

Review for “A benchmark deep learning dataset for the classification of supraglacial lake drainage mechanism across the central-west Greenland Ice Sheet”

Rines et al. 2026

First, I apologize for my delay with this revision – it’s been a hectic summer! This paper presents a software package for tiling satellite imagery, a dataset of supraglacial lakes in CW Greenland, classified by drainage mechanism, and a corresponding ML model trained on the benchmark dataset. Overall, the paper is well written, and I think the dataset will be very helpful for studying Greenland hydrology. However, I do have some major concerns about the contents and organization of the manuscript, outlined below:

- 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.
- 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.
- 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?)

Minor /Technical comments

- L11: replace “cadence” with “temporal frequency” (perhaps consider doing this elsewhere too). I don’t often think of cadence in this context.
- L31: Remove “As dynamic as it is”
- 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
- 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.
- I am curious why Sentinel 1 imagery was not used. It’s not always useful when the firn 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?
- Figure 1: Consider making the grey dots a different color. It’s hard to distinguish from the rock/tundra areas.
- L225: “arrays” → “array”
- L358: How often are ties assigned?
- How often are there large gaps in available imagery? What is the average temporal frequency and how do you think this impacts classifications?
- 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.

.