traces match the remote sensing images (Fig. 4). Second, the time series of terminus variation are in agreement with both Termpicks and CALFIN (Fig. 5). Third, the time series of terminus variations show a clear seasonal signal (refer to the time series data described in section 4.4), which would not be revealed if our terminus traces are unreliable.

Fig. 4 shows only six example traces, and in my regard, checking all 278,239 termini visually manually is also some manual effort, as even if the quality of each trace could be checked in one second, checking all traces would still require at least 10 days. I do not know how many and which images the authors checked, making the assessment not reproducible and subjective. Fig. 5, on the other hand, is a very nice analysis and indicates that there is probably no systemic error in the produced data. Still, this is just a rough hint at the quality and can not replace the test on a test set, which I would like to thank the authors for now providing.

We appreciate the reviewer's comments. We didn't manually check the quality of each trace. Instead, we used the test error and automatically-estimated uncertainty to demonstrate the reliability of the data.

Additionally, an experiment should be conducted to determine whether and by how much the error between prediction and ground truth on the test set is reduced when the screening module is applied versus not applied. In this way, the effectiveness of the screening module can be demonstrated. The same holds for the upsampling of small images (it is not sufficient to visualize the results of one sample, as shown in Fig. 13).

The screening module belongs in post-processing, and is thus not related to network inference or training. Instead, the screening module is for detecting outliers in order to improve data quality. Fig. 3 and the red crosses in Fig. 5 demonstrate the effectiveness of our screening module. For these reasons, we did not see it necessary to validate the effectiveness of the screening module on a test set.

The screening module is part of the complete pipeline which the authors propose. Hence, its effectiveness should, in my regard, be demonstrated in a thorough way as well. This can be done by once using the trained pipeline with and once without the module and reporting the difference in the error metric. Moreover, please also report the differences in the error metric for each sensor, as different sensors are handled differently by the screening module.

We applied the screening module to the test set. For the network trained with TermPicks, the test error is 62 meters after the screening module and 79 before the screening module. The success rate is 90% for the network trained with the TermPicks and 46% for the network trained without the TermPicks.
We added two sentences to describe this:
Line 329: "The success rate of the test set is 90%, and the test error was reduced to 62 meters after the screening module."
Line 404: "That network has a test error of 315 meters and a success rate of 46%, while the network trained with TermPicks has a testing error of 79 meters and a success rate of 90%."
We added a new table (Table S2) to show the test error among the five sensors before and after the screening.

Thank you!

We agree with the reviewer that the test error is needed to demonstrate the effectiveness of upsampling as it is a pre-processing procedure. Thus, we have also conducted an upsampling test. For this test, we randomly select 36 images of five small glaciers to be a test set for size normalization. These images are not included in the training set for the independent evaluation of the size normalization's effectiveness. We add a new table to show the test error and uncertainty from duplicate traces with and without size normalization

(Table S2). The results show that size normalization can effectively reduce test error and uncertainty. The related description is now added in Line 429.

Great work! Thank you! Just the formulation „36 images of five small glaciers that are beyond the training set as the test set“ is hard to follow. Please rewrite as done above.

We appreciate the reviewer's comments and rephrased the sentence as "We randomly select 36 images of five small glaciers as the test set for size normalization. These images are beyond the training set." Line 428

Major Concern 2: Generalizability

The pipeline has to be tested on out-of-sample data (i.e., glaciers not present in the training dataset) and data outside of Greenland to show generalizability to the global scope.

1. Line 451 „Owing to the transferability of deep learning, the entire pipeline has the potential to be applied to many other outlet glaciers around the world“
2. Line 135 „converting the TermPicks terminus data into a training dataset suitable for deep learning highly generalizes the network“

These claims have to be proven on such a test set. As most manually annotated traces available from related work are part of TermPicks and hence, have been used for training, another test set has to be used. For testing on SAR imagery, the dataset provided by Gourmelon et al. could, for example, be used, as it is not incorporated in TermPicks (except Jacobshaven, which probably has overlaps with TermPicks). However, test data for optical imagery might have to be created manually (e.g., from Antarctica or the Russian Arctic). At least, I am unaware of a dataset based on optical imagery that is not incorporated in TermPicks.

The importance of the size of training data in the deep learning field has been well demonstrated. For instance, Sun et al. (2017) showed that the network's performance increases logarithmically based on the volume of training data size. For this reason, we see no need to provide additional proof that our model has the ability to generalize.

