Review of ”Modelling the Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques” by Anilkumar et al.

EGUsphere

1 General comments

Anilkumar and colleagues present a study in which they use multiple machine learning methods to model point glacier mass balance for glaciers in the European Alps. This study is timely, thorough and provides new interesting insights on the use of machine learning to model glacier mass balance. As the authors explain, it provides the next logical step to the mass balance machine learning modelling literature: tackling point mass balance and using other methods than neural networks. Moreover, a recent study in the machine learning community demonstrated that for tabular data (like the one is normally used for glacier mass balance), tree-based models still outperform neural networks for various sizes of datasets[1]. This study corroborates those findings and provides new clues on the best way to model glacier mass balance using machine learning. For all this, I believe it represents a valuable contribution to the community.

Without taking away any of the merits of the study, I still believe there are multiple aspects of the study that could be improved in order to make the results more solid and easily understandable. For this, I will address some of them in the general comments (GC) section, and then I will provide detailed comments for different aspects throughout the text.

1.1 GC1: Separation of train, validation and test datasets

In my oppinion, the main weakness of the study right now is the way the cross-validation has been performed. For what I understood, the authors chose a classic 70% training - 30% test split. But some confusion remains regarding the wording, since the authors sometimes say that they use the test dataset for hyperparameter selection. I have two main issues with this:

1) Have you used the test dataset for anything than just assessing the final performance of the model? Hyperparameter selection should be done only with the validation dataset (i.e. using cross-validation in the training dataset). Using the test dataset for hyperparameter selection is considered a bad practice and will result in a clear model overfitting. Please confirm this and make the necessary changes if otherwise these necessary guidelines have not been followed.

2) Why have you chosen a 3-fold cross-validation for hyperparameter selection? This choice seems extremely arbitrary, and despite being a rather small number (the rule of thumb is more like 5 to 10), it is particularly bothering because it probably implies that the folds have been randomly selected. This means that there is most likely a lot of leaked information in the train/validation folds, since it is quite likely that there are point mass balance data for a same glacier both in the train and validation folds, even for the same years. This information leakage makes the machine learning methods overfit, and could explain the reason why the authors have detected some potential overfitting.

When working with spatiotemporal data, it is essential to respect the spatiotemporal structures in the data (see Roberts et al., 2017[2] for a detailed explanation). This means, that folds should be designed in a way that they correctly separate the spatiotemporal instances that one is trying to model. First, the authors should determine if they aim at simulating point mass balance for unseen glaciers, unseen years, or both at the same time. Once this has been clarified, different strategies should be applied in choosing the folds, namely Leave-One-Group-Out, in order to ensure that there is no overlap in information between train and validation folds. This implies using cross-validation techniques such as Leave-One-Glacier-Out (or multiple glaciers), or Leave-One-Year-Out (or some years). A combination of both can also be used, which is probably what the authors want here. This is clearly explained in Roberts et al. (2017), and it was implemented for glacier-wide mass balance in Bolibar et al. (2020)[3].

I would ask the authors the revise their cross-validation methodology, and to try to design a strategy and clearly presented in a way that it avoids information leakage between train and validation folds. This separation strategy should also be applied to the test dataset, to avoid any overlap in terms of glaciers and years between train and test.

I have seen the authors have chosen to normalize input data between 0 and 1. Have you tried using other types of normalization such as the StandardScaler from Scikit-learn (i.e. substracting the mean and scaling to unit variance)?

Another aspect that would improve the intepretation of the results would be to understand how the errors relate to the target data. Right now, MSE are given for each model in mm w.e./yr. Could you please add a new figure with a histogram of the distribution of the point mass balance data from FoG? This would help understand what is the range of mass balance values and how those relate to the reported errors of the ML models. Having errors of 750 mm w.e./yr is not the same for a region with average MB rates of 100 mm w.e./yr than for regions with MB rates over a meter.

1.2 GC2: Design of the variable training dataset experiment

This is an aspect I particularly appreciated about this study. Such an experiment is very interesting to researchers in the field, since it gives important clues on which machine learning method might be most suitable for each case. Nonetheless, if I understood correctly the experiment design, I think that keeping the 30% test dataset constant and changing the size of the training dataset is not the best way to do this.

I believe that instead the total size of the full dataset (i.e. train + validation) should be changed, in order to respect the 30-70% ratio between train and test. Otherwise, adding new data will produce a different result depending on the correlation between those data points and the ones in the test dataset. This is particularly true in the context of the current (lack of) block cross-validation (see GC1). Since the authors have not correctly separated glaciers and years between the train, validation, and test datasets, this effects will be even more enhanced.

Changing this should be rather straightforward, and would provide more reliable results to this interest experiment.

1.3 GC3: Use of climate data from ERA5-Land

One aspect that is not clear in the manuscript is how the climate data from ERA5 is used in the machine learning models. Since the authors are modelling point mass balance on glaciers, which are located on highly complex terrain, ERA5 is know to not capture well complex topography due to its coarse spatial resolution. It is unclear if the raw information from ERA5 has been used or if any downscaling or preprocessing has been performed.

Have you performed any correction on air temperature and precipitation to adjust to the glacier's altitude? How do you distinguish the different points in a glacier? For small European glaciers all of them probably fall inside the same ERA5 grid cell. If you don't perform any correction to temperature, how can actually extract different climate information for each mass balance point? Please explain this in more detail

These elements will also determine how much you can interpret the feature importance from a physical point of view. It would be interesting to bear in mind the limitations of the input climate dataset when interpreting each one of the machine learning models.

