General comments

The authors' response addresses the questions and concerns raised by both reviewers, and integrates much feedback into the revised manuscript. Such revisions include the incorporation of standardized error metrics (i.e., the median distance deviation between predicted and manually measured termini is ~79m, which is slightly better but on par with existing networks). Other addressed concerns include: reproducibility via provision of the training data; restructuring of the manuscript; clarifications/integration of the figures within the text; elaborations on the training process; and the large number of specific/minor comments.

Reasonable explanations are given for comments that are not fully addressed. However, there are concerns regarding the dataset itself. Otherwise, the manuscript passes general, technical, and presentational criteria to a satisfactory degree.

After review of the author's responses, as well as the changes to the revised manuscript, I can recommend that this submission should be subject to technical corrections before acceptance, at the editor's discretion.

Major comments

1. Regarding previous concerns raised about the quality of the dataset (Reviewer 2 comment 3): The current automated quality control method does not eliminate the majority of errors visible in the as of now final published AutoTerm dataset (Version 3). The specific comments below cover a non-exhaustive list of specific GIDs with errors to address, and potential quality control algorithms/solutions to implement that may help address them.

2. In conjunction with the suggested quality control measures suggested in the specific comments below, perhaps consider that a semi-automated approach is in order. Consider defining and validating several transects/flowlines (either automatically, or manually) for each glacier, then ensure all valid termini pass through all transects/centerlines, without making large spatio-temporal jumps (use Otsu's thresholding to remove time series outliers when measuring any particular termini's advance/retreat metrics). Ensure movement along multiple flowline/transects is relatively consistent (even though the termini will move differently along different transects, the same general trends of advance and retreat should be consistent to within some tolerance, which may be determined empirically/tuned manually).
- Otherwise, ensure by eye that for each glacier there are no visual artifacts or boundary issues (as in the above images, where the termini are cut off by domain boundary, and are in need of correction).
- For the sake of verification/validation, consider producing graphs for all glaciers like Figure 10 was produced for GID 164, but for movement along multiple flowline/transects as well.
- While implementing all of these suggestions may be out of scope for this study, integrating any of them would still be beneficial for the quality of the AutoTerm dataset and the cryosphere community as a whole.

3. If possible, it would also helpful to see the actual classifications output by the AutoTerm neural network, and see how it performs against images with issues like sea-ice mélange, ice tongues, clouds, debris, data gaps (from image boundaries or otherwise).
- Ultimately, the AutoTerm methodology must either successfully detect the glacial termini under those conditions, or flag the detections as incorrect, as it will otherwise result in an impact the

quality of the final dataset. As is, the dataset is of course still very valuable, but further validation will help to improve ease of use and accessibility.

4. Termpicks data provides a standard metadata format that would be good to adhere to, and would provide information such as source image IDs, should users seek to manually verify the termini picks. Fields like Error could also be renamed to Uncertainty/Variance that would better describe the measure, to be more in line with the changes in the manuscript. This may be out of scope and is not necessary, but would be valuable if provided.

Specific Comments

1. GID53 & 54 - These glaciers seem to overlap. Has GID54 not been properly isolated from GID53's domain? Additionally, there are some apparent image artifacts that have persisted through the termini extraction process (see concertation of termini boundaries along centerline – is there a physical explanation for this besides the image boundaries ending within the domain, and getting missed by the automated quality control algorithm?). Furthermore, the highlighted picks should have been removed given inspection of this output data. Perhaps consider removing termini that go beyond the glacial boundaries or intrude too deeply into the ice mask.

2. GID85 - Note the additional erroneous fronts. This data needs to be validated before release. Large amounts of fronts in the ocean and inland indicate potential issues with sea ice and/or clouds. Similar to the above, remove any termini that intrude too deeply into the ocean mask.


