Response to Reviewer 1's comments

General Comments

This paper presents an automated pipeline in Google Earth Engine for glacier terminus tracing together with a so-derived dataset and updated ice/ocean masks. Such a pipeline is highly needed and of great significance to the community. This extent of automation has not been reached in related works. We thank the authors for their valuable contribution!

While this paper employs a sound deep learning architecture in combination with a promising screening module, I have several major concerns, including the technical correctness of the evaluation protocol and, thus, the validity of the proposed study, as the generalizability of the deep learning network still needs to be proven. Furthermore, comparisons to other studies need to be conducted in a technically correct way, and the reproducibility of the study needs to be ensured by making the assembled training dataset publicly available. Lastly, the structure of the manuscript should be improved upon.

We greatly appreciate the detailed and thorough review by Reviewer 1. We have made our best effort to revise the manuscript based on the referee's comments and suggestions. Below is an item-by-item response to the specific comments by this reviewer.

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.

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.

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.

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 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.

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.

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!

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).

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.

Major Concern 4: Reproducibility
Please make your complete assembled training data (including the satellite imagery) publicly available, as only in this way the reproducibility of the results is guaranteed. Moreover, please also provide the manually created reference polygons for each glacier.

All our remote sensing images are freely available on Google Earth Engine. The code for collecting the data is also available on GitHub. TermPicks is also a publicly available dataset. The reference polygons and label polygons converted from TermPicks traces have been included in the AutoTerm dataset now (10.5281/zenodo.7527485).

Thank you for addressing this issue. Recreating a dataset might have produced some slightly different data, which might again have influenced the training of a network.

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.

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.

An explanation of the two uncertainty measures, as given in lines 319 to 323, should be moved to further at the beginning of the manuscript.

The principle of the two uncertainty measures is described fully in Methods (Section 3.6). Lines 344 to 355 explains the uncertainty results and focus on why these two uncertainties are different. Therefore, we believe lines 344 to 355 fit better in the Results section.

I understand. Thank you.

Major Comments:

1. It is unclear to me whether the name „AutoTerm" refers to the automated pipeline or the derived dataset, or both.

We thank the reviewer for pointing this out. We think it is a good idea to use AutoTerm reflect both the data and pipeline. We have changed the title of the manuscript to: "AutoTerm: an automated pipeline for glacier terminus extraction using machine learning and a "big data" repository of Greenland glacier termini."

Yes, thank you.

2. The title of the manuscript does not mention the automated pipeline, which is, in my humble regard, the most significant contribution. Hence, I'd argue for a more suitable title, e.g. AutoTerm: an automated Google Earth Engine pipeline for glacier terminus extraction and „big data" repository of Greenland glacier termini

We thank the reviewer's suggestion. We have changed the title of the manuscript to: "AutoTerm: an automated pipeline for glacier terminus extraction using machine learning and a "big data" repository of Greenland glacier termini."

Thank you.

3. It needs to be clarified whether the region of interest that has to be defined for each new glacier has to be a polygon like in figure 2 or whether it can simply be a bounding box.

Each region of interest is a bounding box. The polygon in Figure 2 is a reference polygon only for converting TermPicks traces into a training label polygon, which is described in the Figure caption. We have updated the text where needed to clarify this point. Line 146.

Thank you.

4. Line 163 onwards: „This allows glaciers with various natural sizes to have a similar image size in computer vision, which largely decreases the complexity of delineating glacier terminus." This statement (the second part of it) needs more explanation or a reference.

We rephrased the text here and added a reference: "*We then normalize the image size, which is commonly adopted in the computer vision field for better capturing object features with various physical sizes (Xu et al., 2017).*" Line 167.

Thank you.

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?

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.

6. Section 3.3:

1. „encoder-decoder structure [...] can obtain sharp object boundaries": Actually, an encoder decoder structure without skip connections would most probably not recover any details and, therefore, no sharp object boundaries. In Chen et al. (2018), they use a more sophisticated method to obtain the sharp boundaries: „A fully connected CRF [conditional random field] is then applied to refine the segmentation result and better capture the object boundaries."
We realize that there is a mistake in our citation. The original DeepLab use the CRF to refine the segmentation results (Chen et al., 2018a, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution and Fully Connected CRFs). Later on, the same author improved the DeepLab and developed DeepLabV3+ (Chen et al., 2018b; Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation), which is the network we use in this manuscript. The DeepLabV3+ adds a simple yet effective decoder module to refine the segmentation results especially along object boundaries (Chen et al., 2018b).
I see, thank you for correcting your reference.

2. „atrous convolution [...] senses multi-scale contextual information": It is not the atrous convolutions alone that make recognition of multi-scale contextual information possible, but the combination of atrous convolutions in ASPP. "Atrous convolution allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. It also allows us to effectively enlarge the field of view of filters to incorporate larger context without increasing the number of parameters or the amount of computation. Second, we propose atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scales." (Chen et al., 2018)
We thank the reviewer's comments and changed "atrous convolution" to "atrous spatial pyramid pooling".
Thank you.

