https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

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

Ritu Anilkumar[1,2], Rishikesh Bharti[2], Dibyajyoti Chutia[1], and Shiv Prasad Aggarwal[1]

[1]North Eastern Space Applications Centre, Department of Space, Umiam, Ri Bhoi
[2]Department of Civil Engineering, Indian Institute of Technology, Guwahati

Correspondence: Ritu Anilkumar (ritu.anilkumar@nesac.gov.in)

Abstract. Glacier mass balance is typically estimated using a range of in-situ measurements, remote sensing measurements, and physical and temperature index modelling techniques. With improved data collection and access to large datasets, data-driven techniques have recently gained prominence in modelling natural processes. The most common data-driven techniques used today are linear regression models and, to some extent, non-linear machine learning models such as artificial neural networks. However, the entire host of capabilities of machine learning modelling has not been applied to glacier mass balance modelling. This study used monthly meteorological data from ERA5-Land to drive four machine learning models: random forest (ensemble tree type), gradient-boosted regressor (ensemble tree type), support vector machine (kernel type) and artificial neural networks (neural type). We also use ordinary least squares linear regression as a baseline model against which to compare the performance of the machine learning models. Further, we assess the requirement of data for each of the models and the requirement for hyperparameter tuning. Finally, the importance of each meteorological variable in the mass balance estimation for each of the models is estimated using permutation importance. All machine learning models outperform the linear regression model. The neural network model depicted a low bias, suggesting the possibility of enhanced results in the event of biased input data. However, the ensemble tree-based models, random forest and gradient-boosted regressor outperformed all other models in terms of the evaluation metrics and interpretability of the meteorological variables. The gradient-boosted regression model depicted the best coefficient of determination value of 0.713. The feature importance values associated with all machine learning models suggested high importance to meteorological variables associated with ablation. This is in line with predominantly negative mass balance observations. We conclude that machine learning techniques are promising in estimating glacier mass balance and can incorporate information from more significant meteorological variables as opposed to a simplified set of variables used in temperature index models.

1 Introduction

We can visualize glaciers as interactive climate-response systems with their response described by changes in glacial mass over a given period (e.g. White et al., 1998). Several studies have reported the impact of climate change on glacier mass at a global and regional scale (e.g. Le Meur et al., 2007; Huss et al., 2008), with repercussions including and not limited to glacial outburst floods and diminishing water supplies. Thus, understanding the response of glacier mass balance to climate change is

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

crucial. Glacier mass balance is most commonly measured via (i) Direct Glaciological Method, where point measures of gain or loss of glacial ice are obtained and extrapolated for the entire glacier (e.g. Kuhn et al., 1999; Thibert et al., 2008; Pratap et al., 2016), (ii) Geodetic Method, where the change in surface elevation between two-time instances for the same portion of the glacier is estimated (e.g. Rabatel et al., 2016; Tshering and Fujita, 2016; Trantow and Herzfeld, 2016; Bash et al., 2018; Wu et al., 2018) and (iii) Indirect Remote Sensing Method, where measured mass balance is correlated with the Equilibrium Line Altitude (ELA) values or Accumulation Area Ratio (AAR) values for time series data (e.g. Braithwaite, 1984; Dobhal et al., 2021). In addition to observational data, simple temperature index-based or sophisticated physics-based energy balance models (e.g. Gabbi et al., 2014) have also been developed. Energy balance models compute all energy fluxes at the glacier surface and require measurements of input variables such as meteorological and other inputs at the glacier scale (e.g. Gerbaux et al., 2005; Sauter et al., 2020). As these models are driven by the physical laws governing energy balance, they provide reliable estimates of glacier mass balance. However, the substantial requirement for ground data to force the model and the computational complexity associated with running the model make it cumbersome to use for large areas. Temperature index models use empirical formulations between temperature and melt (e.g. Radić and Hock, 2011). The simplicity afforded by these models permits extension to large scales effectively. However, using only temperature and precipitation as inputs can lead to oversimplification.

With increasing data points available, a new set of data-driven techniques has gained prominence in various domains of Earth Sciences. For e.g weather prediction (for a review, see Schultz et al., 2021), climate downscaling (e.g. Rasp et al., 2018), hydrology (e.g. Shean et al., 2020) have used data-driven models, particularly, machine learning (ML) and deep learning (DL) models. Cryospheric studies, too, have adopted the use of deep learning in several prediction problems (see review in Liu, 2021). Applications of deep learning in glaciology range from automatic glacier mapping (e.g. Lu et al., 2021; Xie et al., 2021), ice thickness measurements (e.g. Werder et al., 2020; Jouvet et al., 2021; Haq et al., 2021), calving front extraction (e.g. Zhang et al., 2019; Mohajerani et al., 2021), snow cover mapping (e.g. Nijhawan et al., 2019; Kan et al., 2018; Guo et al., 2020), snow depth extraction (e.g. Wang et al., 2020; Zhu et al., 2021), sea and river ice delineation (e.g. Chi and Kim, 2017; Li et al., 2017). The use of ML and DL in glacier mass balance estimation is significantly fewer. Initial data-driven studies used multivariate linear regression to estimate glacier mass balance from temperature and precipitation Hoinkes (1968). Subsequently, several papers have used linear regression methods for varying inputs such as temperature and pressure (Lliboutry, 1974), positive degree days, precipitation, temperature and longwave radiation (Lefauconnier and Hagen, 1990). Recent studies continue to use linear regression for modelling glacier mass balance. For example, Manciati et al. (2014) used linear regression to study the effect of local, regional and global parameters on glacier mass balance, Carturan et al. (2009) used linear regression to incorporate the effects of elevation models in the estimation of summer and winter mass balance measurements. Steiner et al. (2005) was the first to use neural networks to estimate glacier mass balance for the Echaurren glacier. Bolibar et al. (2020) used a least absolute shrinkage and selection operator (LASSO) regression, a linear model, and the non-linear neural network model to simulate glacier mass balance. Despite studies such as Steiner et al. (2005); Vincent et al. (2018); Bolibar et al. (2020, 2022) reporting consistently better performance of non-linear models over linear models, few studies have used ML for modelling glacier mass balance. Most studies using data-driven techniques attempt a linear modelling framework or shallow neural networks. However, there exists a variety of other ML models, such as classification and regression tree, random forests, radial basis function networks, support vector machines. The utility of these models has not been sufficiently explored in glacier mass balance modelling. This limited utilization of ML models is potentially due to the unavailability of large ground truth datasets required for training the ML models and the perceived black-box nature of ML techniques. We aim to address this by assessing the performance of different ML models for varying training dataset sizes. Further, we aim to shed light on the interpretability of ML models by using permutation importance to explain the relative importance of the input meteorological variables. The interpretability of machine learning models is largely dependent on the input variables provided. Existing non-linear neural network models typically use a subset of topographic and meteorological variables. For example, Hoinkes (1968) uses temperature, precipitation and cyclonic/anti-cyclonic activity, Steiner et al. (2005) uses precipitation and temperature, Masiokas et al. (2016) uses temperature, precipitation and streamflow. To the extent of the authors' knowledge, no ML-based study has attempted to use a complete set of meteorological variables associated with the energy balance equation. We expand upon this and assess the monthly contributions of each of these meteorological variables in the estimation of glacier mass balance.

Through this study, we assess the ability of ML models to estimate annual point mass balance. We use an example of each of the following classes of ML models: ensemble regression tree-based, kernel-based, neural network-based and linear models. Under ensemble regression tree-based, we chose one example of boosted and unboosted models. Specifically, we compare the performance of the random forest (RF), gradient-boosted regressor (GBR), support vector machine (SVM) and artificial neural network (ANN) models against a linear regression (LR) model. We also assess the performance for varying dataset sizes as real-world measurements are limited. Finally, to explain the role of the input features on each of the ML models, we use permutation importance described further in Altmann et al. (2010). The input features for the models are the monthly mean of 14 meteorological variables associated with the energy balance equation. We obtained the meteorological data from the ERA5-Land Reanalysis dataset (Muñoz Sabater, 2019, 2021). The labels used for training the ML models are obtained from the Fluctuations of Glaciers database (WGMS, 2021; Zemp et al., 2021) over the second-order region Alps defined by Randolph Glacier Inventory under first-order region 11: Central Europe (RGI, 2017). Section 2 of the manuscript further describes each of these datasets. In this section, we also elucidate the preprocessing steps associated with an ML approach and outline the methodology followed. In sections 3 and 4, we compare the performance of each of the models for various configurations of data availability. We also delve into the interpretability of the models from a feature importance perspective. The specific point we investigate as a part of this study can be summarized as follows:

1. Understand the utility of ML models in the estimation of glacier mass balance using limited real-world datasets
2. Identify specific use cases for different classes of ML models(ensemble tree-based, kernel based, neural network based and linear regression) pertaining to data availability, evaluation metrics and explainability
3. Investigate the ability of ML models to unravel the underlying physical processes
4. Explain the relative importance of meteorological variables contributing to the mass balance estimation on a monthly basis over the year

2 Data and Methods

2.1 Machine learning modelling

ML modelling is a data-driven set of modelling techniques. Here, we used a supervised learning framework for regression where inputs are in the form of monthly meteorological variables and targets are in the form of point measurements of glacier mass balance. The actual point mass balance measurements are the training labels vital to tuning the model parameters. We do this parameter tuning by designing a loss function defining the variation between the actual mass balance measurements, i.e. the labels, and the point mass balance estimates, i.e., the model's output. We start with random initialization of model parameters and finetune the parameters to minimize the loss function. For each of the ML models used in the study, we used the mean squared error (MSE) as the loss function. Further, we obtained the features of importance by assessing permutation importance. Figure 1 depicts the complete workflow used for the study.

The RF model is an ensemble-based algorithm where the base learner used is decision (regression or classification) trees (Breiman, 2001). It relies on the principle of bootstrap aggregating or bagging (proposed by Breiman, 1996) for the generation of multiple training datasets to be used by each base learner (Dietterich, 2000). To illustrate, assume there are N_{data} samples in the training dataset D, and a new dataset $\hat{D}$ is generated by sampling N_{data} samples with repetition. In addition to the generation of bootstrapped datasets, the decision trees are generated using a random subset of input features at every impure node of the tree instead of a complete set of features that standard regression trees use.

