Analysis of environmental, geographical and atmospheric factors on snow cover changes in the Himalayas

Mainack Paul, M.Stat Second Year, Roll No: MD2111
Supervisor: Dr. Sarbani Palit
(Final-Semester Report)

May 17, 2023

Abstract

High Mountain glaciers are an important source of water for the major river systems and provide water to hundreds of millions of people. Strong mass losses are experienced by glaciers and ice caps worldwide creating challenges with availability of water, hydro-power generation, and ecosystems. Understanding and analysing the snow cover in High Mountain Himalayan Ranges is important to anticipate and mitigate the geographical, atmospheric and environmental impacts of the melting glaciers. In this study, the response is **change in snow depth** and we aim to predict the melting of snow as a proxy of the change in snow depth over time. We take into consideration several topographical and meteorological factors and attempted to thoroughly investigate the important features that are responsible for this process of change in snow depth by working with a spatio-temporal data. Machine Learning and Deep Learning models captured a nonlinear response of glaciers to these factors, improving on the different linear statistical models already present. We have found out the best set of predictors for each model at different elevations and all of them are working with a very high accuracy and lastly, we have used them to make the predictions.

1 Introduction

1.1 General Description and Setting the Context

Glaciers are highly climate-sensitive component of the hydrological cycle of the earth. Rapid rise in temperature have impacted the glaciers enormously such as formation of new lakes, permafrost degradation, destabilising effects on mountain slopes, changes in hydrological regimes and challenges in water supply. Yet there exists a critical research gap related to quantifying the melting rate and causes behind it as declared by IPCC (Abram, Carolina, Bindoff, & Cheng, 2019). Melting and accumulation of glaciers vary from one region to others as well as over time; however, the glaciers are getting retreated and they are experiencing negative mass balance in the Himalayas (Bolch et

al., 2012). The cause for these spatio-temporal differences lies in the glacier as response to different climatic and topographic conditions.

1.2 Literature Review

Hock, R. et al.,(2003) worked on the vast majority of glacier evolution models by using a calibrated linear relationship between climatological factors and the melting of ice or snow. Steiner, D. et al.,(2005) and Clarke, G. K. C. et al.,(2009) worked on deep artificial neural networks(ANNs) that offer an alternative approach to these classic methods. However, the use of ANNs remains largely unexplored in glaciology for regression problems, with only a few studies using shallow ANNs for predicting the ice thickness or mass balance of a single glacier. Maussion, F. et al.,(2019) and Zekollari, H. et al.,(2019) have worked on natural processes inducing sudden increase in snow melting and simultaneous reduction in Snow Covered Areas(SCA) in high mountains. They are generally simulated with models requiring several parameterizations and simplifications to operate and efforts are put to improve the natural representation reflected by these models replacing empirical parameterizations with simplified physical models. C. Bouchayer et al.,(2022) developed several machine learning techniques namely Logistic regression, Random Forest and Boosting, to classify surge-type glaciers based on their location, exposure, geometry, climatic mass balance and runoff and compared the performances between the models and based on best model, they found the relative importance of several controlling features. Zhihua He et al.,(2021) has used Random forest algorithm for finding relations in between the response CRC(Contributions of Runoff Components) and the explanatory variables Mean Basin Elevation, Mean Annual Air Temperature, Mean Annual Precipitation, Winter Precipitation Fraction and Glacierized Area Ratio of the glacierized basin.

1.3 Objective of our Present Study

Rapid snow melting leads to glacier induced disaster and thus, our main objective is to model this extreme event beforehand for which we intent to use time series prediction. Since melting depends on various topographical and climatic variables, we are trying to find those factors and also considering the lags of the predictors that are important for modelling the melting. Using the important factors, we are going to fit various Machine Learning Deep Learning models that we will use for prediction. We will compare the models using various accuracy measures and will suggest the best possible model out of those.

Our present study have performed, to the best of our knowledge, the first-ever machine learning and deep learning models for snow depth changes over moderate length times series data taken on dates prior to Shishper Glacier Lake Outburst Flood (GLOF) disasters. With this study, we provide new predictions of glacier evolution in a mountain region where previous GLOF events are frequent over last five years, while investigating the role of nonlinearities in the response of glaciers to multiple topographical and climatic forcings. We have modelled by considering all of these factors together and tried to understand the statistical relationship among each other.

