

Dear Reviewer 1,

Thank you for your interest in our work, and your consideration and comments on our manuscript. We believe that your insight and our corresponding adjustments have strengthened the manuscript. In particular, we have restructured the manuscript to maintain more focus on the dataset, as you pointed out our focus was more dispersed among the dataset, software, and baseline classifier. In particular, we have removed the graphical user interface (GUI) section (formerly section 2.4.3), absorbing only a mention of it into the current section 2.4.3, only highlighting how the GUI supported multiple human labelers in creating this dataset. We have also trimmed down on the description of the sat-tile-stack software, as you suggest in your first major comment. For the baseline classifier, we believe that inclusion of the classifier earns its keep in demonstrating the dataset is “machine learning-ready”, which is the intention of the dataset and its description. However, we agree with the broad consensus across the reviewers, including your second major comment, that the classifier is over-described in the original manuscript. In response, we have trimmed its description, moved to the newly-named “Technical validation” section, and moved all details of the classification to Appendix B. Of note regarding the baseline classifier, we have removed the comparison between Dunmire et al., 2025 and the classifier results and have restricted ourselves to a train/val/test split all within the 2019 season so as to reduce data leakage between the two years, as brought to our attention by Reviewer 2. We include responses (black text) to each of your comments (blue text), both major and minor, below.

Sincerely,
Joshua Rines, on behalf of all coauthors

Major:

The manuscript describes multiple different pieces of software and datasets. A lot of space is dedicated to describing the sat-tile-stack software and I think this takes away from the main objective of the manuscript. My understanding is that this software is designed to be applicable for various uses outside of GrIS hydrology. Much of section 2 describes this software in ways that are not directly relevant to the supraglacial lake dataset, and I think this ends up being confusing and detracts from the paper's main objective. For instance, L196-206, L240, L243-245, L262-264, L397-406, L414-424 describe features specific to the software/GUI but are not necessarily relevant for the dataset presented here. The software sounds really interesting and useful, but perhaps there should be a separate dedicated manuscript for this software in a different journal (e.g. JOSS or equivalent). I was confused a bit when reading sections 2.2-2.4 because the text constantly flips between describing the software and describing how the dataset was created using the software.

Thank you for highlighting this. We agree that mixing the description of the dataset with the capabilities of sat-tile-stack software which are unused in the preparation of this specific dataset is unnecessary. We have remedied this by removing these instances from section 2.4.2, which you pointed out at L196-206, L240, L243-245, L262-264 from the original manuscript.

Regarding your comments about the GUI description in section 2.4.3 (L397-424 in the original manuscript), we believe that some description of the labeling process and the GUI software is necessary, as it provides necessary context for how the labels, which are a key component of this dataset, are made and shared among labelers. We appreciate your comment that the software is over-described in the original manuscript and have thus decided to not allocate an entire subsection to the GUI description (formerly section 2.4.3 in the original manuscript). Instead, we pared down the description of the software, focusing more on how each labeler creates labels using the software, and combine this slimmed-down description into what was formerly section 2.4.4, now 2.4.3 (L379-413), as the former 2.4.3 has been removed in line with this comment.

Similarly, I don't believe that the baseline deep learning classifier (section 3) should be covered in this paper, especially because it is mentioned in L500 that "the model, training code, split files, and trained weights will be released as part of a companion paper". Why is it described here if many of the details will be part of a separate paper? I think section 3 felt a bit haphazardly thrown in and it was too much to have descriptions for the software, dataset, and baseline classifier all in this manuscript.

Thank you for pointing out L500, which we agree should be, and now has been, removed. While we agree that the baseline deep learning classifier was covered in too much detail in the original manuscript, we disagree with the reviewer that it should be removed in its entirety. In light of comments indicating interest in the classifier by the other reviewers, and in particular as stated by reviewer #2 that the baseline classifier demonstrates the utility of the dataset, we believe the classifier earns its keep. However, we do agree with the point that there are too many details included in the manuscript, a perspective that is shared by reviewer #2 as well. In response, we have reduced the description of the classifier and its results to only cover the baseline classifier results for the combined dataset, whereas before we covered results for 'combined' and 'cross-year' experiments we now elect to simply demonstrate that the dataset is structured properly and is directly ingestible by a machine learning model. We believe this sufficiently highlights the utility of the dataset and its structure.

Finally, I question a bit the distinction between the "hydrofracture" and "moulin drainage" classes. To me, these are not "distinct" drainage mechanisms and have substantial overlap. Part of the stated reasoning for distinguishing drainage mechanism is for "resolving distinct physical processes with distinct consequences for ice dynamics..." (L51). Both processes will have similar consequences for ice dynamics as they both deliver meltwater to the ice sheet base, right? If these are kept as separate classes, there needs to be a clear reason why this distinction is important from a glacial hydrology perspective (i.e. why should we care if a lake drains via a newly formed moulin or an already existing moulin?)

While we agree with the reviewer that both of these drainage modes are modes by which surface melt water enters the ice sheet vertically, we disagree in that they are not distinct drainage mechanisms. In particular, many of the moulin drainage events we observed happen laterally away from the lake basin, and indeed are slow drainages. We believe that the rate at

which water reaches the bed is important (Hewitt 2013, Schoof 2010), and so they are distinct classes in the dataset. This is discussed in the introduction and in the labeling protocol as well. However, as mentioned above, we do agree that these could be considered similar enough to group together as (non-crevasse-field) englacial drainages. We thus report the results of the baseline classifier under a four class taxonomy which combines moulin and hydrofracture classes in Appendix B. Thank you for the inspiration to do so.

Minor:

L11: replace "cadence" with "temporal frequency" (perhaps consider doing this elsewhere too). I don't often think of cadence in this context.

Done.

L31: Remove "As dynamic as it is"

Done.

L65: I would remove "such as long short-term memory (LSTM) recurrent neural network architectures" because the final model used was not an LSTM architecture

Done.

L143 – Think of a different way to say "inter-rater" here. It becomes clear what you mean by this later on (when you describe the labeling process), but at this point in the paper it is confusing.

Done. We have replaced this with "validation of the labeling protocol across multiple human-expert labelers" (L130)

I am curious why Sentinel 1 imagery was not used. It's not always useful when the firm is fully saturated, but in some cases (especially in the ablation zone) it can help constrain the timing of lake drainages when optical satellite imagery is not available. I am curious if it could be another band in the dataset?

We agree that the inclusion of Sentinel 1 would be an interesting ablation test of the model. We do not consider it here for simplicity and to build up the model and dataset one step at a time, beginning with a single satellite source before extending to others. Other sources could include Landsat and Planet, both of which would have a different temporal frequency, and could also then help with both drainage type and timing. But we are working one step at a time.

Figure 1: Consider making the grey dots a different color. It's hard to distinguish from the rock/tundra areas.

Done.

L225: "arrays" → "array"

Done.

L358: How often are ties assigned?

Ties are not very common, with only 5 across the total dataset.

How often are there large gaps in available imagery? What is the average temporal frequency and how do you think this impacts classifications?

The temporal frequency is roughly 1-2 days in this region. We believe it does impact the classification, and it is an active area of exploration in our group to figure out how to mitigate this.

L471-480: I don't really understand this paragraph. Maybe a schematic would be helpful? Or maybe that's unnecessary but it would be helpful to think of a simpler way to describe this.

We have removed this paragraph as we have elected for the simplest demonstration of using the dataset in the model, which removes the cross-year and combined splits in favor of training, validation, and test being within one melt season, 2019.