The authors are correct that there is a relationship between dataset size and the performance of a network. This, however, does not mean that every network that is trained with a sufficiently large dataset will have adequate performance. This must be proven on a test set.

Table S1 shows the test errors before and after training with TermPicks traces. The test error was improved from 315 meters to 79 meters. The related description can be found in Line 404.

Thank you!

Further, our study focus is on Greenland alone – not on a global scale. We merely identify the potential for our model to be used at that scale. That said, out of interest we conducted an experiment in which we trained the network with only 1466 training examples prepared manually without including TermPicks. The test error of this network is 315 meters, which is much larger than 79 meters that we have after training with TermPicks. Such an improvement demonstrates the generalization improvements brought by TermPicks. The related description is added in Line 405.

It seems I have misunderstood what the authors meant by „generalizability“. I assumed the meaning was how well the model would predict the termini of images from the target distribution (i.e., all glaciers with termini around the world). However, it seems the authors merely meant how well the model would predict the termini of images from the distribution of data used to train the model (i.e., glaciers that are included in the train set but from future time points given that no other variables significantly change).
I apologize for the confusion. With the evaluation on the test set, the generalizability is proven, and the additional test with TermPicks also nicely proves the sentence from line 135.
However, it has to be stated more clearly (and also in the abstract) that the pipeline is only tested for glaciers in Greenland and not on a global scale and, therefore, can only be applied to Greenland without further precautions!

No need to apologize at all! We appreciate the reviewer's comments as they have significantly improved the manuscript. We have rephrased the sentence in the abstract as "The pipeline has been tested on glaciers in Greenland with an error of 79 meters." Line 16.

Thank you!

Major Concern 3: Comparability
It is not possible to compare the calculated uncertainties of this manuscript to the errors calculated in related works, as done in, e.g., line 304 or line 379. Two totally different metrics are compared here, and studies have been conducted on different datasets. For a valid comparison, the exact same network/pipeline needs to be tested on different datasets, or different networks/pipelines have to be trained, optimized, and tested on the exact same data (a so-called benchmark dataset). Altering both the dataset and the network/pipeline introduces too many changes, and a changed performance could result from either the different dataset (for example, the test set might be easier, and therefore, the performance of an otherwise worse performing network would be better on this test set) or the different network/pipeline.
Concludingly, the claimed improvements 1 and 2 (line 377 „1 increasing the generalization level of the deep learning network to enable more and better quality terminus predictions; 2 deploying size normalization to improve the accuracy of terminus delineation for small glaciers“) are not proven.
One way to show the superior terminus prediction performance on SAR imagery could be the use of the benchmark dataset recently proposed by Gourmelon et al. (2022) (i.e., retraining the pipeline on the train set and evaluating it on the test set using the stated metrics). To the best of our knowledge, there is no equivalent benchmark dataset for optical imagery.

Now that we have performed an error estimate we can more accurately compare our network error to that from other studies (Line 315), and find that our network's performance is comparable with other studies, not superior or inferior.

Please ensure that this is reflected correctly everywhere in the manuscript (e.g., lines 403-407 suggest that your performance is superior).

Although our test error is comparable with other studies, we indeed improve the method regarding the generalization. Converting the TermPicks traces into the training set makes it more representative of the real world and makes the network more generalized, which is demonstrated by the test errors before and after including TermPicks.

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).
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 agree with the review that different networks should be trained and tested on the same benchmark dataset to determine the best. However, such a test is outside of the scope of this study. The objective of this study is not to demonstrate which deep learning network is better but to generate a terminus dataset with spatial coverage and temporal resolution that no prior study has with the addition of further automation in the production pipeline.
For generalization and size normalization, please refer to the response of Major Concern 3 and Major Concern 1, respectively.

With the evaluation on the test set, this is addressed satisfactorily.
I understand that a rigorously correct comparison is outside the scope of this study.

We appreciate the reviewer's comments.