Like the RF model, the GBR model is an ensemble-based algorithm where aggregated base learners of decision (classification or regression) trees provide an estimate. However, it differs from the RF model because it uses boosting instead of bagging to construct ensembles. In boosting-based ensembles, base learners are typically weak learners, and the design of subsequent learners is such that the overall error reduces (Natekin and Knoll, 2013; Friedman, 2001).

The SVM model is a powerful ML tool that relies on Cover's theorem. The theorem suggests that data that might not be linearly separable in a lower dimensional space can be linearly separable when transformed into a higher dimensional space. In the context of classification, the SVM model uses a kernel to transform the data into a higher dimensional space (Cortes and Vapnik, 1995) where linear separability is feasible in the form of a hyperplane and decision boundaries. For this purpose, we use kernels such as polynomial kernel and radial basis function kernel (Vapnik, 1999). In the case of regression, the hyperplane represents the best fit line. Thus, unlike empirical risk minimization, where the difference between the actual and predicted model is optimized, the SVM model for regression uses structural risk minimization by identifying the best fit line.

McCulloch and Pitts (1943) proposed the NN models as mathematical representations of biological neuron interconnections. Hornik (1991) showed that 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. Presently, several applications (Seidou et al., 2006; Moya Quiroga et al., 2013; Haq et al., 2014) using multiple layered NN models demonstrate that NN can infer abstract relationships between features. NN models use weighted combinations of input features in tandem with non-linearity provided by activation functions such as sigmoid, tanh and rectified linear unit (ReLU), resulting in the model output. The weights of the NN model are the model parameters obtained by optimization of the loss function.

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

2.2 Preparation of features and labels

The most crucial component in ML modelling is the availability of labelled data to train the model. The labels used for training should be representative of the entire population. Hence, we chose the Fluctuations of Glaciers (FoG) database (WGMS, 2021; Zemp et al., 2021) that contains measured point mass balance information (46,356 data points) globally. The study area is the Randolph Glacier Inventory (RGI) version 6 (RGI, 2017) second-order region Alps under the first-order region 11: Central Europe. This consisted of 15,727 glacier mass balance point measurements. We performed a first-level preprocessing where we considered only annual mass balance measurements (10,102 data points) and measurements from 1950 (9,595 data points) onward. We then performed an outlier removal where we considered only those points within two standard deviations of the median. This was to avoid the effects of noisy data. We finally used 9166 data points to apply our model.

The second aspect is the input features used by the model to make predictions. We used the ERA5-Land reanalysis dataset (Muñoz Sabater, 2019, 2021) and specifically the variables contributing to the energy balance equation that drives mass balance modelling from a physical standpoint. We considered the monthly mean of each of the following fourteen variables for the modelling: temperature at 2m, snow density, snow temperature, surface net solar radiation, total precipitation, forecast albedo, surface pressure, surface net solar radiation downwards, snowfall, surface net thermal radiation, snowmelt, surface sensible heat flux, snow depth and surface latent heat flux (For details, see Muñoz Sabater et al., 2021). Thus, we have 168 total input parameters. For each of these variables, we extracted the data using the nearest neighbour algorithm, using latitude, longitude and year of the glacier mass balance measurement from the FoG database. Thus the final dataset has 168 input features and 9166 data points. We then normalised the data points using a min-max scaling to ensure the absence of user-conceived bias in the model. We split the dataset into training and testing samples to be utilised by the model. Finally, we rescaled the model's predictions to assess the model metrics, such as root mean squared error (RMSE), in the measured point mass balance units.

2.3 Hyperparameter Selection and Finetuning

In typical ML workflows, we split the complete dataset (set of features and labels) into training, validation and testing. We fit the model to the data using the training subset, tune the parameters using the validation subset, and report the independent performance metrics using the testing subset. In our case, we used a 70%-30% split for training and testing. Rather than using a subset for validation, we used a grid search approach to tune the parameters associated with each model. Table 1 depicts the grid used for estimating the parameters. We estimated the best set of hyperparameters using k-fold cross-validation. Here, we used three-fold cross-validation, i.e., the data was split into three subsets. We used two of the three subsets for training the model and one for testing. This is repeated by considering all combinations of the subsets for training and testing (in this case, three) for each hyperparameter combination. Based on the mean test score, the optimal hyperparameters are selected. We compute the test score as the negative of the root mean squared error after scaling the target labels to a range between 0 and 1. Thus a more negative test score results in a more significant error.

For the RF model, we tuned the number of trees. We maintained the maximum depth as indefinite, leading to tree expansion until all nodes were pure. We considered all features to obtain the best split, ensuring minimum bias. As computation for absolute error is slow at each split, we used the squared error as the splitting criterion. This ensured the minimisation of the variance after each split. For the GBR model, we tuned the number of trees, maximum depth of each tree (which affects the randomness in the choice of features in each tree), and subsampling ratio (for stochastic gradient boosting). Larger values of maximum depth, such as the indeterminate depth of the RF model, are not used as GBR functions with weak learners to increase the randomness. The SVM model hyperparameter finetuning involved kernel selection and a choice of the regularisation parameter. Further, in the case of polynomial kernels, the degree of the polynomial was also tuned. For the NN model, we used a fully connected feedforward network where the hyperparameters of the number of layers and number of neurons in a layer were tuned. The activation function ReLU was used to incorporate non-linearity. We used the adam (Kingma and Ba, 2014) optimiser to minimise the loss function. The training process was performed for 500 iterations with early stopping in the event of convergence before completing the iterations. The NN models for each set of hyperparameters converged before the completion of the 500 iterations.

2.4 Performance Evaluation

The testing dataset evaluation metrics used to assess the models' performances are the coefficient of determination (R^2) which represents the percentage deviation between the target and model predictions, $RMSE$ which represents the absolute deviations between the target and the model predictions. Lower R^2 values suggest that the model does not represent the targets well. Values close to one indicate a strong linear correlation. Lower $RMSE$ values are preferable as this quantifies the variance between the targets and predicted values. Additionally, we report the slope and additive bias using reduced major axis (RMA) regression. We used RMA regression slope and bias to ensure symmetry about the $y = 1$ line. This is preferable as there exist uncertainties in both labels and outputs.

ML models are heavily reliant on the availability of training data. To understand the effect of data availability on the model performance, we split the dataset into subsets of iteratively increasing sizes. We trained the models for each subset and computed the evaluation metrics over the testing datasets.

2.5 Feature Importance

The feature importance is represented using permutation importance described in Altmann et al. (2010). Here, we disregard individual features from the model at each iteration and recorded the reduction in evaluation score. This is repeated for each input feature. We normalize the obtained permutation importance for each model and express the importance of each input meteorological variable as a percentage. A comparative analysis of the obtained feature importance is performed on two counts: (a) Features that are most important. Here the most important 10% of the features are considered. Thus the 17 most important meteorological variables out of 168 used are reported. This is represented in Supplementary material S1. (b) importance associated with the accumulation months and the ablation months are summed and graphically for each model and are represented in Fig. 8.

3 Results

This section describes the major outcomes of the study categorized as the role of dataset size for the effective training of each ML model (see Fig 2), the performance and feature importance associated with each ML model. Figure 3 represents the comparative performance of each of the models in terms of the accuracy metrics $RMSE$, R^2, Slope and Additive Bias. A scatter plot of modelled point mass balance and labels is represented in Fig 4. Figures 5, 6 and 7 represent the hyperparameter tuning associated with the models. The feature importance for all input variables summed over the ablation and accumulation months is represented in Fig 8. The most important meteorological variables (10% of total number of variables) associated with each model are represented in Supplementary material S1.

3.1 Role of Training Dataset Size

The number of samples required for training the ML models depends upon the complexity of the model. Thus each of the models used in this study is variably sensible to the number of training samples. We use the evaluation metrics of root mean squared error and correlation coefficient to assess the requirement of training samples for each of the models. Figure 2 depicts the training and testing metrics varying with the size of the training dataset. The training metrics stabilize after 20-30% of the training dataset size for the LR, RF, GBR and SVM models and at 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 stabilizes for training dataset sizes larger than 50%. This suggests that all models have successfully fit the data.

It is interesting to note that RF, GBR and LR models see an increase in training MAE as opposed to a consistent decrease in testing MAE with increasing training samples. This depicts the tendency of these models to overfit the training samples in the case of smaller datasets. This is evident when observing the order of variation in the training and testing evaluation metric for smaller datasets. E.g. GBR depicts a training MAE of 0.357 mwe and a testing MAE of 1.183 mwe at 10% training dataset size and training MAE of 0.659 mwe and a testing MAE of 0.774 mwe at 100% training dataset size. Thus, care must be taken when using RF and GBR for smaller datasets as they are susceptible to overfitting. The performance of the LR model deteriorates for training, and testing performance is also poor. This is not due to overfitting but due to the inability of the model to explain the complex relationship between the inputs and the target. NN requires larger datasets for the training of the model. When compared with other models, SVM is more robust to the size of the dataset.

3.2 Performance of RF modelling

The best performing RF model resulted in a testing $RMSE$ value of 1.083 mwe and an R^2 value of 0.705. The training $RMSE$ values are 0.934 mwe and R^2 value is 0.804. We observe that hyperparameter tuning is not important, and no major variations were observed upon changing the number of estimators. The slope of RF was closest to 1 with a value of 0.752 for the training samples and 0.744 for the testing samples. Both training and testing additive bias were negative, suggesting the model underestimated point mass balance. This is illustrated in Fig 3. Figure 4 depicts a scatter plot of the testing samples estimated and actual point glacier mass balance.

