Modelling the Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques

Ritu Anilkumar, Rishikesh Bharti, Dibyajyoti Chutia, Shiv Prasad Aggarwal
Correspondance: ritu.anilkumar@nesac.gov.in

We are grateful for the reviewer's critical and insightful feedback on manuscript number **egusphere-2022-1076**: 'Modelling the Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques'. The point-by-point response to the comments is provided below. The comments by the reviewer are quoted in *a black font colour and italicised font style*. The author's response is in blue font colour and normal font style. Text quoted from the revised manuscript is ***blue font colour and bold italicized font style***.

1 Summary and General Comments:

In this paper, the capabilities of different machine learning (ML) models in predicting point glacier mass balance are explored. The used data is composed of monthly meteorological data from ERA5- Land together with direct mass balance measurements in Central Europe from the Fluctuations of Glaciers database. The study is an important next step to explore which ML models are most suitable for applications of mass balance estimates. Further, they assess the data required for the different models and the importance of each meteorological variable. Both are very interesting and important questions for the potential future use of ML models in this field, also in light of increasing data availability in the future.

The study is well designed, but I think parts could be improved to make the results more solid and the manuscript easier to follow by the readers. I have divided my proposed changes into General Comments and specific/Line by Line comments.

COMMENT GC1: *I think it would be good to give more information on the values of the used mass balance observations (e.g. How is the distribution? Are they located in the yearly ablation regions of glaciers or also some in the accumulation regions?). Also, why are only annual mass balance observations used and no seasonal ones? This probably could improve the analysis of Feature Importance performed separately for accumulation and ablation months.*

Thank you for this suggestion. We agree that some more information pertaining to the point mass balance values considered will be beneficial for a complete interpretation of our findings. We will include a figure representing the distribution of the mass balance measurements with a histogram of measurement values to the manuscript.

Regarding the consideration of only annual mass balance observations as opposed to seasonal observations, our decision was purely a consequence of the availability of data. The database of point glacier mass balance observations contains separate entries for annual mass balance observations and seasonal mass balance. For example, we have 9595 points using annual mass balance observations after 1950. For accumulation season, 3281 points are available and for ablation season only 1783 points are available, all of which do not overlap with the accumulation season measurements. While we do agree with the reviewer that separation of mass balance can help bring out the features associated with the accumulation and ablation, a combined measurement of summer, winter and annual point mass balance for the same location was not available using the existing database.

COMMENT GC2: *If just the raw ERA5-Land data is used as input it is probably hard to asses feature importance due to the very complex topography which is poorly represented. In general, how did you deal with the downscaling of the meteorological data to the glacier location? In particular, how do you deal with the height difference between the ERA5-Land grid point and the glacier elevation or how do you deal with poorly resolved precipitation? (Could this be an explanation for why you could not find the expected importance of precipitation* *during the accumulation months?)*

Thank you for your suggestion. We acknowledge that this resolution of 9km/pixel is a poor representation to model glacier mass balance measurements at point scale. Approaches such as using a scaling factor or lapse rates have been attempted by studies (e.g. Radić et al 2014, Maussion et al 2019). However, these studies largely utilize precipitation and temperature as inputs, the scaling of which with elevation is fairly straightforward. Choosing appropriate scaling factors for other meteorological variables (e.g sensible and latent heat fluxes, albedo) is not intuitive. While we accept that the effects of the larger scale of the input variable will persist in the model, we would like to bring to notice that the effects will be consistent across all the models. Thus the effect of the input variable scale is represented by the uncertainty of all models. This will be described further in the subsection 4.1 Comparison of Model Performance and Associated Errors under Discussions.

COMMENT GC3: *Results sections 3.2, 3.3, 3.4, 3.5, 3.6: The last sentences of the first paragraph are not needed and could be incorporated at the end of the sentences where relevant things are discussed, e.g. '(Fig. 3).' at the end of the sentence, like is done in L307. This makes it easier for the reader to check your described findings by themselves in the plots. In the second paragraph, you can point to that this information is available in the supplementary in more detail.*

We thank the reviewer for pointing this out. We accept incorporating this change will improve the readability of the manuscript. The necessary changes will be included in the revised manuscript

COMMENT GC4: *To make it easier for the reader to interpret the Figures you could include subfigure tags (e.g. (a), (b), (c), ...) and describe in the Caption more precisely what is shown in each subfigure. Also, increase the font size where needed.*

Thank you for your suggestion. We will modify all the figures accordingly.