Major Concern 5: Structure of the manuscript
The structure of the manuscript has to be improved upon. There is a mix-up between the training and inference of the pipeline, and some information is given twice at different positions in the manuscript. It is hard to tell when the authors write about the newly derived dataset in contrast to the dataset derived from TermPicks for training the network, e.g., in line 295 („We find an average success rate of 64%“), it is unclear on which dataset the success rate was calculated. I would suggest splitting the manuscript into two main parts as follows, but there could also be another better split-up:

1. Training Pipeline: manual delineated dataset creation (TermPicks + additional manual annotations), neural network (architecture), network training (train-validation-test split, learning rate, number of trained epochs, etc.), screening module, error calculation, uncertainty estimation
2. Inference Pipeline: new data acquisition + pre-processing, uncertainty estimation on this newly derived dataset, ice/ocean mask updates

We thank the reviewer for these comments. We decided to structure the manuscript following the order of data processing. We first collect remote sensing images and conduct preprocessing (section 3.1). Second, we generate the training dataset by converting the terminus traces in TermPicks into label polygons and pairing polygons with the remote sensing images. Third, we introduce the network structure and the training progress in section 3.3. Sections 3.4 to 3.7 are post-processing procedures after applying the well-trained network to make inferences on the 433,721 images collected via Google Earth Engine. That said, we notice that the original section titles may have caused some confusion. Thus, to address this comment we changed the title of Section 3.2 from “Generalizing the network” to “Generating training data from TermPicks”, and the title of Section 3.3 from “Deep Learning Network” to “The Structure and Training of Deep Learning Network”.
Furthermore, we have revised some of the text to make it clearer that we first perform network training and then we apply the well-trained network to all the images collected through Google Earth Engine to generate terminus positions. This is done in the last sentence in Section 3.3 as “*After the training, we apply the well-trained network to the test set for quantifying the test error and to all the images collected via GEE for generating the terminus dataset.*” Line 217.
The success rate is the percentage of terminus traces that pass the screen module. We add one sentence at the end of section 3.4: “*Finally, we estimate the success rate by calculating the percentage of the terminus traces that pass the screening module*.” Line 255.

Thank you. I think adding a paragraph like „The rest of the paper is structured as follows ...“ (as is done above) to the end of the introduction would aid the reader.
Please also provide the success rate for the test set.

We thank the reviewer for the comments. We added one paragraph at the beginning of the method section as the confusion mainly comes from this section. The added paragraph is: “The structure of the method section follows the order of data processing. We first collect remote sensing images and conduct preprocessing (section 3.1). Second, we generate the training dataset by converting the terminus traces in TermPicks into label polygons and pairing polygons with the remote sensing images. Third, we introduce the network structure and the training progress in section 3.3. Finally, sections 3.4 to 3.7 are post-processing procedures after applying the well-trained network to make inferences on all the images collected via Google Earth Engine.”
The success rate is 90% for the test set. We added one sentence in Line 329: “The success rate of the test set is 90%, and the test error was reduced to 62 meters after the screening module.”

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?

In section 4.1, the authors first introduce the ‚success rate‘, which should, however, be introduced in the methods section.

We add one sentence at the end of section 3.4: “*Finally, we estimate the success rate by calculating the percentage of the terminus traces that pass the screening module.”* Line 255.

Thank you.

Moreover, paragraph line 188 to 195 should be moved to limitations.

Line 194 to 201 describe the limitations of the input imagery, not the limitations of our work. These points are discussed here because they explain why we needed to prepare additional training data. Therefore, we think this text fits better in Section 3.2, which describes the training data preparation.

I understand. Does your pipeline still have difficulties with these kinds of images? If so, please additionally add a few sentences about it in the limitations.

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.

Major Comments:

5. The normalization of image sizes is not clear to me. Small images are upsampled, but large images are not downsampled. Hence, do they still have different sizes? I would not call this normalization, then. Moreover, the authors extract patches afterward, so the input size is always equal anyways. Additionally, only showing one figure that shows an improvement for one trace is not sufficient evidence that this upsampling generally improves the delineation performance. Please show the improvement in numbers over a complete, independent test set (refer to major concerns 1 and 2).

Different glaciers are not of the same physical size and therefore don't have similar number of pixels that fall within their fjord walls. This can vary from 300 pixels to 3000 pixels. As a result, the deep learning input (image patch) may sometimes cover more than the entire glacier and may alternatively sometimes only cover a part of a glacier.

After normalization, all small glaciers will be covered by images with a width of about 1000 pixels, regardless of their original image width. In other words, the normalization makes glaciers appear to the deep learning network as if they had a similar physical size. We have attempted to clarify this in the text. Line 168.

Thank you, I understand now that you meant the physical size. So, what is the mean physical size of one pixel now? Moreover, how is upsampling performed? Linear interpolation or a nearest neighbor interpolation?

The normalization will make small glaciers appear to have a larger physical size without changing the physical size of a pixel. For example, if a glacier has a physical width of 1000 meters and is covered by an image with a width of 500 pixels, after normalization, the image width will be 1000 pixels and each pixel will still have the same resolution as before. Thus, the nominal physical width of the glacier becomes 2000 meters.