Feature importance analysis using permutation importance considering the 17 (10% of all features) most essential features indicates the RF model is highly influenced by Downward Solar Radiation in January, Net solar radiation for July, Downward thermal radiation in June, Temperature at 2m in June, forecast albedo in February and December, Snow Depth in January and July, snow density and snowmelt in July, sensible heat flux in December, January, March and May, latent heat flux in August and Surface Pressure in June and July. Permutation importance for the RF model summed over the accumulation months highest importance scores for sensible heat flux followed by downward solar radiation and forecast albedo. Each of these variables depict a summed percentage importance between 6-9%. Snow depth and pressure are also important with a summed percentage importance between 3-6%. For the ablation months, only pressure is observed to have a summed percentage importance greater than 6%. Sensible heat flux, net solar radiation, latent heat flux, snow depth, forecast albedo, snow density and temperature at 2m display summed percentage importance between 3-6%.

3.3 Performance of GBR modelling

Tuning the maximum depth permitted for each weak learner tree was important in estimating the best model, and varying the number of weak learner trees during hyperparameter tuning improved performance in the case of smaller depths of the weak learners. Deeper tree structures did not significantly change the model's performance upon changing the number of trees. Stochastic gradient boosting (subsampling at 0.7) resulted in reduced performance. The hyperparameter combination of the best performing GBR model is 100 trees with a maximum depth of 5 nodes. This is depicted in Fig 5. The best performing GBR model resulted in an $RMSE$ value of 1.071 mwe and an R^2 value of 0.713. The training $RMSE$ values are 0.759 mwe and R^2 value is 0.805. Figure 3 depicts the training and testing performance. Figure 4 depicts the scatter plot of actual versus estimated point glacier mass balance.

The most important meteorological inputs for the GBR model are Snowfall in July, Downward solar radiation in January and December, Forecast Albedo in December, January, February, March and May, Sensible Heat Flux in January, March, May, November and December, Temperature at 2m in June and August, snow depth in June and surface pressure in August. Note the marked importance associated with ablation meteorological variables and the months associated with ablation. Permutation importance expressed as a percentage and summed over the accumulation months depicts the most importance to forecast albedo followed by sensible heat flux, with both variables depicting a summed percentage importance greater than 10%. Among other meteorological variables, downward solar radiation, net solar radiation and snow depth in the accumulation months are also important. The ablation months depict higher summed importance values with forecast albedo in these months prominent. Sensible heat flux, latent heat flux, surface pressure, snowfall, snow depth and temperature at 2m above the surface are also important.

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

3.4 Performance of SVM modelling

The SVM model depicted large fluctuations in the test score with changes in the hyperparameters. This is represented in Fig 6. We considered the hyperparameters of the kernel, degree (for polynomial kernel) and regularisation (penalty) factor. The sigmoid kernel resulted in evaluation metrics markedly poorer than the radial basis function (RBF) kernel and polynomial kernels. The sigmoid kernel was excluded from the graphical representation of the test score to emphasise the variations observed in the other kernels. The polynomial kernel at larger degrees consistently performed better than the RBF kernel in the case of regularisation tuning lower than 1. For larger regularisation parameters, the RBF kernels demonstrated better performance. The best-performing model in this study is the RBF kernel (cost: 10.0). Figure 6 depicts the results of hyperparameter tuning for the SVM kernel. The testing $RMSE$ values for the model are 1.085 mwe and R^2 value is 0.704. The training $RMSE$ values are 0.727 mwe and R^2 value is 0.763. This is represented graphically in Fig 3. Figure 4 depicts a scatter plot of actual versus estimated point glacier mass balance.

The permutation importance associated with Sensible Heat Flux March is most important, as is the sensible heat flux associated with April, May, June and December. Latent heat flux in August and October is important. Snowfall in October and snow density for the months of November, December and January are important. The temperature at 2m above the surface in June and July, downward solar radiation in December and forecast albedo in August, October and December are important. Summing the percentage importance over the accumulation and ablation months, we observe that sensible heat flux in the accumulation months is most important, followed by snow density and downward solar radiation. These three variables depict a summed percentage importance of more than 6%. The temperature at 2m above the ground and forecast albedo depict importance between 3-6% for the accumulation months. For the ablation months, sensible heat flux continues to depict a summed percentage importance of more than 6%. Latent heat flux, snow density, forecast albedo and temperature at 2m above the surface also depict a summed percentage importance between 3-6%.

3.5 Performance of NN modelling

The NN model performance is highly susceptible to hyperparameter selection. We varied the number of hidden layers in the network and the number of neurons in each hidden layer. Figure 7 depicts the variation in performance of the model for each of these cases. On the left is the variation in the number of neurons for a single hidden layer. A larger number of hidden neurons permits more combinations of the inputs that can affect the targets. The improved performance with the increasing size of neurons illustrates the role of the complexity of the model in estimating mass balance. Increasing the number of layers also affects the performance of the NN model, with the best performance obtained using two hidden layers. This further emphasises the importance of incorporating non-linear elements in estimating point mass balance. A larger number of hidden layers did not significantly improve performance as the larger number of parameters demanded a larger training dataset to avoid overfitting and to complete the training. The testing $RMSE$ values for the best performing model are 1.096 mwe and R^2 value is 0.697. The training $RMSE$ values are 0.773 mwe and R^2 value is 0.763. This is represented in Fig 3.

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

The most important meteorological variables in terms of the percentage permutation importance for the NN model are the Sensible Heat Flux for March, April and May, Latent Heat Flux in July, Surface Pressure in February, the Net Solar Radiation in May and September, downward solar radiation in December and forecast albedo in July. Snow Density in December and the snow depth January, February, April, July, September, October and December are important. We see that snow depth across the year dominates the important meteorological inputs for this model. Upon summing the percentage importance for the accumulation and ablation months, we observe that snow depth is the most important for both accumulation and ablation months. Snow density, pressure, sensible heat flux and downward solar radiation are also important in the accumulation months, with a summed percentage importance value between 3-6%. For the ablation months, net solar radiation is also important. Snow density, forecast albedo, latent heat flux and sensible heat flux are also important, with summed percentage importance values between 3-6%.

3.6 Performance of LR modelling

The testing $RMSE$ values for the LR model are 1.248 mwe and R^2 value is 0.577 and the training $RMSE$ values are 1.197 mwe and R^2 value is 0.608. This is depicted in Fig 3.

Snow depth over most of the year is the most important feature for the model, with surface pressure also playing an important role. Other features do not depict as high an importance value. However, relative importance varies across the months.

4 Discussion

4.1 Comparison of Model Performance and Associated Errors

The performance of each of the models was evaluated using an independent test dataset. The GBR model resulted in the best testing performance metrics. It performs marginally better than the RF model. SVM and NN models perform comparably, with the bias performance of NN being better but RMSE being worse. RF, GBR, SVM and NN significantly improve upon the LR model's metrics. The ability of all non-linear models to outperform the linear model is further depicted in each model's scatter plot (Fig. 4). This is in agreement with similar studies in other domains, such as King et al. (2020) who showed that tree-based models such as RF were preferable to LR models for the bias-correction of snow water equivalent and Rasouli et al. (2012) who depicted the efficacy of non-linear models in estimation of streamflow when compared to linear models.

The performance of all models is affected by the uncertainties associated with the input features and targets. Inherent errors exist in point mass balance estimates as heterogeneity is not captured sufficiently by the available measurements (Zemp et al., 2013; Van Tricht et al., 2021). Of the 727 locations with uncertainty estimation performed, we note a mean uncertainty of 0.062 mwe, which can adversely impact performance evaluation. The uncertainty estimates for the remaining point locations are unknown; hence, their impact is not constrained. Input meteorological reanalysis data do not fully reflect point scale data as it has a coarse resolution. Further, reanalysis data can result in bias, especially in locations without sufficient ground stations (Guidicelli et al., 2022). Thus, we suggest using a bias correction step such as that proposed by Cucchi et al. (2020) in the case of RF, GBR and SVM models. Finally, in this study, we did not consider the effect of topography and debris cover for the models. This can lead to inflated RMSE values.

4.2 Role of Training Dataset Availability

The testing performance improves with a larger number of training samples. However, the rate of improvement reduces when including more datasets. This indicates that the training is successful for all models, but improvement in model performance is possible when including more data samples. The RF and GBR models overfit the training samples in the case of smaller datasets. The NN model training and testing metrics depict improved performance with training size. The NN model had the most trainable parameters and hence is most data-intensive. A larger number of training samples is essential for models with a larger number of trainable parameters. The training performance of the LR model deteriorates with increasing training samples. While the graph appears similar to the RF and GBR training graphs, the relatively close training and testing metrics values suggest that overfitting is not the likely cause. Rather, it suggests that the model cannot explain the non-linear relationship between the inputs and the target.

Further, Fig. 2 represents each model's variation in training and testing evaluation metrics. Each model was trained and tested over each dataset size. For each model, the box plots are generated utilising the outcome of the models developed using varying training dataset sizes. The training performance, as expected, is better than the testing performance as the model parameters are tuned to fit this dataset. The range of values is more extensive for the testing errors as a result of overfitting in the case of smaller datasets. In such cases, the use of the SVM model yields better results.

4.3 Unraveling the Physics using Machine Learning-Derived Feature Importance

Assuming a winter accumulation type glacier, we expect the months of November to March to be dominated by accumulation processes and June to September to be dominated by ablation processes. Analysis of the permutation importance (by percentage) of the features of each model was studied month-wise based on a physical understanding of which season-specific features will be most important. Figure 8 represents the summed feature importance for each input variable in the accumulation and ablation months. We sum the percentage importance rather than the feature importance values to permit comparison between models. We expect temperature (2m) for ablation seasons to be significant compared to temperatures in the accumulation season. This is not well reflected when using the LR model. While all the ML models show the reduced importance of temperature in the accumulation months, it is most pronounced in the case of the RF and GBR models. A similar trend is expected for the downward thermal radiation and snowmelt. Here, too the LR model does not reflect the expected outcome. All ML models depict reduced importance in the accumulation months, with a pronounced reduction observed in the RF and GBR models. In the case of snowmelt, all ML models and the LR model follow the expected response. Snow depth throughout the year is important when considering snow density. We expect the depth in the ablation months to be important. All models portray this except the SVM model. We observe that the LR model relies heavily on snow depth to estimate the mass balance. The SVM model reports the exaggerated importance of snow density in the accumulation months. While we expect more importance to precipitation terms such as total precipitation and snowfall in the accumulation months, we do not observe this for any model. The LR model

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

did show a weak reduction in the importance of total precipitation and snowfall. However, the ML models showed only a weak reduction or a weak increase in importance. This can be a result of the effect of lower resolution grid cells associated with the meteorological data. Net solar radiation and albedo are important ablation components. Albedo over snow-covered regions is higher than that of exposed ice or firn. Hence, the role of albedo is less important in the ablation period. The expected trend is observed in the RF and GBR models and inverted in the case of NN, SVM and LR models. Thus we see that the ML models well represent the importance of the ablation features. This is in line with the predominantly negative mass balance observed in in-situ measurements.

