Review of

„Interpretable Deep Learning for Glacier Mass Balance: Temporal Attention Patterns in Central Asia“

by Zafar Avzalshoev and Pang-jo Chun

submitted to *The Cryosphere*

Manuscript summary

The manuscript describes the usage of machine learning (ML) models on glacier mass balance in Central Asia. The introduction of Temporal Fusion Transformers is investigated against more simple ML approaches with the aim of improving glacier mass balance predictions, as well as identifying the specific role of certain environmental conditions on the annual glacier mass balance. A freely available data set of annual glacier mass balances was used in combination with monthly meteorological information from ERA5 and glacier specific geographic information from RGI7. The mass balance data set spanning from 2000 to 2028 was split into a training subset (2000-2016) a validation subset (2017) and a prediction subset (2018). Based on the correlation statistics the quality of the different ML applications was evaluated for predicting future glacier mass balance, as well as providing information about the governing processes (“temporal importance patterns”).

General comments

ML approaches are increasingly used in glaciology for different applications, from feature mapping and mass balance estimations to speeding-up ice dynamic glacier models. Especially if large samples need to be analysed, ML methods might be helpful to detect patterns and improve the process chain. Also for the prediction of glacier mass balance ML techniques have a strong potential to handle large data sets efficiently. This study aims to apply advanced ML techniques on such a large data set of glacier mass balance series in Central Asia. Unfortunately the study contains a series of flaws, wrong assumptions, unclear descriptions and obvious confusions. Because already the basic input data and the accompanied test setup are flawed, I suggest to reject this manuscript to allow a fundamental revision.

Barandun et al. (2021) used volume change time series from 1995 glaciers in the Tian Shan and Pamir mountains as calibration for their simple degree day model for reconstructing mass balance time series. However, they only used 1222 glacier time series in the end, due to constraints in their quality assessment, for a detailed analysis of spatial and temporal variability of glacier mass balances in this region. Provided are annual data sets for these glaciers for the period 2000-2018. This results in 23218 data values. Avzalshoev and Chun state that the use a total of 4 659 264 monthly samples from Barandun et al. (2021). I wonder which data they used.

Unfortunately, there is an obvious misunderstanding that the Barandun et al. (2021) data set consists of annual mass balance data derived from altimetry and photogrammetry, which is not the case. The data are the product of a simple model, calibrated by snow line data and geodetic volume differences. These data, therefore, are themselves reconstructed by using meteorological input from ERA5. Applying an ML routine to reconstruct a mass balance series, which is based on a simple degree day model and ERA5 input by using ERA5 data will provide no additional information at all, besides the ability of the ML routine to mimic a given degree day model. This approach is per se not suitable to investigate any parameters related to real glacier response to meteorological conditions.

Even though the manuscript looks rather well written at the first glimpse, it turns out that it contains a number of flaws at a closer look. Figures are not well described, while the captions are in many cases not related to the figures. E.g. there is no colour coding of the mountain ranges in Fig. 1, Fig. 2 shows no a), b), c) and d), Fig. 4 shows something completely different as the caption suggests, Fig. 5 should be a scatter plot according to the text, while the caption does again not fit to the data shown, Fig. 6-9 also show wrong captions. There are many more inconsistencies, which I can provide to the authors upon request. However, the manuscript needs to be completely rewritten in my opinion which renders it unnecessary to work on such details.

Regarding the scientific value of the manuscript Fig. 10 demonstrates that there likely exists a fundamental misunderstanding of the authors regarding the suitability of models o glaciological science. In line 353 ff. the authors state: “The annual comparison reveals excellent predictive performance with a mean accuracy of 92.6% for the test period (2017-2018).”, which seems a purely model generated metric. The overestimation of the mass loss in Fig. 10 with regards to the simulated mass loss of Barandun et al. (2021) is about 60% which is a poor result and not an excellent predictive performance as stated by the authors. The application of the model to the prediction of one mass balance year only makes the entire discussion of the model performance, predictive processes etc. highly unreliable.

I stop here with my review, because the problems in this manuscript are so serious, that it requires a full reconsideration by the authors.