OK.

We use the cubic upsampling method.

Please add this information to the manuscript.

We agree with the reviewer that test error is needed to demonstrate the effectiveness of upsampling as it is a pre-processing procedure. We randomly select 36 images of five small glaciers that are beyond the training set as the test set for size normalization. We add a new table to show the test error and uncertainty from duplicate traces with and without size normalization (Table S2). The results show that size normalization can effectively reduce test error and uncertainty. The related description is now added in Line 428.

Thank you.

This is Table S3 now. Please correct.

6. Section 3.3:

4. „This network has been proven to have large learning capability, spatial transferability [...]": These are quite big claims based on a train set of two glaciers and a test set of one glacier that are all located in Greenland (Zhang et al., 2021).

The large learning capability has been demonstrated by the paper of DeepLabV3+ (Chen et al., 2018b). Zhang et al. (2021) applied the network to a glacier beyond the training set, showing the network's spatial transferability. In this work, we apply the network to 295 glaciers in Greenland and generate 278,239 glacier termini, of which 17,906 terminus traces are from the training set. This means that 94% of our results are beyond the training set, which further demonstrates the learning capability and spatial transferability of the network.
As this sentence only references Zhang et al. (2021), the authors are not reporting on the present study. Please call it „network architecture“ and not „network“ to make clear what you are referring to and add the reference to Chen et al., 2018b as well.
We thank the reviewer for the comments and revised the related sentence as “This network structure has been proven to have large learning capability, spatial transferability, and the capability of using multi-sensor remote sensing images (Zhang et al., 2021).” Line 213. We also added Chen et al., 2018b to the reference.
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).

8. „The network training takes about a week“: This is quite long and might be due to a sub-optimal learning rate. Please specify not only the training time but also the number of trained iterations over the complete augmented dataset. Also, specify your train-test-split and your metrics for evaluation (refer to major concern 1). Moreover, did you use an early stopping criterion? You might have overfitted during this long training time.
The long training time is caused by the large training dataset. We have 17906 training examples, and the augmentation increases the training set by a factor of four. The training takes more than 600,000 iterations. Among all the training data, we select 5% of images randomly as validation datasets to conduct early stopping. The training will be stopped when the validation error stops decreasing for 3 consecutive epochs. We add these details in Line 214.
Thank you very much. I assume by iteration, inputting one batch into the network and backpropagating its loss to update the weights is meant? Please instead provide the times the network has seen the train set („iterations over the train set“), which is normally defined as the number of epochs. If my assumption is correct, that would be (600,000 * 16) / (17906 * 4).
Out of curiosity, is the checkpoint saved after each batch?
Yes, your understanding of the iteration is correct. The network training stopped at seven epochs. The training examples are counted by the number of images. The remote-sensing images will be split into image patches with identical sizes. In total, we have around 1,300,000 patches after augmentation. Therefore, the number of epochs is (600,000 * 16)/1300000≈7.
The checkpoint will be saved after each epoch if the validation loss decreases.
Thank you!

11. Line 224 „percentile of the data range“: Do you refer to the data range of the generated training data? Is this computed per glacier? Per satellite? The validity of these thresholds needs to be checked on an independent test set (refer to major concerns 1 and 2).
For each glacier, we will calculate the thresholds based on the termini from the same satellite. For instance, we will calculate thresholds for Sentinel-2 traces of GID2 and Landsat-8 traces of GID2 separately. We add one sentence in Line 242 to provide additional clarification: “*The thresholds are calculated automatically based on the results of the same glacier and same satellite.*”
Thank you. It is also now clear that you calculate these metrics on outputs of the network. So the screening can only be done if the pipeline is applied to a time series of images? Or do you store a rolling average of each glacier and each satellite in your pipeline?
Yes, the screening can be done if we apply the pipeline to a series of images and generate a bunch of termini.
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 module belongs to the post-processing, which is not related to network inference or training. The screening module is for detecting outliers and maintaining the data quality. Fig. 3 and the red crosses in Fig. 5 demonstrate the effectiveness of our screening module. So, we did not validate the effectiveness of the screening module on a test set.