3. „multi-scale contextual information [...] [is] helpful for our task since [...] we integrate remote sensing datasets with different spatial resolutions": Multi-scale refers to how many pixels a neuron is able to see (effective receptive field) and not how much square meters one pixel can see. Hence, multi-scale contextual information helps when the calving front covers many versus only a few pixels. Thus, it helps only indirectly with different spatial resolutions of the dataset.
We agree with the reviewer's comment. Multi-scale refers to the network's ability to sense various portions of the images but not sense images with different resolutions. We rephrase the sentence as

”*Sharp boundaries can improve delineation accuracy, and sensing multi-scale information helps indirectly when we integrate remote sensing datasets with different spatial resolutions.*” Line 205
Thank you.

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.

5. „The network is trained with a learning rate of 0.005 [...] as recommended by (Zhang et al., 2021)“: The optimal learning rate for training is highly dependent on the dataset as well as on the batch size (not just the model). Hence, the learning rate has to be treated as a hyperparameter, which has to be optimized on a validation set (not the test set). A sub-optimal learning rate can lead to significantly longer training times until convergence or no convergence at all.
We agree with the reviewer and have tried to train the network with learning rates of 5×10^{-3}, 2×10^{-3}, and 1×10^{-3}. The validation losses for these three learning rates are 0.023, 0.020, and 0.024, respectively. Overall, the validation losses are measured using binary cross entropy and are similar to each other. We chose the learning rate (2×10^{-3}) with the lowest validation error. We rephrased the sentence as: “*Based on the learning rate in Chen et al. (2018b) and Zhang et al. (2021), we train the network with learning rates of 5×10^{-3}, 2×10^{-3}, and 1×10^{-3}, and choose 2×10^{-3}owing to its lowest validation loss.”* Line 210.
Thank you very much.

6. „we choose the largest batch size (16)“: This should be „largest possible batch size (16) on an A100 GPU with 40/80 GB GPU memory“. Please specify whether your A100s have 40 or 80 GB GPU memory.
The sentence is rephrased as follows: “*we choose the largest possible batch size (16) on four A100 GPUs with 160 GB GPU memory in total. We set the batch size to a power of two to take full advantage of GPU processing (Kandel and Castelli, 2020)*” Line 212.
Thank you very much.

7. What exactly is meant by "maximize our computational power“ in line 204?
It means using as much GPU memory as possible. The sentence is now rephrased as “*we choose the largest possible batch size (16) on four A100 GPUs with 160 GB GPU memory in total*.” Line 212.
Thank you.

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?

7. Line 210 „do not have any quality control“: At least Cheng et al. have manual control. So, maybe rephrase it to „do not have any automated quality control“.
We thank the reviewer for pointing that out. Cheng et al. (2021) have 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.
Thank you.

8. Line 215 onwards: Please mention that the screening builds on top of existing works here (Zhang et al. 2021 – terminus curvature screening, Baumhoer et al. 2019 – time series outliers, Gourmelon et al. 2022 – removal of too short termini predictions), but goes one step further, i.e., doesn’t use any manual intervention or prior knowledge of the data.
The related sentences are revised as follows: “*Based on the previous works (Baumhoer et al. 2019 , Zhang et al. 2021, Gourmelon et al. 2022), we develop an automated screening module that forgoes any manual intervention or prior knowledge of the data.*“ Line 232.
Thank you.

9. Line 217 „Terminus length is determined by the sum of the piece-wise length along an individual terminus trace“. Please explain in more detail. This, at least for me, is hard to understand.
Each terminus trace is an ordered set of points. The length is the sum of the length between the two closest points. The description refers to how we calculate the terminus length, which might be confusing to readers. Terminus length is just the physical length of the glacier terminus. As a result of this confusion, we decided to remove this sentence.
Thank you.

10. Line 218 „Terminus curvature is computed between two adjacent points for each point along the terminus and then an average is taken for each terminus trace.“ This is also not completely clear to me. I think an equation would help.
We have rephrased the related text as: “*Terminus curvature is computed among every three adjacent points along the terminus based on Peijin Zhang’s work (https://github.com/peijin94/PJCurvature), and then an average is taken for each terminus traces.*” Line 235.
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?

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.

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.

13. Line 235 „We then repeat this screening procedure ten times to maintain the quality of the terminus product“: What screening procedure is meant here exactly now? All three or just the one with large areas? And does the outcome change when the screening procedure is done several times? If yes, please explain why.
The pipeline of the screening module is shown in Figure 3. For the first time, the inter-quartile range quantifies thresholds based on the distribution of terminus features such as length, and we remove the outlier traces. Such removal changes the distribution of terminus features, and we will have new thresholds for the second time and detect new outliers. We keep doing this ten times or until we don't find any more outliers. We have clarified this in the text. Line 254.
Thank you.

14. Line 245 „Traditional uncertainty quantification for glacier terminus position is conducted by calculating the difference between manually picked termini and automatically-picked termini.“ This is not uncertainty quantification but an error assessment (see major concerns 1).
We agree with the reviewer's comments and have revised the related sentence. See our response to major concern 1.
Thank you.