We can observe that the importance associated with the meteorological variables is not dominated solely by total precipitation and temperature, as with temperature index models. Thus, ML modelling can represent the contributions of a complete set of variables with lesser complexity and ease of use than physical models. This also emphasises the requirement for ML models to use all meteorological variables of interest, as opposed to a subset of them. This is the case with studies such as Bolibar et al. (2020). Further, our results agree with the studies conducted by Steiner et al. (2005) and Bolibar et al. (2022) in that artificial neural networks capture the complexity of the mass balance estimation using non-linear relationships between inputs. However, we propose that other ML models, notably ensemble tree-based methods, can be used for equivalent to improved estimates in case of fewer real-world data samples for training. This has also been observed in other studies (e.g. Bair et al., 2018) For this case, feature importance derived using permutation importance for the ensemble-based models, RF and GBR, represented the expected role of meteorological variables in determining feature importance. The evaluation metrics also emphasise the performance of these models.

5 Conclusions

In this study, we constructed 4 ML models to estimate point glacier mass balance for the RGI order one region 11: Central Europe. We used the ERA5-Land reanalysis meteorological data to train the models against point measurements of glacier mass balance obtained from the FoG database. In addition to the NN model, which is being increasingly utilised for glacier mass balance estimation, we used other classes of ML models, such as ensemble tree-based models: RF and GBR, and the kernel-based model: SVM. We compared these ML models with an LR model commonly used for mass balance modelling. Care must be taken to tune the hyperparameters for the GBR, NN and SVM models. We observe that for these models, hyperparameter tuning was beneficial for improving the estimates of glacier mass balance. For smaller datasets, ensemble models such as RF and GBR depict overfitting. The NN model requires more data samples for effective training. We suggest the use of a kernel-based model in such situations. The SVM model can effectively be used in the case of a smaller number of data samples, which is characteristic of real-world datasets. The LR model is consistently unable to capture the complexity of the data and underperforms. For larger datasets, ensemble models such as RF and GBR perform slightly better in terms of R^2 and $RMSE$. However, NN models depict the least bias. The meteorological variables obtained from reanalysis datasets are associated with high bias. Using NN and LR models permits us to use them directly. For other models, bias correction should be incorporated in the preprocessing. Representation of real-world features is also performed more effectively by RF and GBR models. These models indicate the importance of ablation features dominating the mass balance estimates. This is expected as the mass balance measurements are primarily negative. Further, feature importance suggests that features such as forecast albedo, sensible heat flux, latent heat flux and net solar radiation also play a pivotal role in estimating point mass balance. Thus inclusion of these additional variables might be of importance for future studies.

Code and data availability. The data used for the study is the monthly mean ERA5-Land reanalysis product for inputs features and point mass balance measurements from the Fluctuation of Glaciers database for the labels. The code for processing the data and applying all models used in this study is available at https://github.com/RituAnilkumar/pt-gmb-ml

Author contributions. RA, RB and DJC were involved in the design of the study. RA wrote the code for the study and produced the figures, tables and first draft of the manuscript using inputs from all authors. RB, DC and SPA proofread and edited the manuscript. RA performed the first level of analysis, which was augmented by inputs from RB, DJC and SPA.

Competing interests. The authors report that this study contains no competing interests

Acknowledgements. The authors acknowledge the engaging discussions with peers, most notably Dr Aniket Chakraborty, who always lent a patient ear and sound suggestions to roadblocks along the way.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T.: Permutation importance: a corrected feature importance measure, Bioinformatics, 26, 1340–1347, https://doi.org/10.1093/bioinformatics/btq134, 2010.

Bair, E. H., Abreu Calfa, A., Rittger, K., and Dozier, J.: Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan, The Cryosphere, 12, 1579–1594, https://doi.org/10.5194/tc-12-1579-2018, 2018.

Bash, E. A., Moorman, B. J., and Gunther, A.: Detecting Short-Term Surface Melt on an Arctic Glacier Using UAV Surveys, Remote Sensing, 10, https://doi.org/10.3390/rs10101547, 2018.

Bolibar, J., Rabatel, A., Gouttevin, I., Galiez, C., Condom, T., and Sauquet, E.: Deep learning applied to glacier evolution modelling, The Cryosphere, 14, 565–584, https://doi.org/10.5194/tc-14-565-2020, 2020.

Bolibar, J., Rabatel, A., Gouttevin, I., Zekollari, H., and Galiez, C.: Nonlinear sensitivity of glacier mass balance to future climate change unveiled by deep learning, Nature communications, 13, 1–11, https://doi.org/https://doi.org/10.1038/s41467-022-28033-0, 2022.

Braithwaite, R. J.: Can the Mass Balance of a Glacier be Estimated from its Equilibrium-Line Altitude?, Journal of Glaciology, 30, 364–368, https://doi.org/10.3189/S0022143000006237, 1984.

Breiman, L.: Bagging predictors, Machine learning, 24, 123–140, https://doi.org/https://doi.org/10.1007/BF00058655, 1996.

Breiman, L.: Random forests, Machine learning, 45, 5–32, https://doi.org/https://doi.org/10.1023/A:1010933404324, 2001.

Carturan, L., Cazorzi, F., and Dalla Fontana, G.: Enhanced estimation of glacier mass balance in unsampled areas by means of topographic data, Annals of Glaciology, 50, 37–46, https://doi.org/10.3189/172756409787769519, 2009.

Chi, J. and Kim, H.-c.: Prediction of Arctic Sea Ice Concentration Using a Fully Data Driven Deep Neural Network, Remote Sensing, 9, https://doi.org/10.3390/rs9121305, 2017.

Cortes, C. and Vapnik, V.: Support-vector networks, Machine learning, 20, 273–297, https://doi.org/https://doi.org/10.1007/BF00994018, 1995.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth System Science Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

Dietterich, T. G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine learning, 40, 139–157, https://doi.org/https://doi.org/10.1023/A:1007607513941, 2000.

Dobhal, D., Pratap, B., Bhambri, R., and Mehta, M.: Mass balance and morphological changes of Dokriani Glacier (1992–2013), Garhwal Himalaya, India, Quaternary Science Advances, 4, 100 033, https://doi.org/https://doi.org/10.1016/j.qsa.2021.100033, 2021.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232, 2001.

Gabbi, J., Carenzo, M., Pellicciotti, F., Bauder, A., and Funk, M.: A comparison of empirical and physically based glacier surface melt models for long-term simulations of glacier response, Journal of Glaciology, 60, 1140–1154, https://doi.org/10.3189/2014JoG14J011, 2014.

Gerbaux, M., Genthon, C., Etchevers, P., Vincent, C., and Dedieu, J.: Surface mass balance of glaciers in the French Alps: distributed modeling and sensitivity to climate change, Journal of Glaciology, 51, 561–572, https://doi.org/10.3189/172756505781829133, 2005.

Guidicelli, M., Huss, M., Gabella, M., and Salzmann, N.: Snow accumulation over the world's glaciers (1981–2021) inferred from climate reanalyses and machine learning, The Cryosphere Discussions, 2022, 1–47, https://doi.org/10.5194/tc-2022-69, 2022.

Guo, X., Chen, Y., Liu, X., and Zhao, Y.: Extraction of snow cover from high-resolution remote sensing imagery using deep learning on a small dataset, Remote Sensing Letters, 11, 66–75, https://doi.org/10.1080/2150704X.2019.1686548, 2020.

Haq, M. A., Jain, K., and Menon, K.: Modelling of Gangotri glacier thickness and volume using an artificial neural network, International Journal of Remote Sensing, 35, 6035–6042, https://doi.org/10.1080/01431161.2014.943322, 2014.

Haq, M. A., Azam, M. F., and Vincent, C.: Efficiency of artificial neural networks for glacier ice-thickness estimation: a case study in western Himalaya, India, Journal of Glaciology, 67, 671–684, https://doi.org/10.1017/jog.2021.19, 2021.

Hoinkes, H. C.: Glacier Variation and Weather, Journal of Glaciology, 7, 3–18, https://doi.org/10.3189/S0022143000020384, 1968.

Hornik, K.: Approximation capabilities of multilayer feedforward networks, Neural Networks, 4, 251–257, https://doi.org/https://doi.org/10.1016/0893-6080(91)90009-T, 1991.

Huss, M., Farinotti, D., Bauder, A., and Funk, M.: Modelling runoff from highly glacierized alpine drainage basins in a changing climate, Hydrological Processes, 22, 3888–3902, https://doi.org/https://doi.org/10.1002/hyp.7055, 2008.

Jouvet, G., Cordonnier, G., Kim, B., Lüthi, M., Vieli, A., and Aschwanden, A.: Deep learning speeds up ice flow modelling by several orders of magnitude, Journal of Glaciology, p. 1–14, https://doi.org/10.1017/jog.2021.120, 2021.

Kan, X., Zhang, Y., Zhu, L., Xiao, L., Wang, J., Tian, W., and Tan, H.: Snow cover mapping for mountainous areas by fusion of MODIS L1B and geographic data based on stacked denoising auto-encoders, Computers, Materials and Continua, 57, 49–68, https://doi.org/10.32604/cmc.2018.02376, 2018.