Fig. 3 and Fig. 5 are insufficient proof of the effectiveness. The screening module is part of the complete pipeline which the authors propose. Hence, its effectiveness should, in my regard, be demonstrated in a thorough way as well.
We applied the screening module to the test set. For the network trained with TermPicks, the test error is 62 meters after the screening module and 79 before the screening module. The success rate is 90% for the network trained with the TermPicks and 46% for the network trained without the TermPicks.
We added two sentences to describe this:
Line 329: "The success rate of the test set is 90%, and the test error was reduced to 62 meters after the screening module."
Line 404: "That network has a test error of 315 meters and a success rate of 46%, while the network trained with TermPicks has a testing error of 79 meters and a success rate of 90%."
We added a new table (Table S2) to show the test error among the five sensors before and after the screening.
Thank you!

12. Line 227 „For outliers in terminus length, we remove both the lower and upper thresholds (Eqns. 1 and 2) because we do not anticipate large changes in terminus length in either direction (bigger or smaller)." As far as I understood, these thresholds were calculated on data for Greenland. Hence, the optimal thresholds for, e.g., Antarctica, might completely deviate from the ones calculated for Greenland. This might hinder the global applicability of the pipeline (refer to major concerns 1 and 2). This should be added to the limitations.
We agree with the reviewer that these thresholds differ for Greenland and Antarctica glaciers. They are different among glaciers in Greenland. However, these thresholds are determined **automatically** based on the distribution of termini from each satellite and each glacier. Therefore, we believe it is feasible to apply the method globally. We have clarified this in the text. Line 242.
I understand now. However, please indicate that the screening has not been tested on glaciers outside of Greenland.
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.

17. The results do, at some points, not validate the conclusions. No correlation was calculated (or it was not stated in the manuscript), and even a correlation would not necessarily induce causality. Please rephrase the conclusions to hypotheses.

1. Line 307 „glaciers with less training data will have larger uncertainties and lower success rates"
Rephrased as "*glaciers with less training data will probably have larger uncertainties and lower success rates*" Line 331.
2. Line 309 „since they have the highest spatial resolution"
Rephrased as "*Among the five datasets used, Landsat-8 and Sentinel-2 have the lowest average uncertainties, probably because they have the highest spatial resolution*." Line 333.
3. Line 310 onwards „The reasons for the Landsat-5 uncertainty are twofold [...]"
Rephrased as: "*The reasons for the Landsat-5 uncertainty might be twofold*." Line 335.
4. Line 314 „The higher uncertainty of Sentinel-1 images is due to its low image quality, coarse
resolution, and the lower volume of training data derived from this sensor."
Rephrased as "*The higher uncertainty of Sentinel-1 images could be due to its low image quality, coarse resolution, and the lower volume of training data derived from this sensor*." Line 338.

Thank you for addressing my points here. Please also go over the complete manuscript to search for further such claims (e.g., line 327 „However, the network does struggle to delineate termini in many wintertime Sentinel-1 images because of blurry boundaries and the lack of sufficient training data specifically using Sentinel-1 imagery"). There is some mix-up of results and discussion (e.g., line 325 „Such variations are largely caused by the uneven distribution of the training data—glaciers with more training data have higher success rates." Here again, causality is not proven. The authors **observe** that glaciers with

more training data have higher success rates, and **conclude** that the uneven distribution of training data might cause this.) which should be addressed. Please separate plain results and conclusions into two sections (i.e., the conclusions should be moved from the results section to the discussion section).

Thanks for the comments. We rephase the following sentences:

• Line 331 "Such variations could be caused by the uneven distribution of the training data---glaciers with more training data have higher success rates."

• Line 332 "However, the network does struggle to delineate termini in many wintertime Sentinel-1 images, probably because of blurry boundaries and the lack of sufficient training data specifically using Sentinel-1 imagery."

• Line 340: "The duplicate trace uncertainty varies between glaciers along with success rates might be because the training data is not evenly distributed for each glacier"

We move the third paragraph of section 4.1 to the discussion section (section 5.4 Difference of the two types of uncertainties).

Thank you.

18. Line 325 „the uncertainty from duplicate traces is more representative of Landsat-7 and Sentinel-2 than other datasets" - Is it not only representative of these two datasets, as it was only calculated for these?

For each glacier, we average the uncertainties from all duplicated traces and use the mean to represent the uncertainty of that glacier. Each glacier has one value of uncertainty from duplicate traces, and that value is more representative of Landsat-8 and Sentinel-2 as the value comes from these two satellites. 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*."

Still, my question stands: Is the uncertainty calculated with duplicate traces (and just this uncertainty, not the MC dropout one) not only representative of these two sensors, as it was only calculated for these? (Line 355)

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.