3. GID 44 & 45 - Perhaps implement an additional automatic quality control metric, deviation from closest/mean termini. Any termini which deviates a lot from the average/mean/nearest termini should be flagged for exclusion. Note also that the apparent Error (uncertainty) metadata is low, despite other valid termini having higher Error (uncertainty) values.
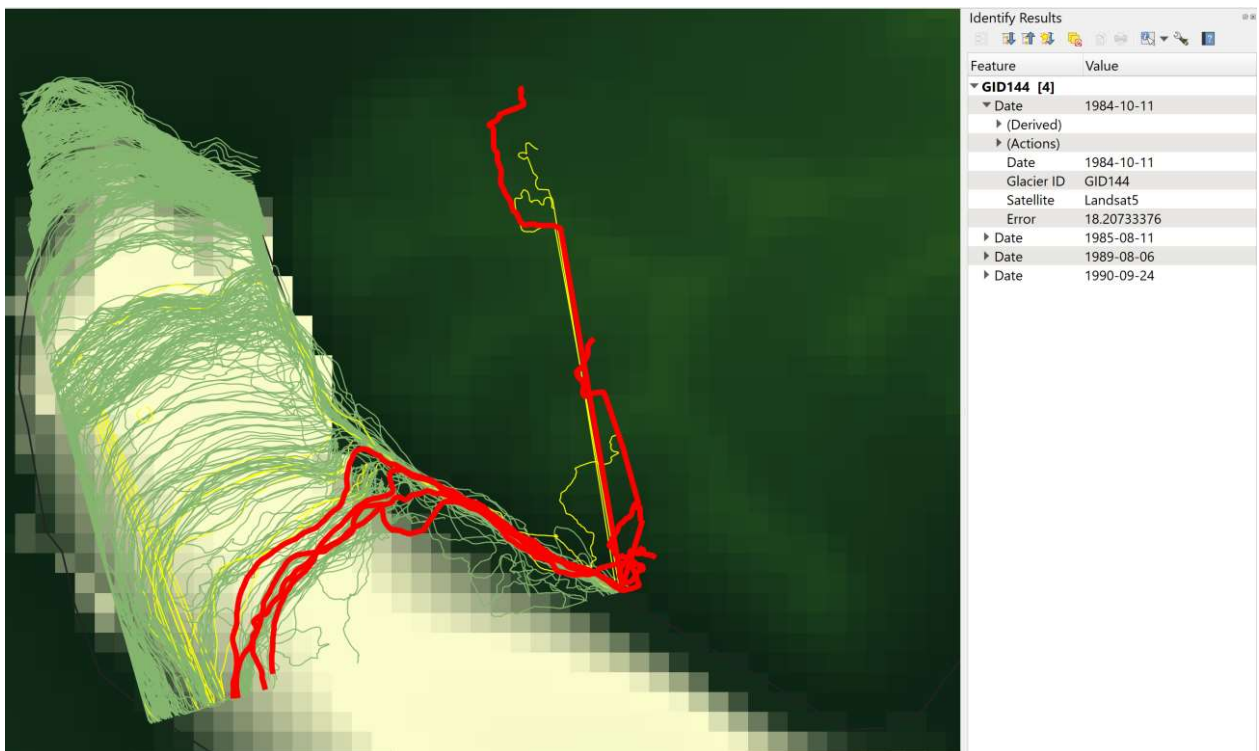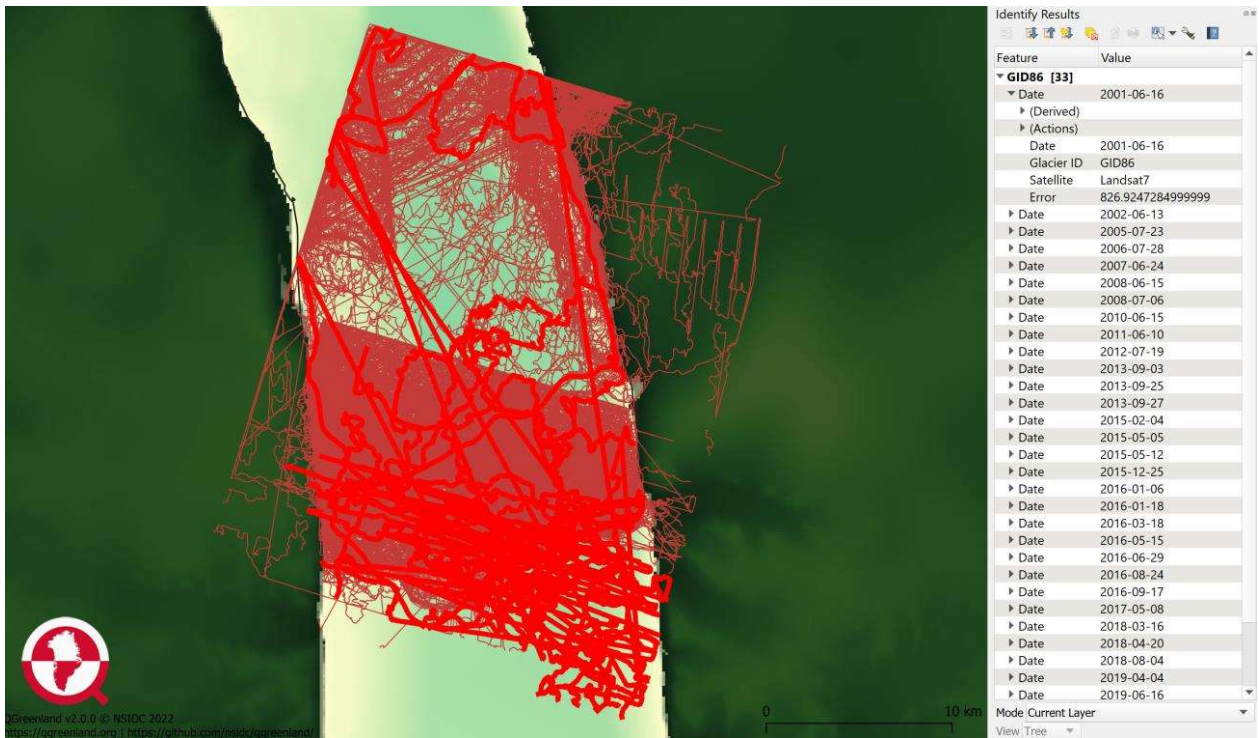
4. GID 86 & GID 144 - Another quality control metric could account for the distance between points in the detected line (to detect large