King, F., Erler, A. R., Frey, S. K., and Fletcher, C. G.: Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada, Hydrology and Earth System Sciences, 24, 4887–4902, https://doi.org/10.5194/hess-24-4887-2020, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, https://doi.org/10.48550/ARXIV.1412.6980, 2014.

Kuhn, M., Dreiseitl, E., Hofinger, S., Markl, G., Span, N., and Kaser, G.: Measurements and models of the mass balance of hintereisferner, Geografiska Annaler: Series A, Physical Geography, 81, 659–670, https://doi.org/10.1111/1468-0459.00094, 1999.

Le Meur, E., Gerbaux, M., Schäfer, M., and Vincent, C.: Disappearance of an Alpine glacier over the 21st Century simulated from modeling its future surface mass balance, Earth and Planetary Science Letters, 261, 367–374, https://doi.org/https://doi.org/10.1016/j.epsl.2007.07.022, 2007.

Lefauconnier, B. and Hagen, J.: Glaciers and Climate in Svalbard: Statistical Analysis and Reconstruction of the Brøggerbreen Mass Balance for the Last 77 Years, Annals of Glaciology, 14, 148–152, https://doi.org/10.3189/S0260305500008466, 1990.

Li, J., Wang, C., Wang, S., Zhang, H., Fu, Q., and Wang, Y.: Gaofen-3 sea ice detection based on deep learning, in: 2017 Progress in Electromagnetics Research Symposium - Fall (PIERS - FALL), pp. 933–939, https://doi.org/10.1109/PIERS-FALL.2017.8293267, 2017.

Liu, L.: A Review of Deep Learning for Cryospheric Studies, chap. 17, pp. 258–268, John Wiley and Sons, Ltd, https://doi.org/https://doi.org/10.1002/9781119646181.ch17, 2021.

Lliboutry, L.: Multivariate Statistical Analysis of Glacier Annual Balances, Journal of Glaciology, 13, 371–392, https://doi.org/10.3189/S0022143000023169, 1974.

Lu, Y., Zhang, Z., Shangguan, D., and Yang, J.: Novel Machine Learning Method Integrating Ensemble Learning and Deep Learning for Mapping Debris-Covered Glaciers, Remote Sensing, 13, https://doi.org/10.3390/rs13132595, 2021.

Manciati, C., Villacís, M., Taupin, J.-D., Cadier, E., Galárraga-Sánchez, R., and Cáceres, B.: Empirical mass balance modelling of South American tropical glaciers: case study of Antisana volcano, Ecuador, Hydrological Sciences Journal, 59, 1519–1535, https://doi.org/10.1080/02626667.2014.888490, 2014.

Masiokas, M. H., Christie, D. A., Le Quesne, C., Pitte, P., Ruiz, L., Villalba, R., Luckman, B. H., Berthier, E., Nussbaumer, S. U., González-Reyes, A., McPhee, J., and Barcaza, G.: Reconstructing the annual mass balance of the Echaurren Norte glacier (Central Andes, 33.5° S) using local and regional hydroclimatic data, The Cryosphere, 10, 927–940, https://doi.org/10.5194/tc-10-927-2016, 2016.

McCulloch, W. S. and Pitts, W.: A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics, 5, 115–133, https://doi.org/10.1007/BF02478259, 1943.

Mohajerani, Y., Jeong, S., Scheuchl, B., Velicogna, I., Rignot, E., and Milillo, P.: Automatic delineation of glacier grounding lines in differential interferometric synthetic-aperture radar data using deep learning, Scientific reports, 11, 1–10, https://doi.org/https://doi.org/10.1038/s41598-021-84309-3, 2021.

Moya Quiroga, V., Mano, A., Asaoka, Y., Kure, S., Udo, K., and Mendoza, J.: Snow glacier melt estimation in tropical Andean glaciers using artificial neural networks, Hydrology and Earth System Sciences, 17, 1265–1280, https://doi.org/10.5194/hess-17-1265-2013, 2013.

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Muñoz Sabater, J.: ERA5-Land monthly averaged data from 1981 to present., https://doi.org/10.24381/cds.68d2bb3, (Accessed on 7-DEC-2021), 2019.

Muñoz Sabater, J.: ERA5-Land monthly averaged data from 1950 to 1980., https://doi.org/10.24381/cds.68d2bb3, (Accessed on 27-DEC-2021), 2021.

Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, Frontiers in neurorobotics, 7, 21, 2013.

Nijhawan, R., Das, J., and Raman, B.: A hybrid of deep learning and hand-crafted features based approach for snow cover mapping, International Journal of Remote Sensing, 40, 759–773, https://doi.org/10.1080/01431161.2018.1519277, 2019.

Pratap, B., Dobhal, D. P., Bhambri, R., Mehta, M., and Tewari, V. C.: Four decades of glacier mass balance observations in the Indian Himalaya, Regional Environmental Change, 16, 643–658, https://doi.org/https://doi.org/10.1007/s10113-015-0791-4, 2016.

Rabatel, A., Dedieu, J. P., and Vincent, C.: Spatio-temporal changes in glacier-wide mass balance quantified by optical remote sensing on 30 glaciers in the French Alps for the period 1983–2014, Journal of Glaciology, 62, 1153–1166, https://doi.org/10.1017/jog.2016.113, 2016.

Radić, V. and Hock, R.: Regionally differentiated contribution of mountain glaciers and ice caps to future sea-level rise, Nature Geoscience, 4, 91–94, https://doi.org/https://doi.org/10.1038/ngeo1052, 2011.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414-415, 284–293, https://doi.org/https://doi.org/10.1016/j.jhydrol.2011.10.039, 2012.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

RGI: Randolph Glacier Inventory (RGI) – A Dataset of Global Glacier Outlines: Version 6.0. Technical Report, https://doi.org/https://doi.org/10.7265/N5-RGI-60, 2017.

Sauter, T., Arndt, A., and Schneider, C.: COSIPY v1.3 – an open-source coupled snowpack and ice surface energy and mass balance model, Geoscientific Model Development, 13, 5645–5662, https://doi.org/10.5194/gmd-13-5645-2020, 2020.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200097, https://doi.org/10.1098/rsta.2020.0097, 2021.

Seidou, O., Ouarda, T. B. M. J., Bilodeau, L., Hessami, M., St-Hilaire, A., and Bruneau, P.: Modeling ice growth on Canadian lakes using artificial neural networks, Water Resources Research, 42, https://doi.org/https://doi.org/10.1029/2005WR004622, 2006.

Shean, D. E., Bhushan, S., Montesano, P., Rounce, D. R., Arendt, A., and Osmanoglu, B.: A Systematic, Regional Assessment of High Mountain Asia Glacier Mass Balance, Frontiers in Earth Science, 7, https://doi.org/10.3389/feart.2019.00363, 2020.

Steiner, D., Walter, A., and Zumbühl, H.: The application of a non-linear back-propagation neural network to study the mass balance of Grosse Aletschgletscher, Switzerland, Journal of Glaciology, 51, 313–323, https://doi.org/10.3189/172756505781829421, 2005.

Thibert, E., Blanc, R., Vincent, C., and Eckert, N.: Glaciological and volumetric mass-balance measurements: error analysis over 51 years for Glacier de Sarennes, French Alps, Journal of Glaciology, 54, 522–532, https://doi.org/10.3189/002214308785837093, 2008.

Trantow, T. and Herzfeld, U. C.: Spatiotemporal mapping of a large mountain glacier from CryoSat-2 altimeter data: surface elevation and elevation change of Bering Glacier during surge (2011–2014), International Journal of Remote Sensing, 37, 2962–2989, https://doi.org/10.1080/01431161.2016.1187318, 2016.

Tshering, P. and Fujita, K.: First in situ record of decadal glacier mass balance (2003–2014) from the Bhutan Himalaya, Annals of Glaciology, 57, 289–294, https://doi.org/10.3189/2016AoG71A036, 2016.

Van Tricht, L., Huybrechts, P., Van Breedam, J., Vanhulle, A., Van Oost, K., and Zekollari, H.: Estimating surface mass balance patterns from unoccupied aerial vehicle measurements in the ablation area of the Morteratsch–Pers glacier complex (Switzerland), The Cryosphere, 15, 4445–4464, https://doi.org/10.5194/tc-15-4445-2021, 2021.

Vapnik, V.: The Nature of Statistical Learning Theory, Springer science & business media, https://doi.org/https://doi.org/10.1007/978-1-4757-2440-0, 1999.

Vincent, C., Soruco, A., Azam, M. F., Basantes-Serrano, R., Jackson, M., Kjøllmoen, B., Thibert, E., Wagnon, P., Six, D., Rabatel, A., Ramanathan, A., Berthier, E., Cusicanqui, D., Vincent, P., and Mandal, A.: A Nonlinear Statistical Model for Extracting a Climatic Signal From Glacier Mass Balance Measurements, Journal of Geophysical Research: Earth Surface, 123, 2228–2242, https://doi.org/https://doi.org/10.1029/2018JF004702, 2018.

Wang, J., Yuan, Q., Shen, H., Liu, T., Li, T., Yue, L., Shi, X., and Zhang, L.: Estimating snow depth by combining satellite data and ground-based observations over Alaska: A deep learning approach, Journal of Hydrology, 585, 124 828, https://doi.org/https://doi.org/10.1016/j.jhydrol.2020.124828, 2020.

Werder, M. A., Huss, M., Paul, F., Dehecq, A., and Farinotti, D.: A Bayesian ice thickness estimation model for large-scale applications, Journal of Glaciology, 66, 137–152, https://doi.org/10.1017/jog.2019.93, 2020.

WGMS: Fluctuations of Glaciers Database, https://doi.org/DOI:10.5904/wgms-fog-2021-05, 2021.

White, I. D., Harrison, S. J., and Mottershead, D. N.: Environmental systems: an introductory text, Psychology Press, 1998.