2 Data used for Model Preparation

2.1 Study Area

The study area around Shishper Glacier is situated within 36.2° - 36.6°N latitude and 74.3° - 74.9°E longitude. Recurrent GLOFs have happened from 2017 to 2022 around April to July months. However, the explicit linkage of snow cover dynamics with multiple topographical and climatic changes in the Shishper region is rarely investigated. Therefore, in the present study, we have attempted to understand the dynamics of seasonal snow cover and also the control of climate and topography on it. The relative significance of these factors is also assessed.

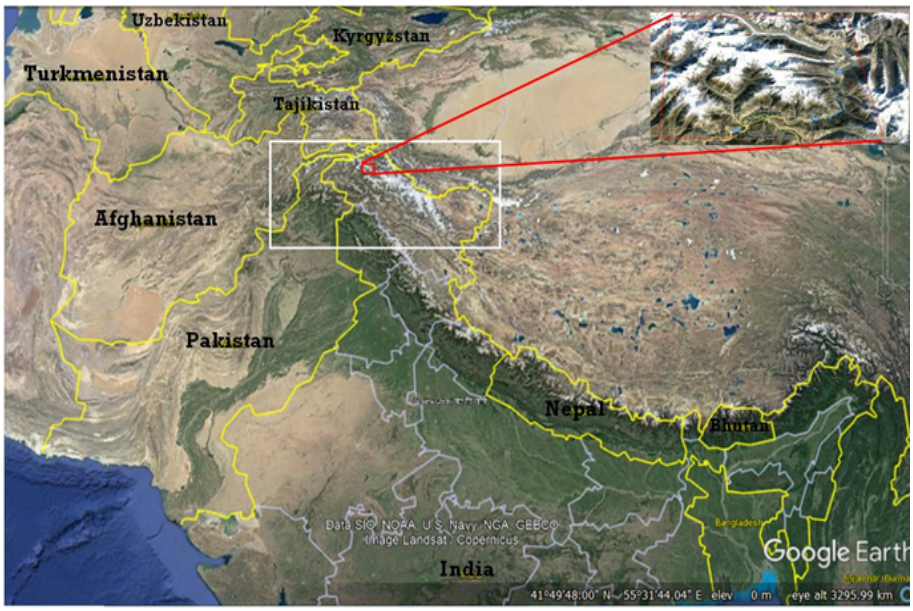


Figure 1: Shisper Region

2.2 Data Description

In this study, the influence of six topographical and five climatic variables over the study area are explored. The topographic variables (also the static variables) namely **Elevation, Slope, Aspect, Roughness, Terrain Ruggedness Index(TRI) and Topographic Position Index(TPI)** are derived from the digital elevation model data obtained by the Shuttle Radar Topography Mission at US Geological Survey website. It has a spatial resolution of 30m. The climatic variables (also the dynamic variables) viz. **Surface Radiative Temperature(SRT), Surface Air Temperature(SAT), Precipitation, Specific Humidity(SH) and Surface Pressure(SP)** have a spatial resolution of 0.01° and temporal resolution of 1 day, that has been obtained from the NASA Giovanni website (`https://giovanni.gsfc.nasa.gov/giovanni/`).

2.3 Data Management (Process of Collection and Extraction)

Data has been collected from the NASA Giovanni website as rasters in geotiff format. For each variable particuarly for each day, rasters has been downloaded and it is repeated for each variable. Then, using QGIS software, we have converted this rasters into usable data points. Then, we have filtered out the points that are necessary for our analysis.

3 Methods

3.1 Data Preprocessing

Any data needs to be verified for its consistency before starting the analysis. This verification of the data is known as scrutiny. We scrutinize the data to check if there are any missing values.

Next, data normalization for the data is conducted. Data normalization is the technique of transforming the features on a similar scale, which can improve the performance and training stability of the model. In this study, we have used **Min-Max scaler**. The transformation is given by

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where, x_{min},x_{max} and x_{new} referred to the minimum, maximum and new values of the x variable.

