Review of ”A minimal machine learning glacier mass balance model ” by Marijn van der Meer and al.

This paper presents miniML-MB, a new machine learning (ML) model designed for estimating annual point surface mass balance (PMB) on glaciers, specifically tailored for situations with very limited datasets. By employing an XGBoost architecture, miniML-MB addresses the challenges associated with data scarcity in glaciological research, providing a data-driven tool to predict PMB at specific sites within the Swiss Alps. A significant feature of miniML-MB is its ability to discern influential climatic drivers of PMB, which the authors identified as mean air temperature (May–August) and total precipitation (October–February). The model demonstrated a high level of accuracy across multiple decades (1961–2021), achieving a mean absolute error (MAE) of 0.417 m w.e., and outperformed a traditional positive degree-day (PDD) model, which had an MAE of 0.541 m w.e.

A number of prognostic glacier models suffer from significant uncertainties in Surface Mass Balance (SMB) modeling. One of the most widely used models, the Positive Degree-Day (PDD) model (and its variations), fundamentally struggles with calibration issues: PDD parameters can vary considerably across different glaciers and regions. Typically, calibration is performed through basic regression techniques. The advent of machine learning (ML) techniques in glaciology, which offer advanced regression capabilities and greatly simplify implementation, is therefore pivotal in addressing these calibration challenges. Beside initial efforts by J. Bolibar and colleagues to model mass balance processes with ML, there has been a need to tackle the challenge of deriving a local ML-based SMB model, driven by the glacier modeling community's shift towards distributed modeling (as opposed to glacier-wide models). To my knowledge, this is the first paper to address this issue—a crucial step for the community given the demand for more accurate models. Overall, I enjoyed reading this paper; it is well-written, clear, original, and presents promising results.

I don't have any major concerns, but some comments, or suggestions that I hope will help the authors to improve their manuscript.

- You do not discuss the choice of architecture. I assume you must have tried a series of different ML models, and I think it would be valuable to share some of your experience with this process. Additionally, it would help to justify your final choice beyond simply opting for a lightweight model due to the limited data availability.
- The explainability analysis you conduct to identify the most influential predictors (i.e., summer temperature and winter precipitation, as expected) is definitely interesting. However, in a straightforward ML approach, I would have used a large number of candidate variables, allowing the model to select automatically those with high explanatory power while disregarding automatically those with little impact. Isn't one of the primary benefits of ML methods, as black boxes, their ability to

automatically assess the relevance of each input during training? In short, let's include everything; the irrelevant ones won't be used anyway. The fact that adding many predictors (e.g., n=24) appears to degrade the performance of miniML (Fig. 6) is surprising to me. Would this result differ for an ML model based on a neural network? (I must admit I am not very familiar with decision trees.)

- An important limitation of PDD models is the variability of PDD factors in the literature, necessitating sensitivity tests of these parameters in modeling applications, which can lead to significant variability in results. If I understand Fig A1 correctly, you show that, besides improving predictability over PDD, miniML reduces this variability significantly. In the context of using miniML for glacier evolution modeling, wouldn't this result from Fig A1 be a particularly strong advantage? I think it should be further highlighted.
- Another point not mentioned in the paper is the computational efficiency of the ML model, which is virtually instantaneous (l. 156). With the ice flow model component being parallelized on GPUs, the SMB model, such as PDD, can become the bottleneck, especially when applied to large grids, since PDD models with snow layer tracking require temporally sequential (and thus non-parallelizable) sub-steps, see (Jouvet, 2022). This is an advantage of your approach that you should emphasize.
- You have chosen MAE as the loss function, which is more tolerant of outliers. Out of curiosity, did you try using Mean Squared Error?
- Section 5.4: I must admit I was somewhat skeptical about retraining on 2022 to predict 2023, as this involves very little data. You may possibly frame this experiment and its results with more caution. What are the risks of working with so little data? What if you were switching 2022 and 2023?
- l. 401: "we used the sum of daily average temperatures ..." To my understanding, all the thing of ML is to do this for you; not sure I see the point of reporting this experiment here.
- It would be beneficial to have a "prospective" section in the conclusions. The paper contains important results, and I would like to understand the main challenges to make this "minimal" ML model generalizable so that it can be embedded in a glacier evolution model (GEM). What are the next steps in this regard? How strong is the "data bottleneck"? I would appreciate some insights on both the model's generalizability and its direct applicability within a GEM. Even though the model is customized, it would be interesting to see if it could be compatible with GEMs. Running an ensemble based on the 28 miniML models could possibly yield GEM results with reduced uncertainty compared to using a PDD model with factors varying within literature-based ranges?

And a few minor comments:

- l 35-38 : Consider broadening the scope of the literature.
- Section 2.2, title : Capitalize the first letter : "Point ..."
- l 161 : "an ML", remove 'n'