COMMENT GC5: *You should use the same units in the text and figures, e.g. in the text*

L241 it says RMSE value of 1.071 mwe, but in Figure 3 the y-axis shows 1071 (with no unit given).

Thank you for pointing out this oversight. For the next revision, we will ensure that all units are specified and consistent throughout the manuscript.

2 Specific Comments:

1 **Reviewer Comment:** *L190: define which months are accumulation months and which months are ablation months, should be done earlier in the manuscript (is only defined in L335)*

Author Response: Thank you. We agree with the suggestion. We will modify lines 189-191 as follows:

(b) Percentage importance associated with the accumulation months (November to March) and the ablation months (June-September) are summed and graphically represented for each model in Fig. 8.

2 **Reviewer Comment:** *L204: How is stabilizing the training metrics defined? We can not see this from Figure 2, maybe include a similar subplot as the right one for training performance.*

Author Response: Thank you for bringing this to our notice. We will include the Figure **R1** as a subplot to Figure 2 of the manuscript for clarity.

3 **Reviewer Comment:** *L207: Also here, how is stabilizing defined? 'This suggests that all models have successfully fit the data.': Doesn't it only shows that the results do not get better if we give the models more data than 50%, and it tells us nothing about how successful the fit is?*

Author Response: Thank you for bringing this to our notice. The term stabilizing is defined as "no significant change in the metric." Here, we have considered the change

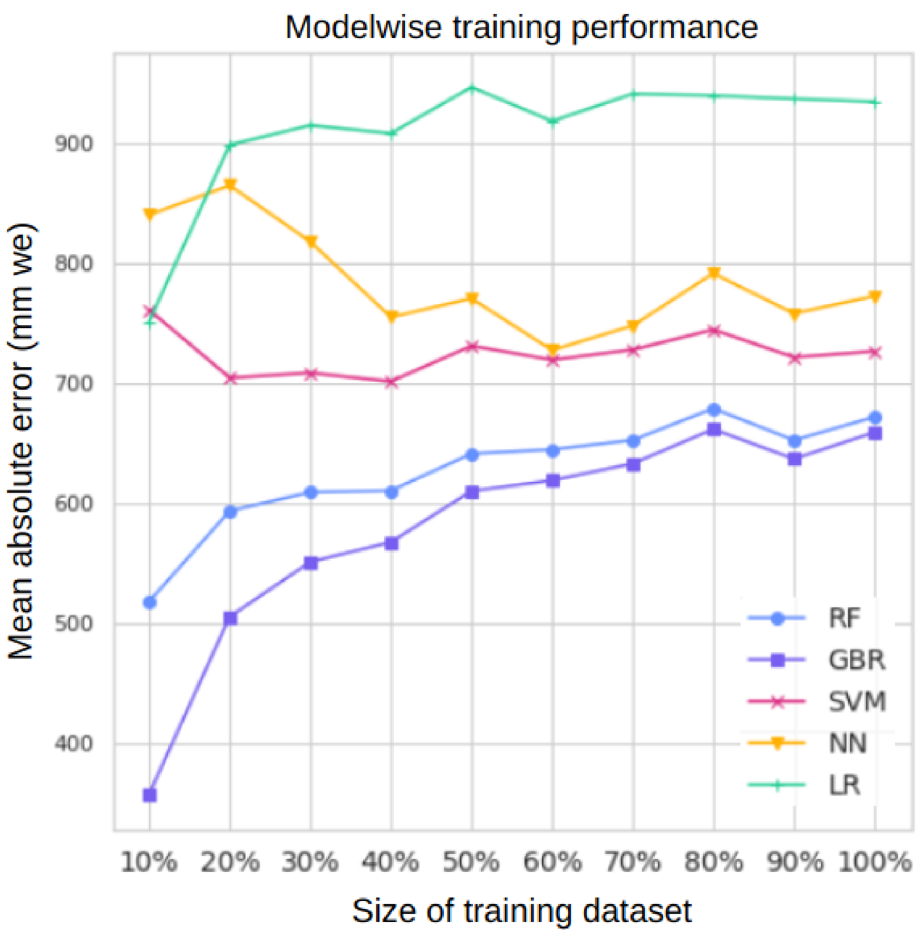
Modelwise training performance
Mean absolute error (mm we)
Size of training dataset
900
800
700
600
500
400
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
RF
GBR
SVM
NN
LR