20. Line 355 „The metadata contains the date in YYYY-MM-DD, Glacier ID, source image satellite, and the uncertainty of each trace by averaging the two types of uncertainties provided": I thought the uncertainties were not available for every single trace, as they were only calculated for some of them due to computational limitations? Please clarify.

Each glacier has six measures of uncertainty: one from duplicate traces and the other five estimated by MC dropout for each sensor. The uncertainty of each trace is estimated by averaging the uncertainty from duplicate traces and the MC dropout uncertainty of its satellite sensor. We added one sentence to clarify in 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*."

Please additionally indicate in line 386 that the average is taken as the uncertainty measure for each of the glacier's traces.

Moreover, an additional question arises: In lines 293-296, the authors describe the procedure for how the dataset is generated. But how is it configured in the automatic pipeline if, say, it is used for glaciers in Antarctica? Will MC dropout uncertainty be calculated for each image, or how are images randomly chosen here?

We rephrase the sentence in Line 379: "The entire record of uncertainties is provided in a spreadsheet. Each glacier has six averaged uncertainty measures, including one from duplicate trace uncertainty and five from MC dropout uncertainties of different satellites."

Thanks.
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?

21. Line 434: „The pipeline can alert us of its failure based on the success rate within the screening module.": With your limitation that the screening module might not provide valid results for glaciers with few training examples, this alert might not trigger.
When most of the results are of good quality, the terminus features, such as length, will have a Gaussianlike distribution, where most of the terminus lengths are within the thresholds determined by the screening module, and the success rate will be high. For glaciers with few training examples, their results might be of poor quality. In that case, the distribution of the terminus features will be relatively scattered because many terminus lengths could be unreasonably long or short. As a result, the screening module will detect many traces as outliers, and these glaciers will have a low success rate, which can be used as an alert. We have clarified the text to point this out more readily: " *The network's failure will result in many termini not passing the screening. The pipeline can use the low success rates to alert us to prepare more training data for the corresponding glaciers.*" Line 464.
As the thresholds of the screening module are calculated for each glacier individually based on the network's predictions, the screening module will not filter out wrong predictions of a glacier where only wrong predictions ever occur (e.g., due to very few training examples). I assume all three variables (terminus length, curvature, and enclosed area) would roughly be uniformly distributed (as the predictions would be somewhat random), which would lead to a smaller Q1 and a bigger Q3 and, therefore, to a lower T_L and a higher T_U. This, in turn, reduces the amount of filtered-out predictions. Please correct me if I have an error in my logic.
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.

22. Please revise the color scheme of your figures, as red and green should not appear in the same plot (https://www.the-cryosphere.net/submission.html#figurestables).
Thank you for bringing this to our attention. We have modified the color schemes of Figure 6, Figure 8, and Figure 11, and checked figures through https://www.color-blindness.com/coblis-color-blindnesssimulator/.
Please also do the same for the remaining figures (5, 7, 10).
We are sorry for the confusion. We have changed the color scheme for figures 5,7, and 10. We have merged the original figures 3 and 5. So, figures 6, 8, and 11 in response actually mean figures 5,7, and 10
Please revise all figures, as there are still several that show green and red in one plot.

24. Figure S5: Visually, this does not appear to be a linear relationship. Have you done a correlation test?
Yes, the uncertainties from duplicate traces and MC dropout do not show a linear relationship. The differences in the two types of uncertainties are caused by their quantification methods and source images. We explained the details of the differences in the last paragraph of section 4.1, Data Quality.
I don't understand. Why does the caption then claim that there is a linear relationship?
Our previous response is somewhat misleading. There is a linear relationship between the two uncertainties but they are not exactly the same. The linear relationship is more clear in Landsat-8 and Sentinel-2 but less obvious among Landsat-5, Landsat-7, and Sentinel-1. Since uncertainty values vary drastically across glaciers, we now use the natural logarithm to show the comparison

between the two types of uncertainties (see the updated Fig. S5). The correlation coefficient is 0.69 for Landsat-8 and Sentinel-2, and 0.43 for the rest three satellites.
Thank you.

Response to Reviewer 2's comments

Major Comments:

- 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.
  I had not realized that there would be a bias towards better values, as Reviewer 2 here correctly states. This should be addressed somewhere in the text, even if the authors now additionally use the commonly used error metric on a test set, as this bias still exists for the uncertainty measure and should be handled with care.
  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.