straight jumps), as well as detecting the number/magnitude of large turns in the termini lines. Similar to the above, be sure to eliminate any detections that intrude too far into the grounded ice mask.
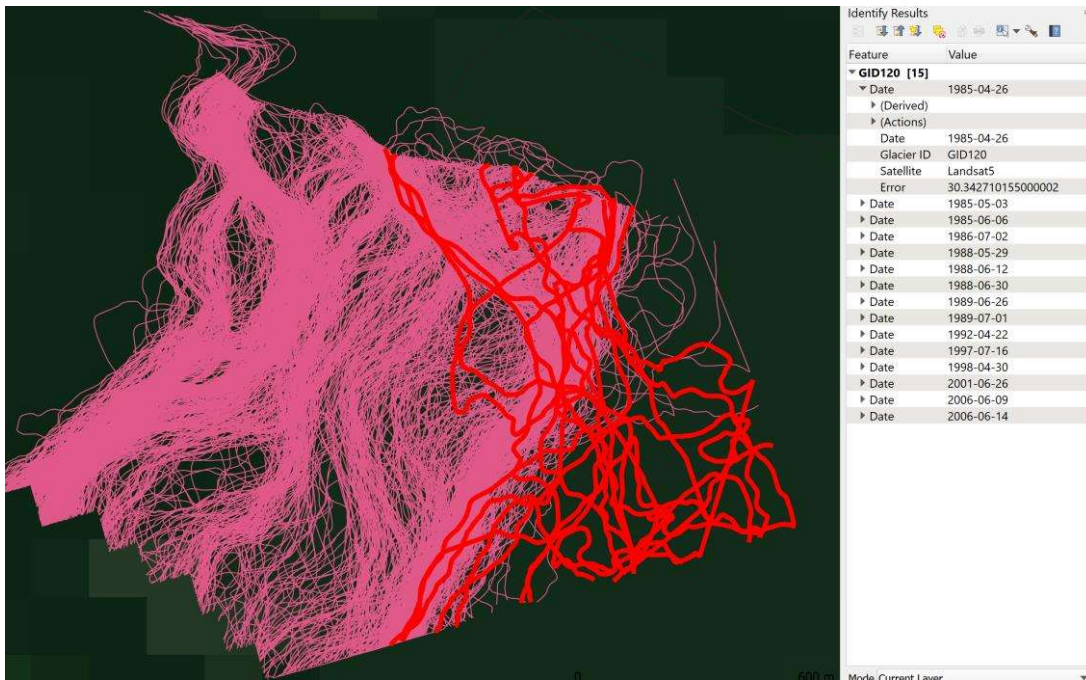
5. GID 120, GID150-153 - While this glacial domain looks fine for the majority of the termini, there are still issues with the domain boundary cutoff and termini that should be removed along those boundaries. Consider checking the proximity of the termini to the domain boundaries, and eliminating them if they have significant portions that are close to the edges & differ from the other termini.
