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 detailed and insightful comments 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 color and italicised font style*. The author's response is in blue font color and normal font style. Text quoted from the revised manuscript is ***blue font color and bold italicized font style***.

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. 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 COMMENT 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.

Thank you for pointing out the lack of clarity in the manuscript pertaining to the validation and testing dataset. We have used a 70%-30% split for training and testing. Here, 30% is reserved for assessing the model performance only. It has not been used for hyperparameter selection. For hyperparameter selection, we use a 3 fold cross validation using the 70% training split. The modifications will be included in the revised manuscript. Line 146 in the original manuscript 'We split the dataset into training and testing samples to be utilised by the model' is revised to: ***We have split the dataset using a random split where 70% of the total dataset is used for training the model and 30% is used for testing the model performance. The training split is used in a 3 fold cross validation process for tuning the hyperparameters as described further in Section 2.3***

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 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-GroupOut, 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).

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.

Thank you for your suggestion. We have maintained a 70-30 ratio for the training and testing datasets. The reason we went for a 3 fold cross validation is to maintain the data folds in a manner that best replicated the ratio used for the overall training and testing. Among the 3 folds, for a given iteration, 2 folds are used for training and 1 for validation. This 2:1 ratio of training and validation is the closest representation of the 70:30 split of the training and test datasets.

Regarding the cross validation and testing data split methodology, we accept the point made by the reviewer that spatially and temporally structured datasets would benefit from a manually designed blocking strategy such as the Leave One Glacier Out and Leave One Year Out strategy as depicted in Bolibar et al 2020. Acknowledging the merits of this technique in validation and testing split generation, we would like to bring to the attention of the reviewer that through this study, we aim to perform a comparative assessment of a number of machine learning models available using present day data driven techniques and explain the feature importance associated with a machine learning modelling of the glacier mass balance. As the testing and validation splits will result in similar effects in all the models, performing an additional exercise in blocking strategies of the data split might dilute the information we wish to convey. However, for cases where a single model is to be used to estimate glacier mass balance, we accept that the Leave One Glacier Out and Leave One Year Out techniques are vital. We propose including this in subsection 4.4 Relevance to future studies under Discussions (see response to GC4) and incorporating a section in the supplementary material with a comparison of the effect of the blocking strategy and the random split performance.

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)?

Thank you for this point. We will perform this exercise and include this as an additional exercise in the supplementary material demonstrating the role of the min-max 0 to 1 scaling compared to the StandardScaler that performs a mean subtract and scaling by standard deviation.

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.*

We are grateful to the reviewer for bringing this to our attention. We can generate the figure of the glacier mass balance measurements as a histogram to provide a complete representation of the mass balance and the errors to the readers. Further, we can include an additional error metric normalized root mean square error (nRMSE) which depends upon the root mean square error (RMSE) and the standard deviation of mass balance measurements (σ_{MB}) defined as:

$$nRMSE = \frac{RMSE}{\sigma_{MB}} \tag{1}$$

to represent the errors including the context of the natural variation in the target data. This will be included in the revised manuscript.

1.2 COMMENT 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.*

We thank the reviewer for pointing out the ambiguity in the text explaining how the training and testing split was undertaken. In fact, we have maintained the ratio of a 70:30 split consistently for varying dataset sizes to ensure a complete separation between the training and testing samples. In order to explain this, we modify line 180-182 in the original submission as follows:

"To understand the effect of data availability on the model performance, we perform an experiment on varying the training sizes. We split the original dataset into subsets of iteratively increasing sizes. We partition each subset into training and testing partitions using a 70:30 ratio. For each subset, we train all the models using the training partition and computed the evaluation metrics over the testing partition. "

1.3 COMMENT 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.

We thank the reviewer for pointing out the the lack of clarity in the text explaining how the climate datasets have been used. We agree ERA5 is known to be erroneous in complex terrains due to its course spatial resolution (31 km/pixel). We have used the ERA5-Land product which is generated by integrating the ECMWF land surface model driven by the downscaled meteorological forcing from the ERA5 climate reanalysis dataset. This included an altitude correction and is described further in Muñoz-Sabater et al 2021. The final product we have used has a spatial resolution of 9 km/pixel.