Figure **R1**

in performance greater than 50mm we to be a significant change in mean absolute error. Further, regarding the sentence 'This suggests that all models have successfully fit the data,' we agree with the reviewer. For clarity, we rewrite the lines 204 to 208 as follows ***The training metrics do not show significant change after 20-30% of the training dataset size for the LR, RF, GBR and SVM models and after 40% for the NN model. This illustrates the larger number of trainable parameters resulting in the requirement of larger datasets for artificial neural networks for training. The testing performance of each of the models do not show significant change for training dataset sizes larger than 50%. We observe that while a downward trend is evident with the addition of new data, the rate of improvement is slower.***

4 **Reviewer Comment:** *L209: also here a plot suggested under L204 would be helpful to see the explained increase in training MAE*

Author Response: Thank you for this suggestion. We will incorporate Figure **R1** as a subplot to Figure 2 of the manuscript.

5 **Reviewer Comment:** *L217: How do you see this? (Smaller box in Figure 2 left?)*

Author Response: We can see from the plot on Figure 2 (right) that random forest, gradient-boosted regression and support vector machines depict the best performance for smaller datasets. We show in line 209 the tendency of random forest and gradient-boosted regression to overfit in the case of smaller datasets. Thus we conclude that support vector machines are better-suited algorithms in case of fewer datasets. For clarity, we modify line 217 as follows: ***Figure 2b depicts the superior performance of RF, GBR and SVM in the event of limited dataset availability. However, we have seen that RF and GBR show a marked increase in training MAE with increasing training samples which suggests overfitting to limited datasets. Thus SVM is more robust to smaller datasets.***

6 Reviewer Comment: *L240: Instead of 'This is depicted in Fig 5.' just right '(Fig. 5).' at the end of the sentence*

Author Response: Thank you. We accept this suggestion.

7 Reviewer Comment: *L247: define somewhere in the manuscript what are 'ablation meteorological variables'*

Author Response: Thank you for this suggestion. We will incorporate an explanation of which variables contribute to ablation and accumulation in the data and methods section 2.2

8 Reviewer Comment: *L261: is 'cost' the same as 'penalty'? If so you should be consistent and use one or the other throughout the manuscript.*

Author Response: Thank you for bringing this to our notice. This is a remnant of an earlier iteration of manuscript preparation. We will correct all occurrences of this oversight.

9 Reviewer Comment: *L304: How do you conclude this ranking? From Figure 3 and Figure 4, it looks like RF and SVM are closer than SVM and NN.*

Author Response: Thank you for bringing this to our notice. We were interpreting using Figure 2. Here, a consistent shift in the performance of mean absolute error of random forest and SVM is evident. To improve clarity, we modify lines 303-305 to ***The GBR model resulted in the best testing performance MAE, RMSE and R^2 values outperforming the RF model, SVM and NN models. Neural networks resulted in better bias performance.***

10 Reviewer Comment: *L326: To which graphs are you linking here? Maybe include the figure number.*

Author Response: Thank you for bringing this to our notice. We meant the graph of LR model in Figure 2. We will specify the figure number to avoid confusion.

11 Reviewer Comment: *L348: Probably you could not find the expected importance of precipitation because it is poorly resolved in the climate input data (see GC2).*

Author Response: Thank you. This is likely. We will include the line ***This is possibly a result of the scale of the meteorological variables used.***

12 Reviewer Comment: *Table 1: Why is 'Number of trees' listed two times?*

Author Response: We apologize for the lack of clarity in the representation of the table. We have corrected the design of the table to include horizontal separators as depicted in Table 1:

Table 1: Grid of settings used for hyperparameter tuning of each of the models

Machine learning model	**Hyperparameter**	**Values**
Random Forest	Number of trees	10,20,50,100
Gradient Boosted Regressor	Number of trees	50,100,200
	Subsampling	0.7, 1.0
	Maximum Depth	3,5,10
Support Vector Machine	Cost	0.1, 1, 10, 20
	Kernels	Sigmoid, Radial Basis Function, Polynomial
	Degree (polynomial kernel)	2, 3, 4, 5
Artificial Neural Network	Number of layers and nodes	**1:** 10, 50, 100, 200, 300, 400, 500, **2:** (100, 50), (200, 100), (400, 200), (200, 400) **3:** (400, 200, 100), (500, 200, 100), (200, 100, 50), (100, 50, 10), **4:** (200, 300, 400, 500), (300, 200, 100, 50), (200, 100, 50, 10)