Wu, K., Liu, S., Jiang, Z., Xu, J., Wei, J., and Guo, W.: Recent glacier mass balance and area changes in the Kangri Karpo Mountains from DEMs and glacier inventories, The Cryosphere, 12, 103–121, https://doi.org/10.5194/tc-12-103-2018, 2018.

Xie, Z., Asari, V. K., and Haritashya, U. K.: Evaluating deep-learning models for debris-covered glacier mapping, Applied Computing and Geosciences, 12, 100 071, https://doi.org/https://doi.org/10.1016/j.acags.2021.100071, 2021.

Zemp, M., Thibert, E., Huss, M., Stumm, D., Rolstad Denby, C., Nuth, C., Nussbaumer, S. U., Moholdt, G., Mercer, A., Mayer, C., Joerg, P. C., Jansson, P., Hynek, B., Fischer, A., Escher-Vetter, H., Elvehøy, H., and Andreassen, L. M.: Reanalysing glacier mass balance measurement series, The Cryosphere, 7, 1227–1245, https://doi.org/10.5194/tc-7-1227-2013, 2013.

Zemp, M., Nussbaumer, S. U., Gärtner-Roer, I., Bannwart, J., Paul, F., and Hoelzle, M.: Global Glacier Change Bulletin Nr. 4 (2018-2019), Tech. rep., World Glacier Monitoring Service, Zürich, https://doi.org/10.5167/uzh-209777, 2021.

545 Zhang, E., Liu, L., and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13, 1729–1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

Zhu, L., Zhang, Y., Wang, J., Tian, W., Liu, Q., Ma, G., Kan, X., and Chu, Y.: Downscaling Snow Depth Mapping by Fusion of Microwave and Optical Remote-Sensing Data Based on Deep Learning, Remote Sensing, 13, https://doi.org/10.3390/rs13040584, 2021.

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)

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

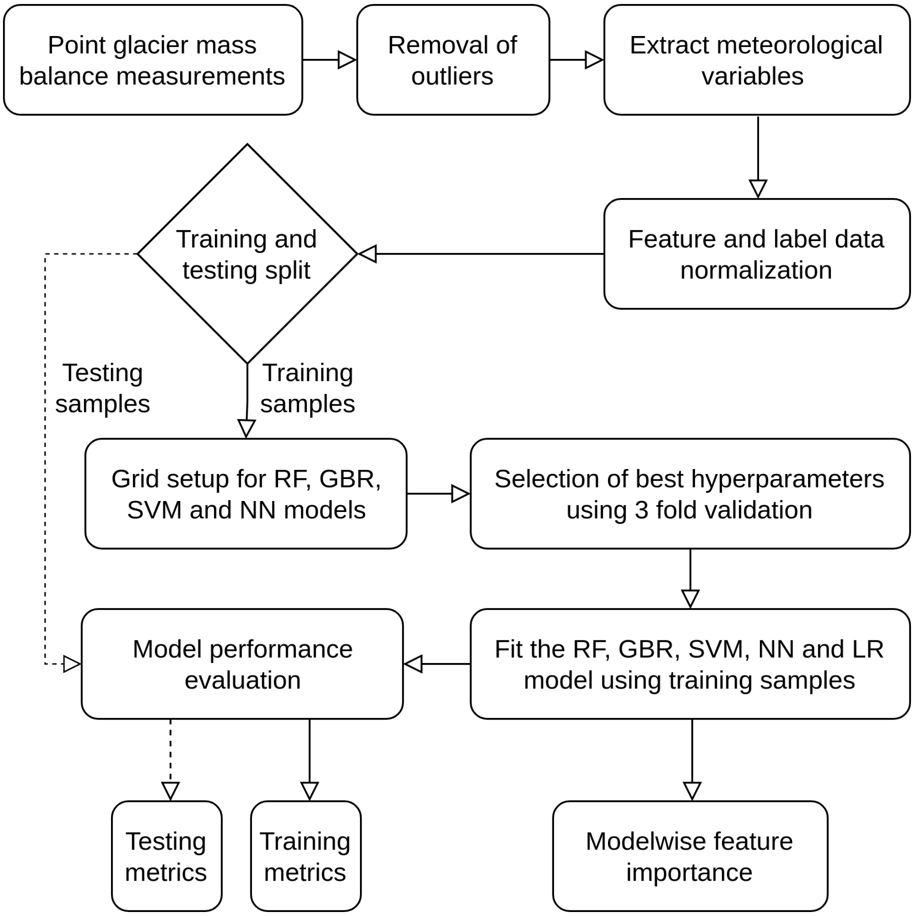
Point glacier mass balance measurements
Removal of outliers
Extract meteorological variables
Feature and label data normalization
Training and testing split
Testing samples
Training samples
Grid setup for RF, GBR, SVM and NN models
Selection of best hyperparameters using 3 fold validation
Fit the RF,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

Figure 1. Flowchart of the methodology

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

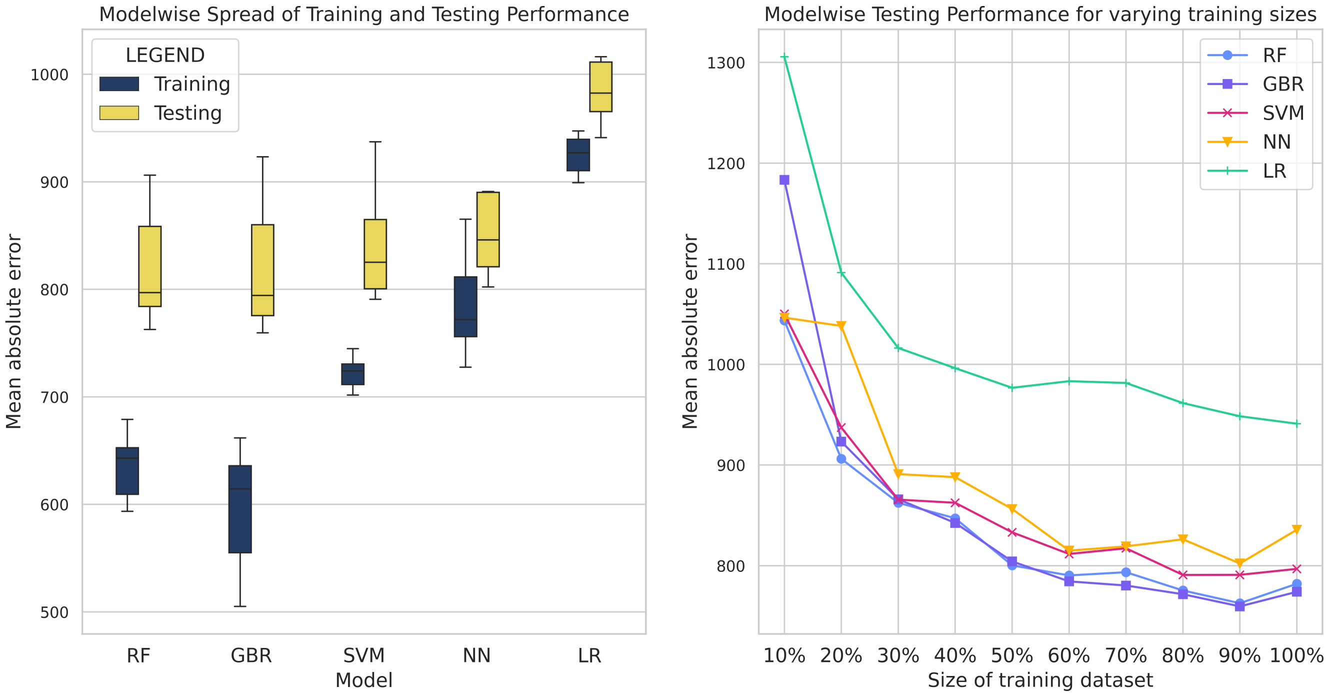

Figure 2. Training and testing $RMSE$ (in mm we) and r values for varying the size of the training dataset for each of the models: Random Forest (RF), Gradient Boosted Regression (GBR), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Linear Regression (LR). The training dataset size is expressed as a percentage of the largest size of the training dataset i.e. 6416 data points

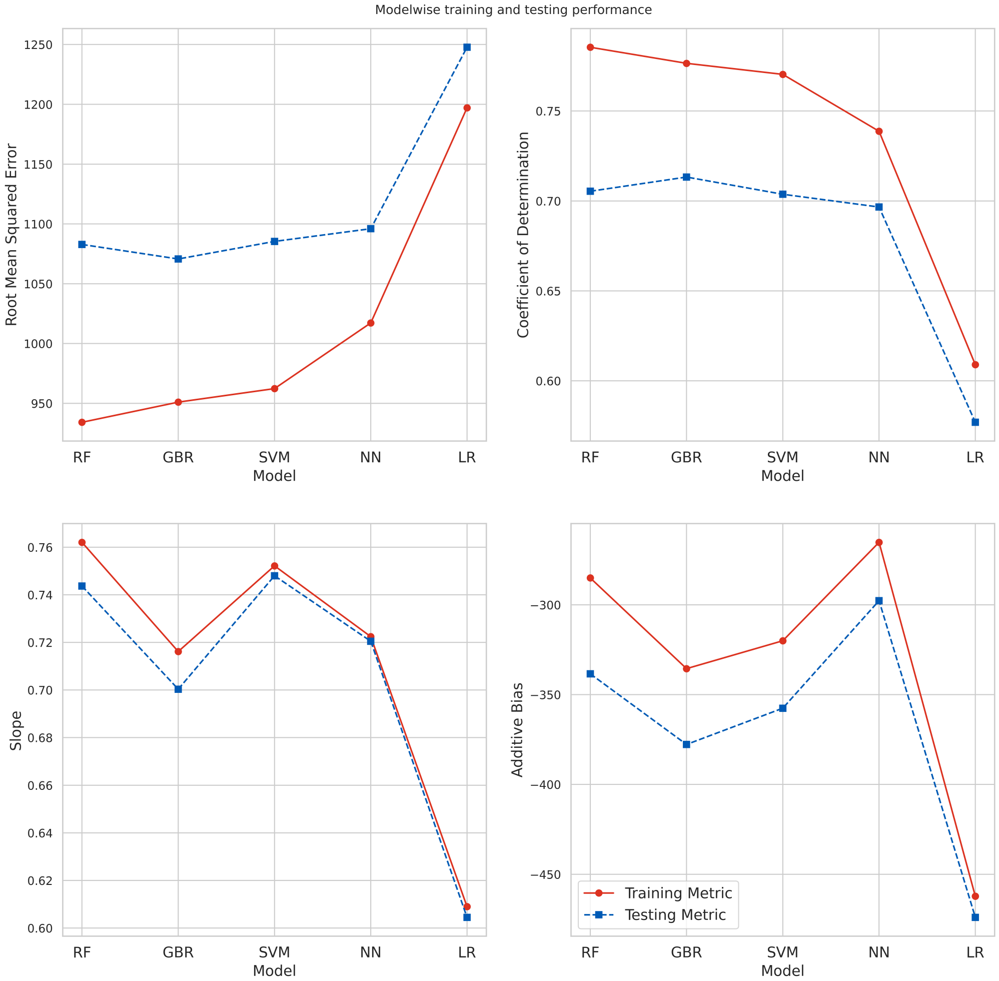

Figure 3. Training and testing performance of each of the models: Random Forest (RF), Gradient Boosted Regression (GBR), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Linear Regression (LR)