We acknowledge that this resolution of 9km/pixel is also large as we are using point glacier mass balance measurements. Applying a scaling factor can be straightforward in case of features such as temperature. However, 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 note 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.

1.4 COMMENT 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.*

Thank you for bringing this point to our notice. As suggested, we will incorporate a subsection under Discussions titled '4.4 Relevance to future studies' in the revised manuscript which will include the following points:

- The continued importance of tree based methods for tabular data structures keeping in mind the suggestions provided in Grinsztajn et al 2022
- Guidelines on the use of machine learning techniques for future studies
- Extension of the study to other datasets available
- Extension of the study to other data sparse RGI regions with emphasis on the role of transfer learning.
- Reducing uncertainties of the models by downscaling of input variables.
- Understanding the role of feature importance for different glaciated regions and the importance of local, regional and global inputs.

2 Specific Comments:

1 **Reviewer Comment:** *Title I believe the title would sound better as "Modelling Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques".*

Author Response: Thank you. We agree the title Modelling Point Mass Balance for the Glaciers of Central European Alps using Machine Learning Techniques sounds better. We will incorporate this change in the updated manuscript.

2 **Reviewer Comment:** *L14-15 I'd rather present the RMSE (or MSE) in the abstract than the r^2, since it provides more information*

Author Response: Thank you. We agree with the comment and will include the RMSE in the abstract as well.

3 **Reviewer Comment:** *L35-36 I would also add the great number of parameters to calibrate.*

Author Response: Thank you. We will modify the lines in the revised manuscript as follows: ***However, the substantial requirement for ground data to force the model, the sizeable number of parameters to calibrate and the computational complexity associated with running the model make it cumbersome to use for large areas***

4 **Reviewer Comment:** *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]

Author Response: Thank you for directing us to this study. We agree, the variations in DDFs spatially and temporally are significant. We will include the following lines in the text: ***However, using only temperature and precipitation as inputs can lead to oversimplification. Further, the degree day factors (DDF) considered in temperature index models are often invariant. But studies such as Gabbi et al 2014, Matthews and Hodgins 2016, Ismail et al 2022 have observed a decreasing trend in DDF, particularly in higher elevations. Ismail et al 2022 also report the sensitivity of the DDF under the influence of the changing climate, particularly to to solar radiation and albedo.***

5 **Reviewer Comment:** *L57 For me the sentence would read better as "and a nonlinear neural network..."*

Author Response: We agree. The sentence is fixed to read as: ***Bolibar et al. (2020) used a least absolute shrinkage and selection operator (LASSO) regression, a linear model, and a nonlinear neural network model to simulate glacier mass balance.***

6 **Reviewer Comment:** *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"*

Author Response: Thank you for bringing this to our notice. We have modified lines 57-62 in the updated manuscript as follows: ***Steiner et al. (2005); Vincent et al. (2018); Bolibar et al. (2020, 2022) are some of the few studies reporting consistently better performance of non-linear models over linear models. These studies have largely used neural networks. However, a gamut of ML techniques such as ensemble-based and kernel-based techniques exist which have largely been under-utilized for the purpose of modelling glacier mass balance.***

7 **Reviewer Comment:** *L68 Why use a linear regression example after mentioning NNs?*

Author Response: Thank you for pointing out this oversight. We meant to represent the inputs used in data driven models, not neural networks specifically here. We have corrected line 67 in the original manuscript to: ***Existing data-driven models typically use a subset of topographic and meteorological variables.***

8 **Reviewer Comment:** *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.*

Author Response: Thank you. We agree. All instances of training labels will be replaced by target data.