15. Line 262 „instead of quantifying the uncertainties of terminus traces, [Hartmann et al.] use the multiple inferences of MC dropout as extra information to retrain the network. “ This is not quite correct. Hartmann et al. use the model uncertainty on one specific input as additional information for a second network with dropout. This second network then outputs several predictions again from which uncertainties could be calculated - but instead, to make it more robust, the predictions are averaged to eliminate this uncertainty.
We thank the reviewer's comments and rephrase the sentence as "*Hartmann et al. (2021) applied MC dropout to glacier terminus delineation and built a two-stage approach. They used the uncertainty of the first network as additional information to train the second network. The multiple outputs of the second network are averaged to eliminate the uncertainty and get the final prediction.*" Line 283.
Thank you.

16. Line 267 „To strike a balance between computational cost and the reliability of the MC dropout, we randomly chose ten images from all the sensors and make three inferences for each of them“: It is not quite clear to me. Are ten images of each sensor taken? „in total each glacier will have two measures of uncertainty“ – So, also ten images of each glacier?
Our original description is somewhat misleading. For each glacier, we will randomly select ten images for each sensor, and we will have 50 images in total. Using the ten images from the same sensor, we conduct MC dropout to quantify one uncertainty for that sensor. We have two measures of uncertainty, one is from duplicate traces, and the other is from MC dropout. The ten images from the same sensor are only for quantifying MC dropout uncertainty. We rephrase the text as "*To*

strike a balance between computational cost and the reliability of the MC dropout, we randomly chose ten images from each of the five sensors and made three inferences for each of the images. 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." Line 289.
Thank you.

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).

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)

19. Line 308 „Among the five datasets used, Landsat-8 and Sentinel-2 have the lowest average uncertainties": Please give the exact numbers here. A table showing the different values for different data subsets would be good.
The numbers are shown in Figure 8. The missing information is now added to the figure caption.
Thank you.

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?

Line 423: „additional training data will be required to improve the data quality“: or an improved network/pipeline.
Revised as suggested.
Thank you.

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.

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).

23. Figure 9: Is the number on the bottom left the average? Moreover, it would be good to state between which sensors the duplicates were calculated in the description of the figure.
Yes. We have added the missing information in the figure captions.
Thank you.

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?

Specific Comments:

1. Line 56 onwards: Heidler et al. 2022 (Deep Active Contour Models for Delineating Glacier Calving Fronts), Loebel et al. 2022 (Extracting glacier calving fronts by deep learning: the benefit of multispectral, topographic and textural input features), Gourmelon et al. 2022 (Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery), and Davari et al. 2022 (Pixelwise Distance Regression for Glacier Calving Front Detection and Segmentation) are missing.
We thank the reviewer's comments, and all the missing citations are included.
Thank you.

2. Line 188: „Although TermPicks covers a range of conditions and brings great diversity to the training set, additional training data would improve the accuracy of the network in difficult situations." Please rephrase more cautiously (e.g., „... would presumably improve ..."), as you have no hard evidence that further training data would really improve the accuracy in this situation.
The sentence is revised as suggested.
Thank you.

3. Line 205: GPU -> GPUs
Revised as suggested.
Thank you.

4. Line 220 „With these three metrics, we calculate the lower (TL) and upper thresholds (TU) for each based on the inter-quartile range:" - The sentence structure is hard to follow. So, you compute the thresholds for each individual criterion?
Yes. The sentence is rephrased as "*For each of the three metrics, we calculate the lower (TL) and upper thresholds (TU) based on the inter-quartile range*:"
Thank you.

5. Line 417 „120 GB of GPU memory": I guess you mean 120GB RAM? There are only a 40GB and an 80GB A100 version as far as I know, and 4 (=number of GPUs) times 40 GB is already 160 GB.
I mean 120 GB of GPU memory. We have four A100 GPUs with a total memory of 160 GB, and we use 120 GB of memory. We didn't use all the GPU memory since we wanted the batch size to be a power of 2. When setting the batch size to 16, our network will need 120 GB of GPU memory.
I see, thank you.

6. Line 444: Remove the word „fully", as you still have some manual steps like defining the region of interest.
The region of interest is only manually defined once at the beginning, and it won't interrupt the pipeline for continuous producing termini. Also, the region of interest will be **automatically** defined by the intersection between the terminus and the flowline in the future. So, we still think our pipeline is fully automated.
OK.

7. Table 1 includes abbreviations that were not introduced.
The information in the last column is not necessary so it was removed.
Ok, thank you.

8. Figure S2: Please name the conditions in the figure's description as well, referencing (a) to (e).
The names of the conditions are included in the figure caption.

Thank you.

Response to Reviewer 2's comments

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.

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.

- 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 doublechecked 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 5×10^{-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 5×10^{-3}, 2×10^{-3}, and 1×10^{-3}, and choose 2×10^{-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-blindnesssimulator/ 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."*