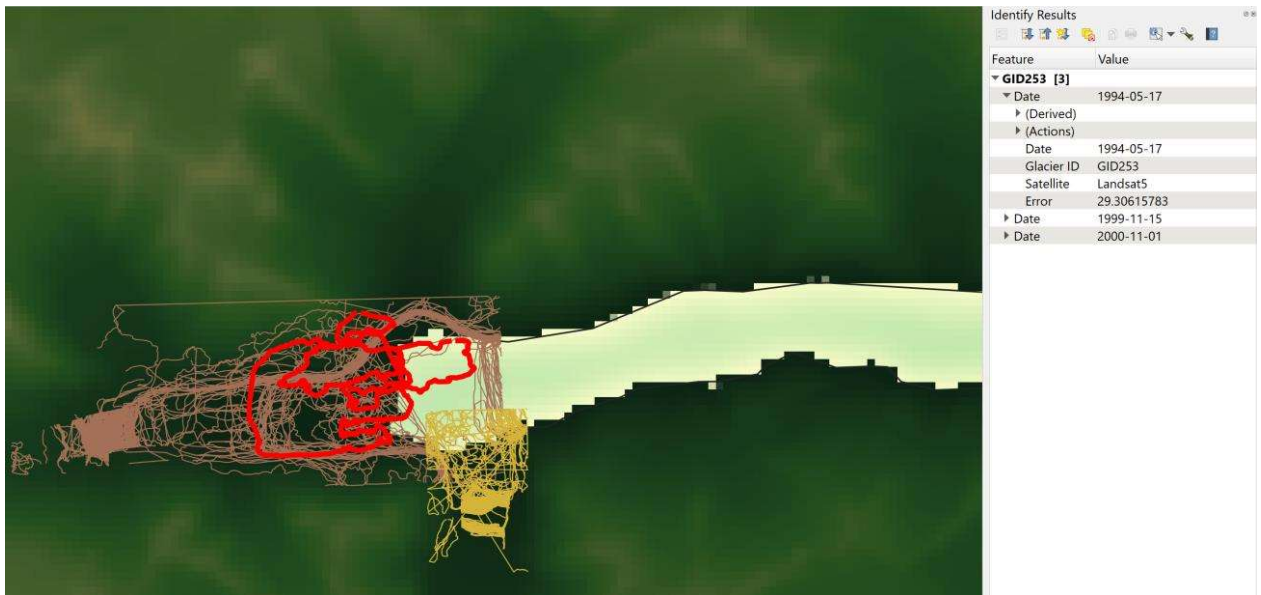
6. GID 262, 253, 210, 211, 214, 215, etc.- Overlapping and erroneous/boundary issue termini. While enlarging the domains and re-evaluating the termini may be out of scope, perhaps add commentary on potential solutions to this issue for future work (such as consolidating domains/GIDs and reprocessing, merging termini across GIDs from the existing dataset, or using the overlapping smaller inset/subset domains to cut away overlapping portions of the larger domain's termini).