13 Reviewer Comment: *Figure 2: See GC4. In the caption also explain which quantiles are shown in the box plot on the left. And explain how the two plots are connected (are the yellow boxes on the left representing the quantiles of the lines in the right plot?) Add the unit to the y-axis. Currently wrong caption: 'Training and testing RMSE (in* *mm we) and r values for varying the size of the training dataset for each of the models:' but only shown is MAE.*

Author Response: Thank you for bringing this to our notice. We agree with the concerns raised. We will update the caption in line with this suggestion.

14 Reviewer Comment: *Figure 3: See GC4. In caption: e.g. how are training and testing data split in this plot,70%/30% or different, include (a), (b), (c) and (d) and explain also in the caption which performance measure is shown in which subplot. Add units to the y-axis where needed.*

Author Response: Thank you for bringing this to our notice. We have corrected this figure to reflect the suggestions by the reviewer.

15 Reviewer Comment: *Figure 4: See GC4. Maybe you can include the information of Figure 3 into this figure and delete Figure 3 (e.g. "RMSE: 0.95/1.08 mwe" and include a legend at the empty subplot space lower right with "RMSE: Training/Testing"). For the y-equations don't write y=0.744x + (-338.433) instead write y = 0.744x – 338.433. Is the high precision of numbers with three decimals meaningful for the RMA regression?*

Author Response: Thank you for this suggestion. In fact, the initial draft of this manuscript included the information of Figure 3 in Figure 4 exactly as suggested by the reviewer and another iteration represented in the form of stacked line plots of all the metrics in the empty panel of Figure 4. However, both options appeared cluttered. To improve the readability a separate plot with the training and testing metric was included.

16 Reviewer Comment: *Figure 5: hard to distinguish in the legend what are the solid lines and what are the dashdotted lines. In the caption mention which test score is shown and explain briefly what the negative scaled RMSE is.*

Author Response: Thank you for bringing this to our notice. We will update the legend and caption to reflect this suggestion.

17 Reviewer Comment: *Figure 6: In the caption mention which test score is shown and explain briefly what the negative scaled RMSE is.*

Author Response: Thank you. We will incorporate this suggestion. We have used the negative of the root mean squared error after scaling the target labels to a range between 0 and 1 as the test score. This makes the assigning of ranks to the hyperparameter combination setting more intuitive.

18 Reviewer Comment: *Figure 7: increase the font size, In Caption mention which test score is shown. Also include the test score name in the y-axis (currently only 'Test score').*

Author Response: Thank you for bringing this to our notice. We have corrected the font size and the test score details.

19 Reviewer Comment: *Maybe you could combine Figures 5, 6 and 7 into one Figure.*

Author Response: Thank you. Yes, we agree. Figures 5, 6 and 7 will be merged.

20 Reviewer Comment: *Figure 8: increase the font size. Because the x-axis is limited to 13 maybe add the numbers in the plot for features which go beyond this limit. Maybe include the abbreviations of meteorological variables in the caption or the text, so you can understand the plot without having a look in the supplementary. And you can also use the abbreviations in the result sections.*

Author Response: Thank you for your suggestion. To improve the representation of the feature importance and the meteorological variables without having to look at the supplementary file, we reformatted the image in the form of a RADAR plot. The tentative figure is as depicted in Figure **R2**.