9 **Reviewer Comment:** *L100 Same with "labels"*

Author Response: Thank you. This is fixed.

10 **Reviewer Comment:** *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.*

Author Response: Thank you for pointing out the lack of consistency in this usage. We have fixed all occurrences to reflect the correct term.

11 **Reviewer Comment:** *L104 This would read better as "is a decision (regression or classification) tree".*

Author Response: Thank you. This modification has been incorporated.

12 **Reviewer Comment:** *L106 This would read better as "To illustrate this".*

Author Response: Thank you. This modification has been incorporated.

13 **Reviewer Comment:** *L122 The subject of the sentence is missing (i.e. "a neural network").*

Author Response: Thank you. We have corrected this sentence as follows: ***Hornik (1991) showed that neural networks with as few as a single hidden layer with a sufficiently large number of neurons, when used with a non-constant unbounded activation function, can function as universal function approximators.***

14 **Reviewer Comment:** *L125-126 "Nonlinearity" should be "nonlinearities".*

Author Response: Thank you. This modification has been incorporated.

15 **Reviewer Comment:** *L129 Same with "labelled data" and "labels".*

Author Response: Thank you. This modification has been incorporated.

16 Reviewer Comment: *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.*

Author Response: Thank you for this suggestion. We considered annual mass balance observations purely due to availability of data. The database of point glacier mass balance observations contain separate entries for annual mass balance observations, summer and winter 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. 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 this database.

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

Author Response: Thank you. This point is addressed in the response to GC3.

18 Reviewer Comment: *L146 As per GC1, please explain this better and make the corresponding changes.*

Author Response: Thank you. We have incorporated the suggestion as described in the response of GC1.

19 Reviewer Comment: *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.*

Author Response: Thank you for bringing this to our notice. We will correct all instances of the misuses of terms hyperparameters and parameters

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

Author Response: Yes, we have modified the lines 151-156 (Rather than using...optimal hyperparameters are selected) for improved clarity as follows: ***We have considered a hyperparameter grid with all combinations of values that each hyperparameter can take (see Table 1). Rather than using a fixed ratio subset for validation as was the case with the testing, we divided the training data subset into three equal folds. Two folds are randomly selected as the training set and the third fold is used for validation. The validation score is noted and the process is then repeated for the other fold combinations. The mean validation score for each hyperparameter setting obtained from the grid is used for selection of the optimal hyperparameters.***

21 Reviewer Comment: *L154 Following GC2, please better explain this and make the necessary changes.*

Author Response: Thank you. The changes have been incorporated as specified in the previous comment response.

22 Reviewer Comment: *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.*

Author Response: Thank you for pointing this out. Yes, we did intent to write validation score. We have corrected it at all occurances of this error.

23 Reviewer Comment: *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.*

Author Response: In the k fold validation technique, we have considered all permutations of folds as training and testing subsets. For example, for our case where 3 folds were used, First fold 1 and 2 were used for training and 3 for assessment (validation). Then 2 and 3 for training and 1 for assessment. Finally 1 and 3 were used for training and 2 for assessment. There are thus 3 assessment values that we obtain. The mean score of this is the final validation score which is used to represent each hyperparameter setting. While the rescaled RMSE provides an estimate of errors that is useful at the time of reporting accuracies, for a comparative analysis, the relative values are sufficient. Hence the scaling back to original units was not undertaken. Further, we used negative of the RMSE purely for intuition in assigning ranks to the hyperparameter combination setting. Settings with higher RMSEs perform poorly. Thus settings with higher negative RMSE perform well and can be ranked better.

24 **Reviewer Comment:** *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?*

Author Response: Thank you for bringing this up. We did try alternate runs with Parametrized Leaky ReLU on PyTorch. We did not include it in the final version of the manuscript for brevity. We can include the sample runs as well as the code for the same in the Supplementary material.

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

Author Response: Thank you. This is fixed.