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

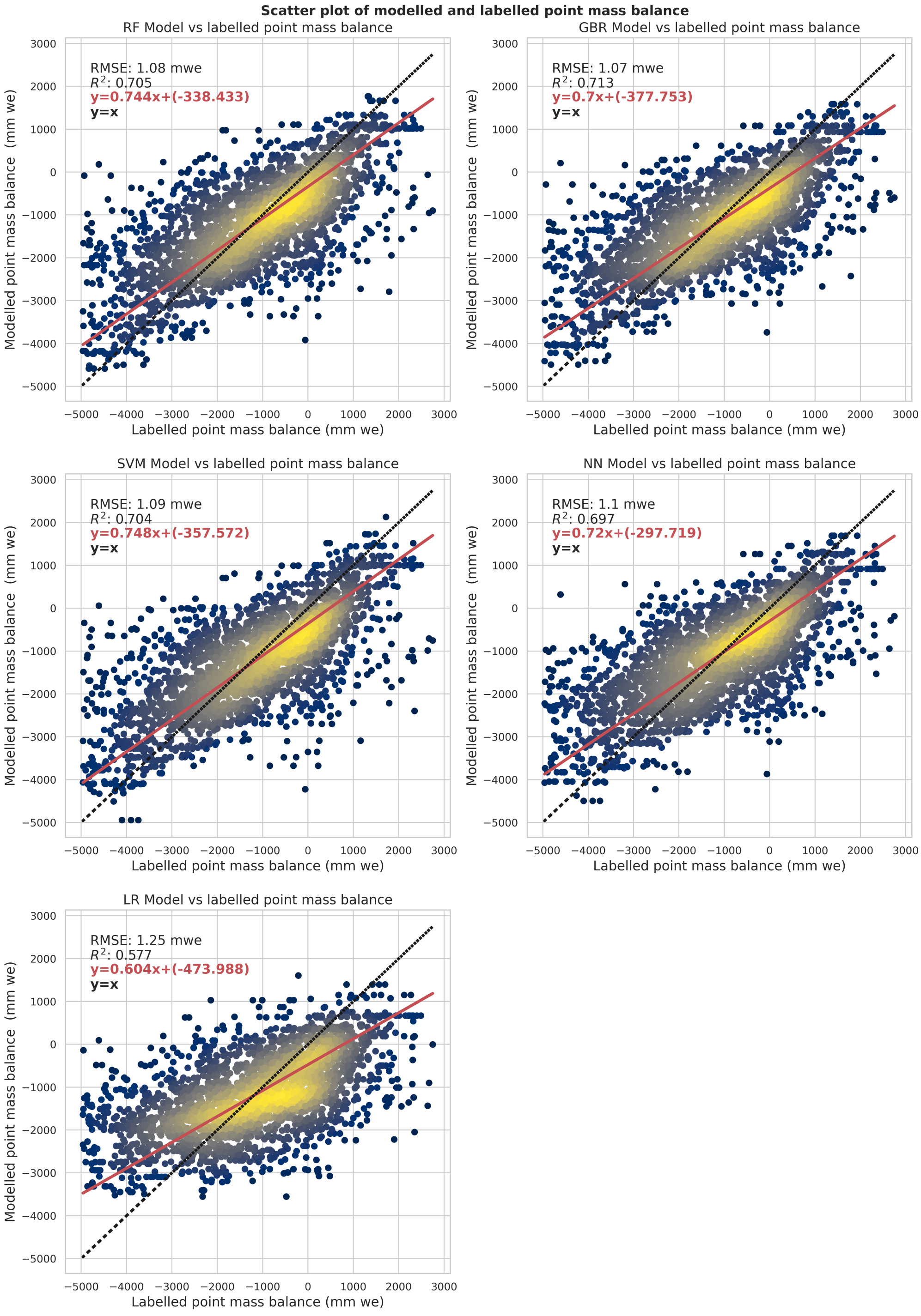

Figure 4. Testing scatter plot depicting the performance for each of the models: Random Forest (RF), Gradient Boosted Regression (GBR), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Linear Regression (LR)

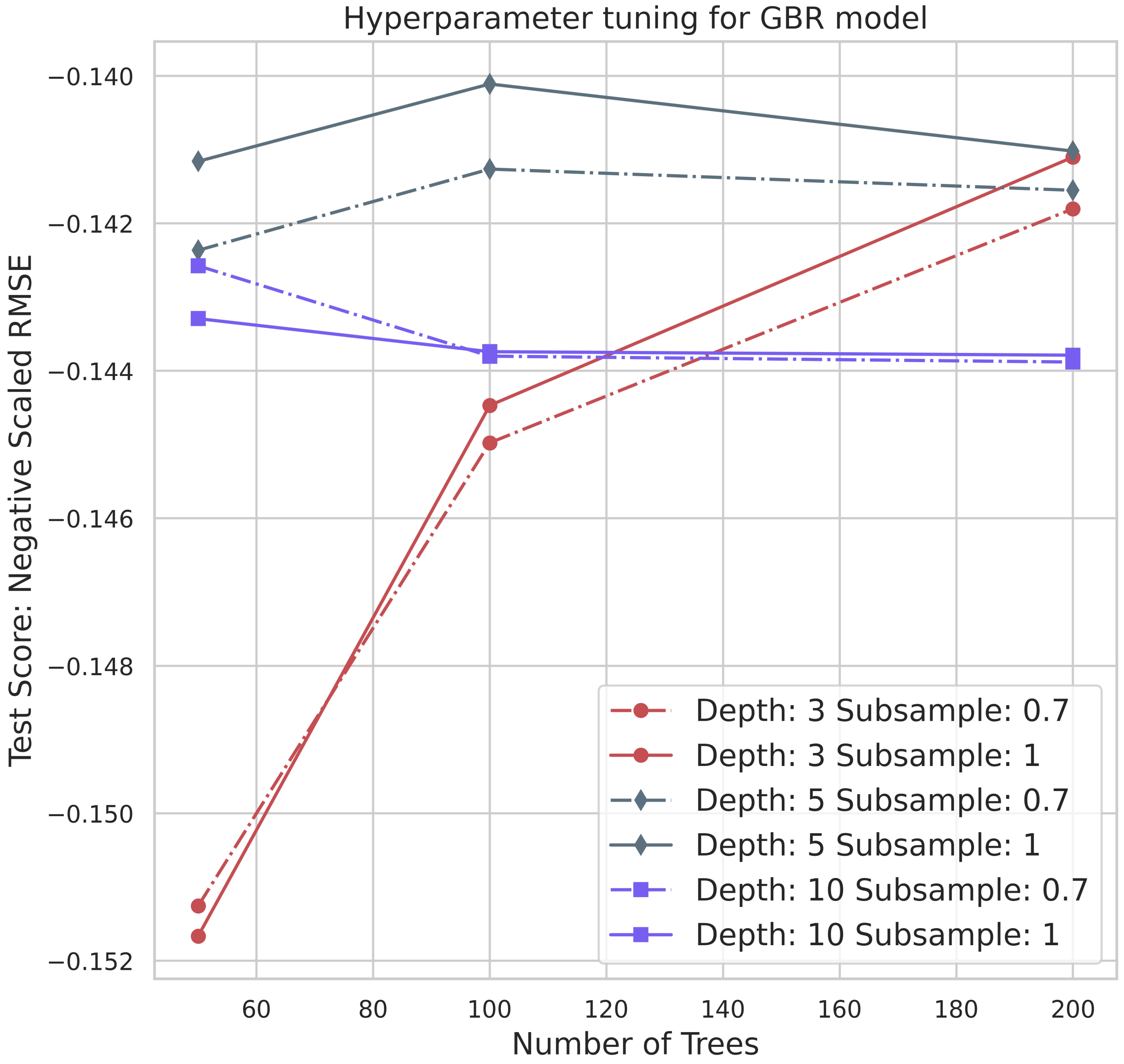

Figure 5. Hyperparameter tuning for the gradient boosted regressor model varying the number of trees, maximum depth of each tree and subsampling fraction