21 Reviewer Comment: *Supplementary S1:*

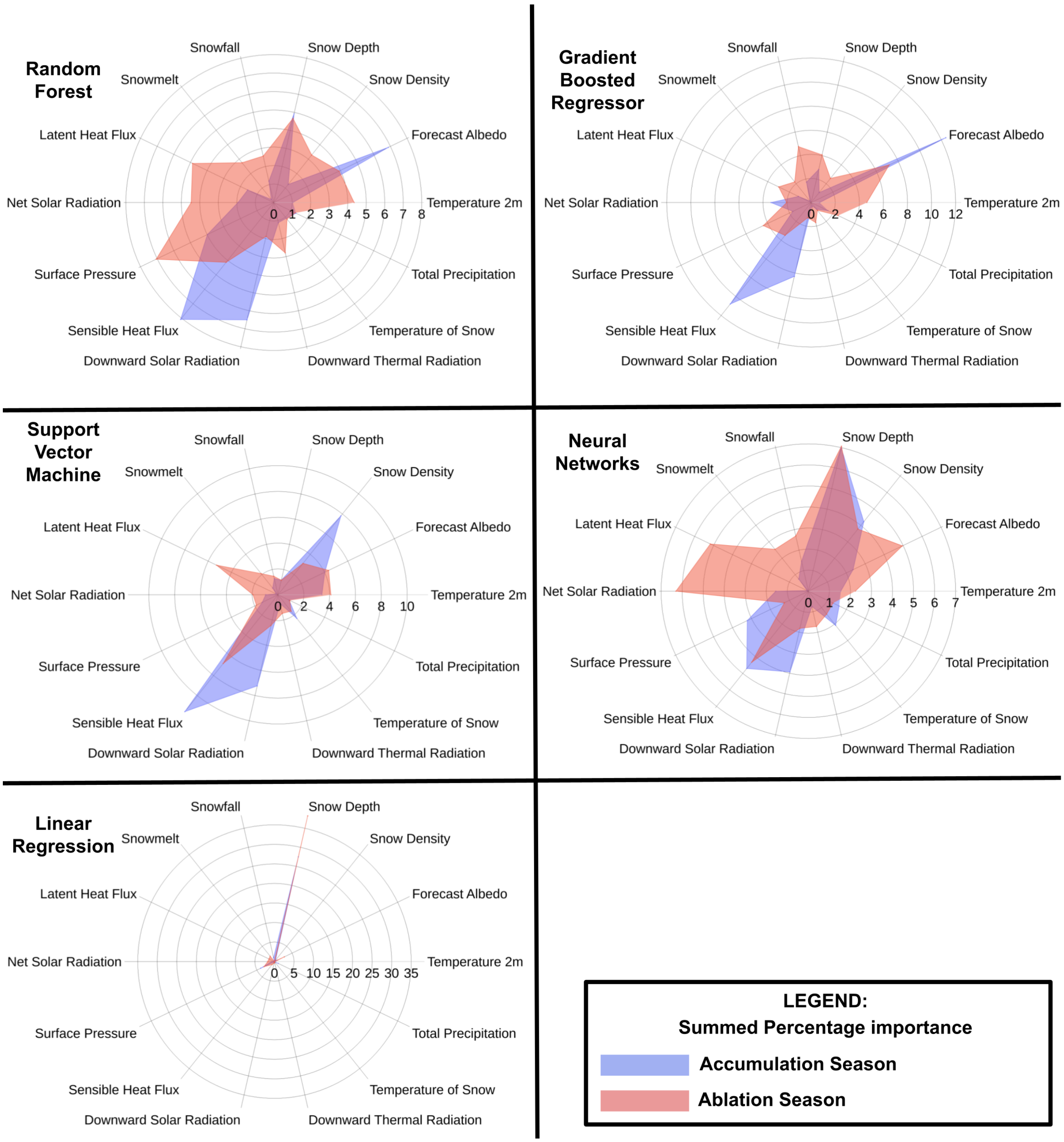

Figure **R2**: Radar plot depicting the percentage importance of all features summed over the accumulation and ablation season for the models: Random Forest, Gradient Boosted Regression, Support Vector Machine, Artificial Neural Network and Linear Regression. The radial axis represents the summed percentage importance and the angular axis represents the input features.

- *general: give more meaningful names to the individual sheets*
- *sheet3: no explanation of what is shown on this sheet, include references in the text or delete this sheet*

Author Response: Thank you for bringing this to our notice. Sheet 3 will be deleted. Sheet 1 will also be deleted as the new Radar plot (Fig **R2**) contains the full name of the meteorological variables.

References

1. Radić, V., Bliss, A., Beedlow, A.C., Hock, R., Miles, E. and Cogley, J.G., 2014. Regional and global projections of twenty-first century glacier mass changes in response to climate scenarios from global climate models. Climate Dynamics, 42, pp.37-58.

2. Maussion, F., Butenko, A., Champollion, N., Dusch, M., Eis, J., Fourteau, K., Gregor, P., Jarosch, A.H., Landmann, J., Oesterle, F. and Recinos, B., 2019. The open global glacier model (OGGM) v1. 1. Geoscientific Model Development, 12(3), pp.909-931.