26 **Reviewer Comment:** *L209 Please see GC2*

Author Response: Thank you. We have incorporated the changes described in response to GC2 at line 180.

27 **Reviewer Comment:** *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?

Author Response: These errors are the same as those depicted in Figures 3 and 4 of the original submission. It is different from Figure 2 as the mean absolute error is reported in Figure 2 not the Root Mean Squared Error. We will include the mean absolute errors in addition to the RMSE in the revised manuscript text for all models.

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

Author Response: Thank you for bringing this to our notice. For clarity, we will rewrite the lines 320-322 as follows: ***The testing performance improves on increasing the number of training samples. We observe that for larger number of data points, marginal improvement is observed upon increasing the number of samples further. The reduction in rate of improvement for all models suggest that all models have been successfully trained. However, the marginal improvements observed suggest a potential improvement in model performance is is possible when including more data samples.***

29 **Reviewer Comment:** *L324 I think you mean hyperparameters here.*

Author Response: Thank you for your comment. We mean parameters here as we are referring to the weights and not the hyperparameters such as number of layers, number of neurons or activation.

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

Author Response: Thank you. Yes, here, we mean the weights associated with the network and hence parameters was used. Other erroneous misuse of the terms parameter and hyperparameters have been corrected in the revised version of the manuscript.

31 **Reviewer Comment:** *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?*

Author Response: Tree based models do provide a feature importance analysis. However, these use a mean decrease in impurity (RMSE, MSE for regression or gini for classification). Strobl et al 2007 report a skewed representation of features in such cases as a result of varying scales of the data and correlation between the input features. In our study, normalization is performed. Thus, the varying scales of data will not be a hindrance. However, correlation is observed between the input variables. This renders tree based importance metrics less accurate. This issue is resolved using permutation importance. Thus we selected permutation importance . An additional advantage of using permutation importance is to be able to use a model-agnostic explainability metric.

32 **Reviewer Comment:** *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.*

Author Response: Thank you for bringing this to our notice. We correct this to the following: ***Albedo over snow-covered regions is higher than that of exposed ice or firn. At higher elevations and in summer months, we expect the lower values of albedo. Thus variations in albedo are are significance. The expected importance of the albedo is observed in the RF, GBR, NN and SVM model. LR models, in contrast, depict a very low importance of albedo for the accumulation months.***

33 **Reviewer Comment:** *L362-364 This is a very interesting finding and in line with*

recent studies from the machine learning community regarding ML for tabular data (cite1)

Author Response: Thank you for your comment. In line with GC4, we are including a subsection in the Discussions where we will describe the importance of tree based ensembled in working with tabular datasets.

34 Reviewer Comment: *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.*

Author Response: Thank you. Yes, we will delete this sentence as the next sentence explains this better.

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

Author Response: Thank you. We have fixed this. The table now appears as depicted in Table 1:

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

Author Response: Thank you. We have fixed this. The figure is as depicted in Figure **R1**:

37 Reviewer Comment: *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.*

Author Response: Thank you. Yes, here, we used the mean absolute error and the errors reported in the text are the root mean squared error. For clarity, we will include mean absolute error in the text for each model.

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)

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

Author Response: Thank you, the labelling of figures will be corrected. The errors specified here are different from figure 2 because here, we depict the root mean squared error as opposed to mean absolute error in Figure 2.

39 Reviewer Comment: *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.*

Author Response: Thank you. We agree. We will merge them into a single figure for the final manuscript.

