# Review of "Interpretable Deep Learning for Glacier Mass Balance: Temporal Attention Patterns in Central Asia", by Zafar Avzalshoev and Pang-jo Chun

EGUsphere

Jordi Bolibar and Kamilla H. Sjursen

## 1 General comments

Avzalshoev and Chun present a study where they apply different machine learning (ML) methods, including Temporal Fusion Transformers (TFT V2), to simulate surface mass balance of glaciers in Central Asia. While the paper introduces some interesting concepts, like new methods for modelling surface mass balance or attempting to relate it to glacier-related hazards, the methods have fundamental flaws that discard the validity of the results. Importantly, the methodology and results do not support the objectives and conclusions stated in the paper. In our opinion, the problems are so fundamental that a complete revision of the study is necessary for the results and arguments to be valid. Therefore, we recommend rejecting the paper and encourage the authors to re-submit it when the fundamental problems in the methods have been addressed.

Here we will go over the main reasons why we recommend a rejection and what should be improved in the methods in order to draw scientific conclusions from them. We will address these points in different general comments (GCs).

### 1.1 GC1: Training a machine learning model on the outputs of another model

The first fundamental flaw of this paper is the fact that the authors claim that they train their ML models on "geodetic mass balance data" (e.g. L7-8, L74, L94-96, or "measurements" on L349). However, these training data are in fact the output of a temperature-index model calibrated on geodetic mass balance and snowlines (Barandun et al., 2021). This strongly influences the conclusion that can be drawn from their work and creates a misalignment between the research objectives, methodology and analysis. Training an ML model on the outputs of another model can indeed make sense in some cases, such as when one intends to emulate an expensive physical model at a fraction of the cost by "compressing" it using statistical learning. That is not the case here. Barandun et al. (2021) already performed the task that the authors intend to do in this study, that is, calibrating a surface mass balance model based on different

types of observations to reconstruct the full surface mass balance series of the whole region for a period of time. The authors here are emulating a temperature-index model, which is a very simple model with only two parameters, by using a rather complex deep learning model with thousands of parameters. This doesn't make sense from a scientific point of view, because the primary reason one would emulate a physical model is to make it faster, which is not the case here. For the study to make scientific sense and to provide added value, the authors could instead try to replicate what Barandun et al. (2021) did in their study, and try to improve the accuracy of a model by training on the same observations (i.e. the snowlines and the geodetic mass balance) as target data. That would provide an added value, and results could be compared to the ones of a temperature-index model to see if any extra physical processes or nonlinearities are being captured by the more complex deep learning model. With the use of modelled data it is also unclear if this approach is feasible for real-world observational data, which is noisy and sparse. The reported performance is not representative for such data and it is unclear whether the insights gained from the TFT actually provide new insights or just an interpretation of the "inner workings" of the temperature-index model.

In light of the current goals of the paper, we recommend the authors to rethink their study and to train the model **only** based on observations, and not based on the outputs of other models.

### 1.2 GC2: Fundamental statistical issues in the training

The second fundamental flaw we see in this paper concerns several statistical problems related to the training of the model, which strongly damage its validity and performance. The main problems are related with how the dataset is divided into train, validation and test, which crucially determines what the model will learn. The authors mention that their intention is to make the model capable of making future predictions in time, so they perform a temporal split. The fact that they chose to use just a single year for validation and a single year for test is clearly not enough. Glacier surface mass balance is characterized by interannual variability linked to the forcing climate. If the performance of the model is assessed on a single year (both for validation for training and test for independent performance assessment) it is impossible to evaluate if the model has learnt to capture that interannual variability. The claims that the model demonstrates accuracy over different time periods and climatic conditions and is robust for long-term predictions ( Section 3.5 and Fig. 10) are therefore not supported by the test performance. Moreover, if the chosen hydrological year happens to have a climate very close to the median, the model model performance will come across as much better than if that year represents an outlier with extreme conditions.

This has several implications for the study. First, Figure 10 clearly shows that the model is not working, with poor performance for the two years that are evaluated (validation and test). This shows that metrics alone do not provide a complete picture, and the model likely does not capture the right physics. Additionally, the benefits of moving to larger and more complex models such as TFTs are reported only in terms of $R^2$, which is generally not a good metric for these cases. This is even more problematic taking into account that the metrics are computed based on a single year. This means that $R^2$ is just capturing the spatial variability between the surface mass balance of glaciers across regions, rather than the interannual variability, which

was reported to be the main goal of the modelling. Considering the other reported metrics (i.e. RMSE and MAE), we can see almost no added value in using these more complex models. This indicates that there is limited learning, perhaps because the training procedure is flawed. It would be interesting to see the training vs validation errors to assess if the model(s) are overfitting or underfitting.

Finally, the claims about model interpretability linked to temporal attention are not true. The methods used to assess the importance or contribution of features to given years or the global performance of the model can be assessed for any of the other models, using tools such as SHAP values. Moreover, tree-based methods are generally known to be much more interpretable than neural networks, and the same analysis of understanding the importance of individual months or features could also be obtained with the right design of the model and the training process.

### 1.3 GC3: Link to glacier hazards

The connection between the results of the ML models and glacier hazards has limited scientific explanation or statistical grounds, such that the strong statements regarding this link (e.g. Section 4, L12-14) are not sufficiently supported. Importantly, the statistical results presented are not sufficiently explained and based on unpublished data without reference (Section 3.6). The ML model is meant to predict glacier surface mass balance, which per se is not necessarily directly linked to glacier-related hazards. The authors use the attention weights from the model predicting surface mass balance to argue that they can also indicate glacier-related hazards. This connection is weak and not sufficiently supported by evidence. The attention weights are only indications of the contribution or importance of those months to explain/predict surface mass balance, not hazards.

### 1.4 GC4: Overall quality of the figures and references

A bit less relevant, but also important, is the overall quality of the figures. Many of them are difficult to read, with panels overlapping text and other basic issues which hamper their readability and the communication of the results. Moreover, the authors mention that there is supposed to be a scatter plot in Figure 5, which is not the case. A direct comparison with the ground truth data is never shown.

Several statements lack references (e.g. L278-279, L333-335 and L338-340) and the discussion section fails to relate the findings of the study to the current state of research (a single reference is mentioned in the six page discussion) and lacks a discussion of the limitations of the approach. In some places references do not support the statements (e.g. it is not clear why Huss et al. (2008) and Rastner et al. (2019) are referenced on L42 for geodetic mass balance, and why Farinotti et al. (2019) is referenced as an example of machine learning to estimate ice thickness on L48?).

### 1.5 GC5: Closed code

Finally, the authors do not seem to be willing to open-source their code. We understand that they might not want to share it until the paper is published, but open-sourcing scientific work,

including code, should be a must.