Manuscript Review for 'GlaUnTI: A hybrid physics–machine learning model enables transferable glacier surface mass balance estimation'

April 12, 2026

The manuscript 'GlaUnTI: A hybrid physics–machine learning model enables transferable glacier surface mass balance estimation' is a very interesting read. The authors have tested the performance of (i) a differentiable temperature index model, (ii) a pure data-driven gated recurrent unit model, and (iii) a modified temperature index model with a neural network corrector. Two settings of (iii) are used, which are with and without a glacier facies map. This is quite an innovative approach, which attempts to allow the interpretability of temperature index models while allowing non-linear interactions between the meteorological variables at the grid level and the glacier level. The domain of hybrid machine learning and temperature index models is quite novel in studying glacier mass changes and is well aligned with the scope of this journal. That being said, I have a few suggestions that can strengthen the contributions and impact of this study. I divide my comments into Major and Minor Comments.

Major Comments

The main selling points in this manuscript are as follows:

1. Development of the differentiable form of the temperature index model
2. Development of a machine learning based corrector to temperature index models
3. Spatial transferability of the models developed

However, some aspects regarding transferability and performance are not yet convincingly supported.

Transferability

Regarding the spatial transferability, the manuscript starts compellingly with descriptions of how differentiable models will aid transferability, but does not provide systematic quantification of it in the results. It would strengthen the paper to show that the differentiable TI model achieves comparable accuracy to conventionally calibrated (e.g., grid-search) TI

models, ensuring that the smoothing approximations (softplus, sigmoid) do not degrade performance. The authors have described the advantages of the fully differentiable approach to include interpretability and transferability, which will be well justified with this comparison.

Validation Strategy

While the authors define six stratified folds, performance is reported for only one designated test fold. A full cross-validation framework where each fold serves as the test set in turn would yield more reliable and generalisable performance estimates, particularly given the limited test set size of 13 glaciers.

Clarification on modelling framework

Further clarification is needed on both the model inputs and the SMB datasets used. It is unclear whether inputs are provided at daily time steps with the loss computed over aggregated annual or seasonal periods, and whether this aggregation follows hydrological year boundaries. Additionally, it is not specified how glacier-wide SMB estimates were derived — whether through spatial interpolation of point mass balance measurements or from geodetic estimates — nor how glacier-wide predictions are obtained from the models.

Glacier Facies Map

The descriptions of the glacier facies map included as a predictor take the focus away from the primary selling points and don't add much value to the manuscript. The fewer data points available for the validation constrain robust statistical analysis. Further, one can argue that it is a rather arbitrary inclusion, with more physically relevant ones possible, such as shortwave radiation, albedo, and wind.

Minor Comments

This section is divided into subsections that distinguish between suggestions in technical content and those associated with minor rephrasing and flow.

Sentences that can be rephrased for clarity and flow

1. Line 14-18: 'Including glacier facies... SMB trajectories.' These sentences are a bit abrupt and disconnected. The flow can be improved for readability.

2. Line 46 onwards: For the sake of completeness, it might be worth talking about how geodetic datasets are used for calibration as well.

3. Lack of clarity and consistency in equation symbols. For example, cell superscript is used in certain equations like eq (5) and skipped in others. Perhaps it is not essential as i and j subscripts represent the cell uniquely? Is A_{ij} coming from the elevation grids? m and n are not defined in eq (17)

4. Line 133: Should be 30th September, not 31st.

Specific Technical Recommendations

1. Abstract: The ability of the models to be spatially transferable is not represented clearly here.
2. Line 9-10: What do you mean by heterogeneous regions?
3. Line 94-98: Include the time periods during which the other studies were carried out as well.
4. Line 165: How was the precipitation downscaling performed?
5. Line 166: What are the uncertainties associated with Z? What are the sensitivities to the annual elevation maps for each of the models?
6. Glacier facies map: It is not clear what the classes generated by the algorithm is? What prediction algorithm was used to generate it?
7. What is the procedure by which the glacier-wide mass balance was estimated?
8. Line 215 to 220: What's the basis for this?
9. Could authors confirm that normalisation was performed before training of the ML models?
10. It's unclear what the training process for the GRU is compared to the TI models. Is it daily inputs with losses computed the same way over time?
11. Equation (18): Can the coefficient of determination be used in addition to or in place of Pearson's r? The Pearson r is a suitable metric if you aim to evaluate the linear association between variables. However, as a goodness-of-fit test, the coefficient of determination is the preferred metric as it explains how much of the variance in the observations (or reference set) is being represented by the model. Note that the coefficient of determination is not always the square of Pearson's r and can take negative values when the model performs worse than a model that naively predicts the mean.
12. Line 332-335: Might be worth checking if GPU is being used during the training. For 50 epochs on the small parameter size and dataset size, training is typically much faster. If GPUs are being used, check if jit pipeline breaks occur during training.
13. Note on the number of epochs trained. 50-100 epochs used for training, as depicted in Table 2, and grid search training with a couple of epochs, as depicted in line 312, is generally not sufficient for most machine learning training, which can explain the poorer performance of GRU. This can be checked by plotting the loss vs epoch curve for training and validation. If both the training and validation losses continue to decrease, then training is not complete. If limited epochs were selected due to the time taken for training, see the previous comment. This may suggest that GPU acceleration or JIT compilation is not being fully utilised

14. What do the model predictions look like over the glaciers? Can this be represented as a figure over a few of the test glaciers?

15. I also recommend that the integrated gradients-based explainability described in Appendix A2 can be included as a part of the main manuscript to highlight the advantages of this differential form of the temperature index model.

***