40 Reviewer Comment: *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.*

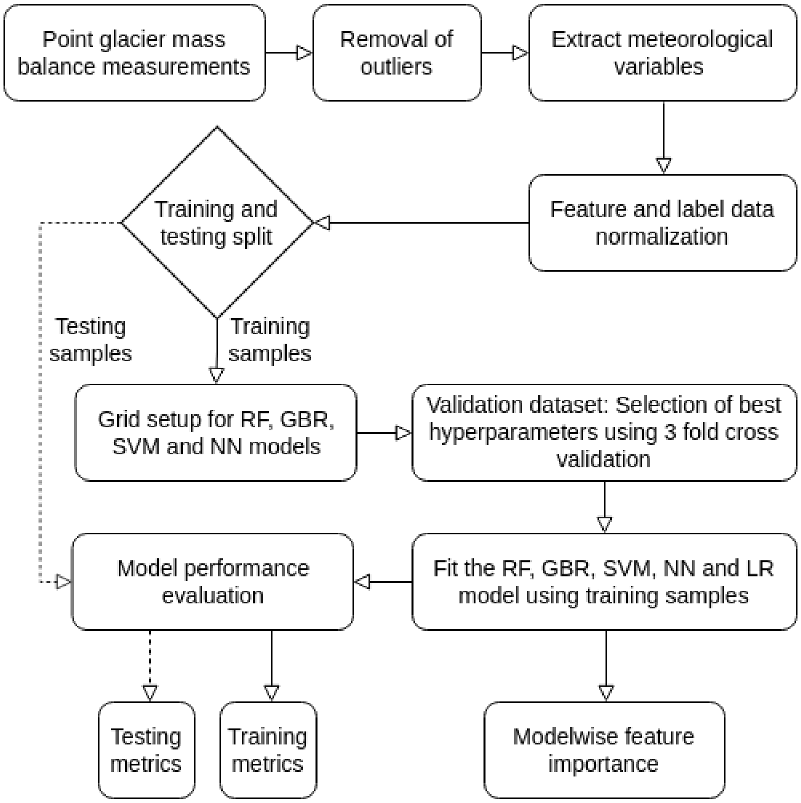
Point glacier mass balance measurements
Removal of outliers
Extract meteorological variables
Feature and label data normalization
Training and testing split
Testing samples
Training samples
Grid setup for RF, GBR, SVM and NN models
Validation dataset: Selection of best hyperparameters using 3 fold cross validation
Fit the RF, GBR, SVM, NN and LR model using training samples
Model performance evaluation
Testing metrics
Training metrics
Modelwise feature importance

Figure **R1**: Flowchart of the methodology

Author Response: Thank you. We agree. To improve readability, we reformatted the image in the form of a RADAR plot with the labels on the right. The tentative figure is as depicted in Figure **R2**.

References

1. Muñoz-Sabater, J, Dutra, E, Agustí-Panareda, A, Albergel, C, Arduini, G, Balsamo, G, Boussetta, S, Choulga, M, Harrigan, S, Hersbach, H and Martens, B (2021) ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), pp.4349-4383. doi: 10.5194/essd-13-4349-2021

2. Gabbi, J, Carenzo, M, Pellicciotti, F, Bauder, A and Funk, M (2014) A comparison of empirical and physically based glacier surface melt models for long-term simulations of

glacier response. J. Glaciol., 60(224), 1140–1154. doi: 10.3189/2014JoG14J011

3. Matthews, T., and Hodgkins, R. (2016). Interdecadal variability of degree-day factors on Vestari Hagafellsjökull (Langjökull, Iceland) and the importance of threshold air temperatures. Journal of Glaciology, 62(232), 310-322. doi:10.1017/jog.2016.21

4. Ismail, M. F. and Bogacki, W. and Disse, M. and Schäfer, M. and Kirschbauer, L. (2022). Estimating degree-day factors based on energy flux components. The Cryosphere Discussions 1-40, doi:10.5194/tc-2022-64

5. Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data?. arXiv preprint arXiv:2207.08815.