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

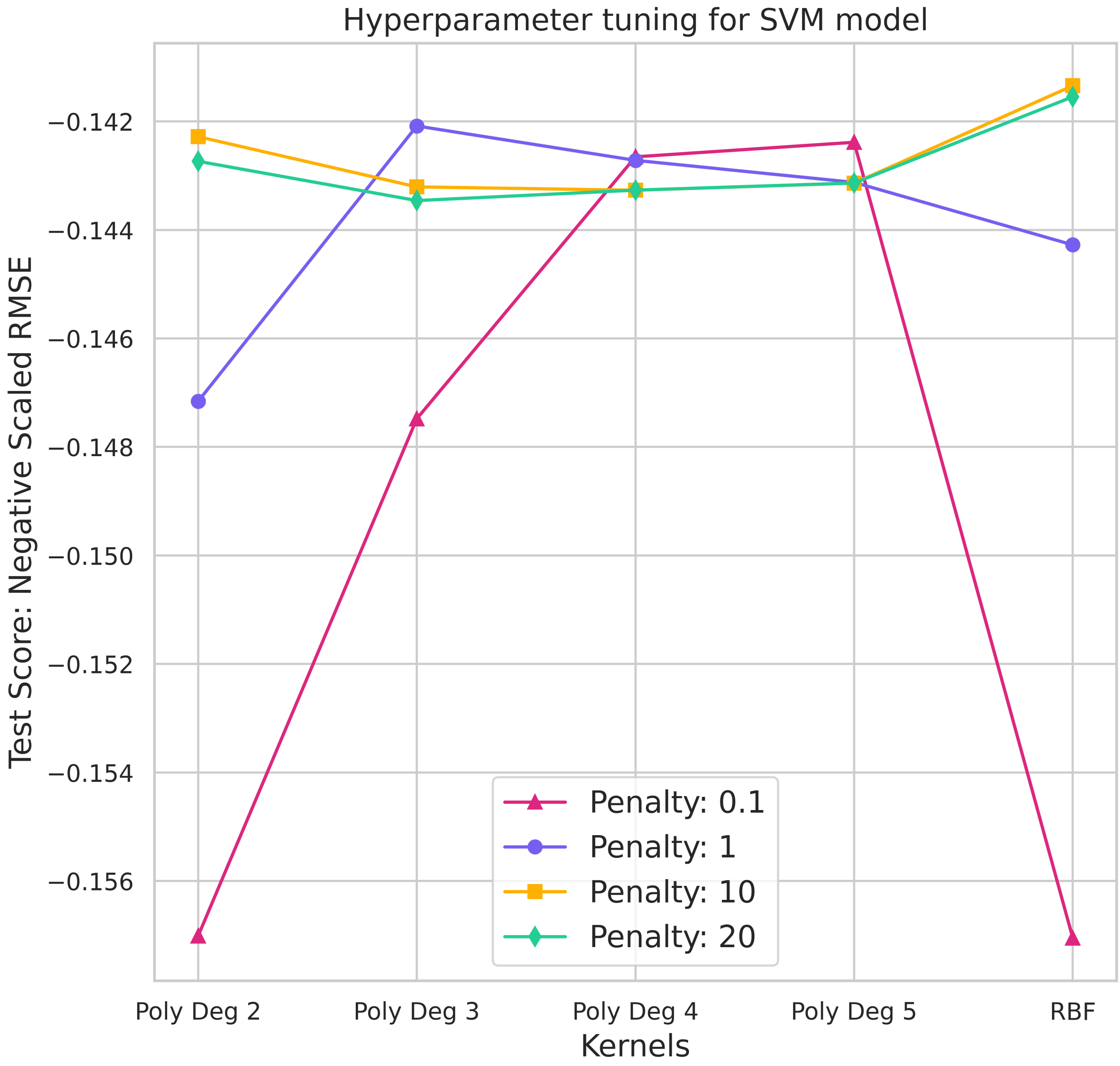

Figure 6. Hyperparameter tuning for the support vector machine model varying the kernels and regularization values as well as degree in case of the polynomial kernel.

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

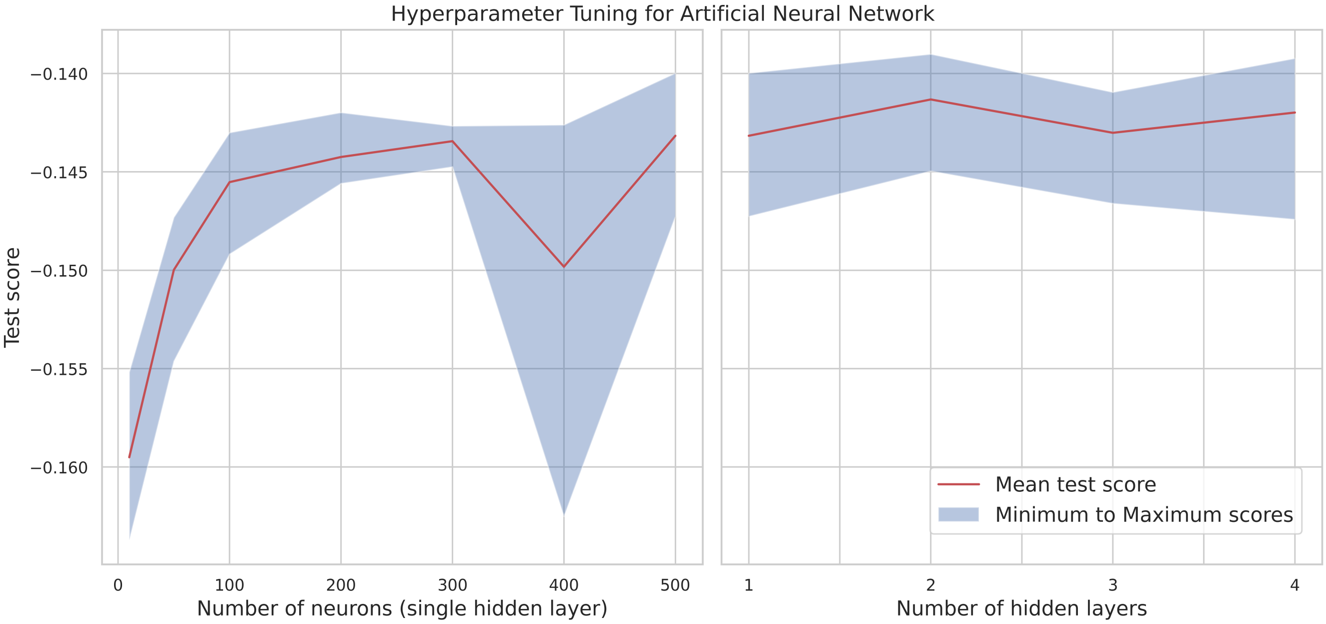

Figure 7. Hyperparameter tuning for the artificial neural network model varying the number of hidden layers and the number of neurons in the layer

https://doi.org/10.5194/egusphere-2022-1076
Preprint. Discussion started: 10 November 2022

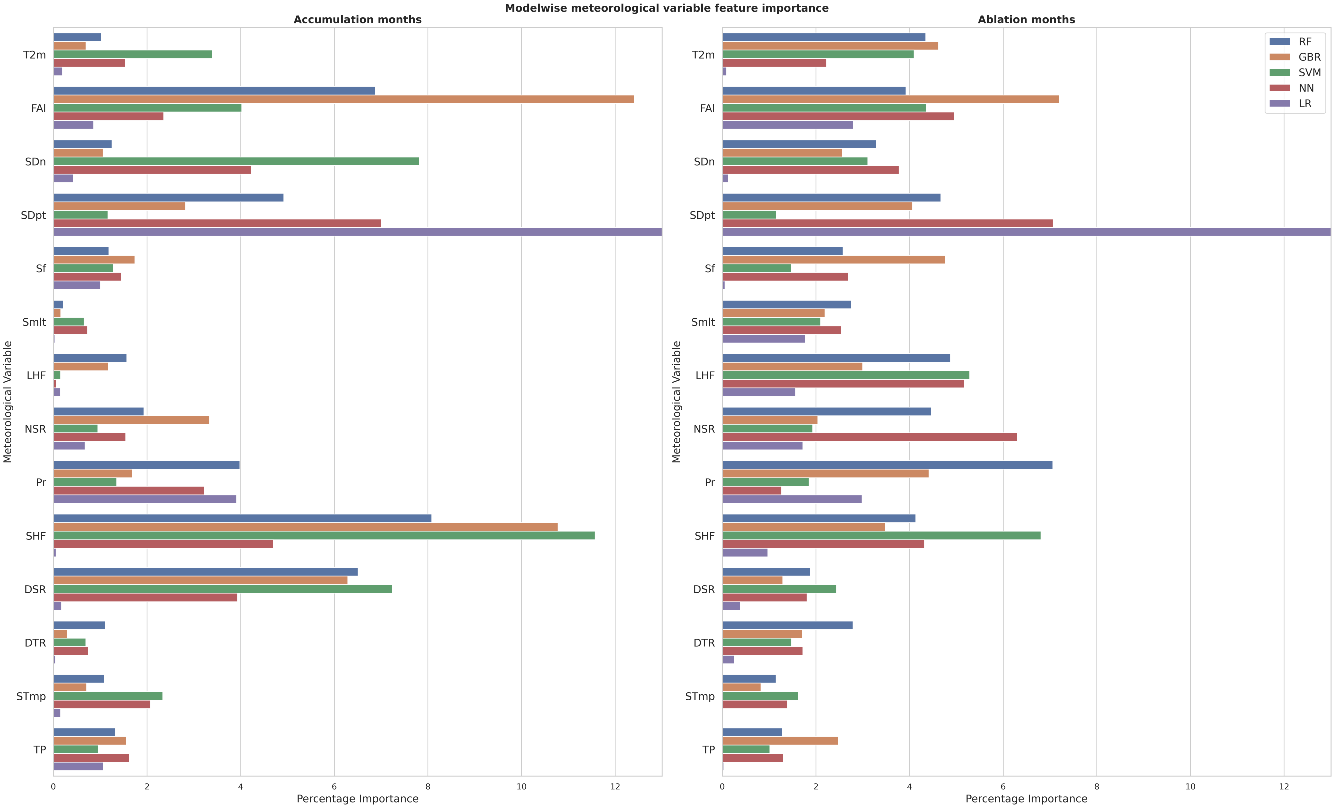

Figure 8. Percentage importance of all features summed over the accumulation and ablation season for the models: Random Forest (RF), Gradient Boosted Regression (GBR), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Linear Regression (LR). The figure has an x-axis limited to 13 for representation. The abbreviations used in the figure are expanded in Supplementary file S1.