1.4 GC4: Lack of perspectives

This study introduces new methods, but offers almost no perspectives on what is the reason of their success and which new possibilities are opened by these. I would appreciate adding a section in which these aspects are discussed, and where the authors suggest the next steps, the main potential future bottlenecks, and what are the greatest opportunities following this study. Applying this to even more different glaciological regions will be challenging, especially in terms of cross-validation and hyperparameter tuning. How would you face those problems? Is there enough data available to apply this at a global scale? Answering such questions could be very useful for the community.

2 Specific comments

- **Title** I believe the title would sound better as "Modelling Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques".
- **L14-15** I'd rather present the RMSE (or MSE) in the abstract than the r^2, since it provides more information.
- **L35-36** I would also add the great number of parameters to calibrate.
- **L38-39** I would also point out the fact that for simulations over large temporal periods, temperature-index models (i.e. degree-day factors) are prone to be oversensitive to climatic changes[4].
- **L57** For me the sentence would read better as "and a nonlinear neural network..."

- **L59** This sentence is confusing. Artificial neural networks ARE machine learning models. I would reformulate, as you do in the abstract, to "have used the full diversity of different types of machine learning methods"

- **L68** Why use a linear regression example after mentioning NNs?

- **L98** I wouldn't call this training labels. This is a jargon more related to classification problems. I would just call them target data or reference data.

- **L100** Same with "labels".

- **L101** Regarding the parameters: that's the case for the NN only, right? Tree-based models don't really have parameters, mostly just hyperparameters to be tuned. Make sure that you really mean parameters and not hyperparameters.

- **L104** This would read better as "is a decision (regression or classification) tree".

- **L106** This would read better as "To illustrate this".

- **L122** The subject of the sentence is missing (i.e. "a neural network").

- **L125-126** "Nonlinearity" should be "nonlinearities".

- **L129** Same with "labelled data" and "labels".

- **L134** Why only annual mass balance observations and not seasonal? This is something that surprised me quite a lot, since dividing mass balance into accumulation and ablation season can definitely help to better calibrate melt vs accumulation features.

- **L143-144** Following GC3, please develop these aspects to make them clearer.

- **L146** As per GC1, please explain this better and make the corresponding changes.

- **L150** It's not the parameters which are tuned (e.g. the NNs weights), it's the hyperparameters. It's important not to confuse both.

- **L152** This is in fact cross-validation. So instead of just using a subset for validation you divide into folds.

- **L154** Following GC2, please better explain this and make the necessary changes.

- **L156** Do you mean the validation score? The test score can only be accessed once at the end, once you have selected the hyperparameters. Using the test dataset for selecting hyperparameters is a bad practice.

- **L157-158** What is the advantage of doing this? An advantage of the RMSE is that it keeps the units and it is therefore interpretable in terms of magnitude.

- **L168** Did you try any other activation functions? ReLu is known for vanishing. Did you try other improved activation functions such as Leaky ReLu or softplus?

- **L202** RMSE: Once the acronym has been introduced, you should use it to keep things brief.

- **L209** Please see GC2.

- **L284** Why are all the test performances given in these sections higher than the ones reported in the figures? Could you please explain and fix if this is an issue?

- **L320-321** This sentence is not clear, and seems somewhat contradictory. Could you please elaborate?

- **L324** I think you mean hyperparameters here.

- **L325** Please, revise the concepts of parameters and hyperparameters and make sure to use them correctly throughout the text.

- **L336** Tree-based models also provide a feature importance analysis in order to understand the most important input features. Did you compare the outpout of these with the permutation analyses? Are the results similar?

- **L352-353** This sentence is not clear, and seems somewhat contradictory. If you say that albedo is very important for the ablation season, why do you then say that is not important? Surface albedo is critical in summer, since the transition between snow, firn and ice drives important nonlinear spatial responses in terms of melt patters and the total annual mass balance.

- **L362-364** This is a very interesting finding and in line with recent studies from the machine learning community regarding ML for tabular data (cite[1])

- **L376-377** "We suggest the use of kernel-based model in such situations": This sentence appears out of the blue and it is not clear. Please merge with the following one to make your point clear.

- **Table 1** Please clearly separate each line in order to make it easy to see which hyperparameters are related to which model.

- **Figure 1** Here you should mention the validation dataset and call it 3-fold cross-validation, not validation.

- **Figure 2** Why are the errors reported here substantially lower than the ones reported in the text? Are you talking about different errors? Also, please report the units of the error in the vertical axis.

- **Figure 4** Please use target or reference data instead of "labelled". Why are the errors in here different than in Fig. 2?

- **Figure 5-7** These figures are not that interesting by themselves. I would either merge them in a single figure or move them to a supplementary material.

- **Figure 8** Instead of giving the abbreviations in the supplementary material, I think it would be better for the reader to have them in the legend. This should take that much space and it would increase readability.

References

1. Grinsztajn, L., Oyallon, E. & Varoquaux, G. *Why do tree-based models still outperform deep learning on tabular data?* 2022. https://arxiv.org/abs/2207.08815.
2. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. en. *Ecography* **40,** 913–929. ISSN: 09067590. http://doi.wiley.com/10.1111/ecog.02881 (2019) (Aug. 2017).
3. Bolibar, J. *et al.* Deep learning applied to glacier evolution modelling. en. *The Cryosphere* **14,** 565–584. ISSN: 1994-0424. https://www.the-cryosphere.net/14/565/2020/ (2020) (Feb. 2020).
4. Ismail, M. F., Bogacki, W., Disse, M., Schäfer, M. & Kirschbauer, L. Estimating degree-day factors based on energy flux components. *The Cryosphere Discussions* **2022,** 1–40. https://tc.copernicus.org/preprints/tc-2022-64/ (2022).