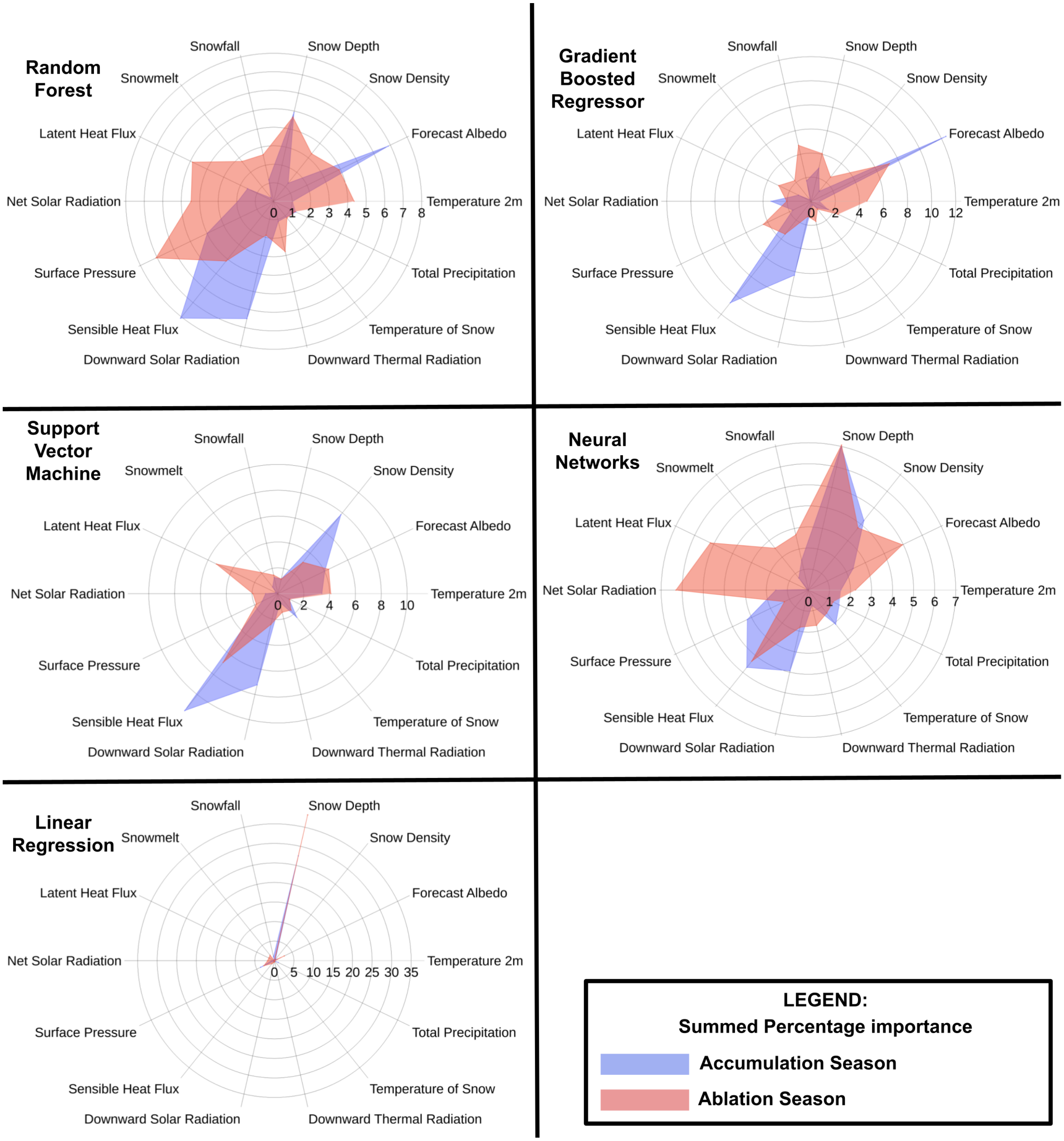

Figure **R2**: Radar plot depicting the percentage importance of all features summed over the accumulation and ablation season for the models: Random Forest, Gradient Boosted Regression, Support Vector Machine, Artificial Neural Network and Linear Regression. The radial axis represents the summed percentage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