3.2 Training and Testing Data Sets

The training dataset is used for training the machine learning model. The model learns the relationship between the input features and the output variable based on the examples in the training dataset. The testing dataset is used for evaluating the performance of the trained machine learning model. The model is applied to the testing dataset, and its performance is evaluated by comparing its predictions with the actual values of the output variable. The testing dataset should be separate from the training dataset and should not be used in any way during the training process.

In this project, we have worked on a daily time series data of 45 days from 1 April - 15 May, 2022 for each of the 4 dynamic predictors and snow depth. For each timepoint, we have 2220 spatial data points which are chosen in such a way that spatial collinearity is negligible. For better understanding the data, we have divided the spatial points with respect to elevation(measured in m) since elevation is one of the most significant topographical variable. We have created 5 groups of elevation: 3000-4000m, 4000-5000m, 5000-6000m, 6000-7000m and finally 7000-8000m. Using this elevations, we have calculated lags for the variables which will be used for prediction.

For the prediction purpose, we have used those locations where high melting has been observed significantly and for that locations, we have used the daily time series data for the analysis. The first 37 data points, we have taken as training data and the last 8 points as the testing data.

3.3 Statistical Inter-Relationship versus Known Physical Relationships among variables (taken all together as well as elevation-wise)

For finding out the statistical inter-relationship among dynamic and static variables separately, we have used correlation and pair-wise scatterplots. For final conclusion. we have used Variance Inflation Factor, which is a measure of the degree of multicollinearity among the predictor variables in a multiple regression model. VIF is calculated for each predictor variable and indicates how much the variance of the estimated regression coefficient is increased due to multicollinearity. The formula for VIF is as follows:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where VIF_j is the j^{th} predictor variable and R_j^2 is the coefficient of determination obtained from a regression model that regresses the j^{th} predictor variable on all the other predictor variables in the model.

After finding the multicollinearity by taking the whole data at once (for both static and dynamic variables), then, we have divided the data elevation-wise and checked the same.

High humidity melts snow and ice much faster because the humid air squeezes more air molecules out onto the snow surface, where it cools and condenses. This phase change releases heat, which further melts the snow. In dry air, there is very little condensation, so the snow melts much more slowly. Again, high temperature is directly proportional to the melting. Thus, we expect a positive correlation in between humidity and melting and also in between temperature and melting.

3.4 Cross-correlation

Cross-correlation measures the degree of linear dependence between two time series at different time lags. The cross-correlation function is a mathematical function that quantifies the similarity or dissimilarity between two time series as a function of the lag between them. Let's consider two time series, x_t and y_t, with t denoting time. The cross-correlation function between x_t and y_t at a lag τ is given by the following formula:

$$R_{xy}(\tau) = \frac{\sum_{t=1}^{n-\tau} (x_t - \mu_x)(y_{t+\tau} - \mu_y)}{(n - \tau)\sigma_x \sigma_y}$$

where μ_x and μ_y are the means, σ_x and σ_y are the standard deviations of the x and y time series respectively, n is the length of the time series.

After finding the cross-correlation by taking the whole data at once (only dynamic variables), then, we have divided the data elevation-wise and checked the same.

3.5 Machine Learning Models

Various machine learning models were constructed for the purpose of feature selection and prediction in modelling our response variable which is **change in snow depth** namely **Multiple Linear Regression, Regression Tree, Random Forest, Generalized Boosting and Artificial Neural Network.**

3.5.1 Multiple Linear Regression

A linear model is used to describe the relationship between the response variable and the explanatory variables using a linear function. The multiple linear regression model is a linear regression model with multiple explanatory variables. The equation for multiple linear regression is:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i;\ i = 1, ..., n$$

where β_0 is the intercept, β_j is the partial regression coefficients of y on the j^{th} explanatory variable $j = 1, ..., p$ for fixed values of the remaining explanatory variables; and ϵ_i is the error term, p is the total number of explanatory variables and n is the data size.
Estimation of a multiple linear regression model is done **with the least squares criterion**. With this criterion, the β_i coefficients are chosen that minimize the sum of squared vertical distances between observed y_i and the fitted values of the response, by:

$$\hat{\beta} = min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}))^2$$

where $\hat{\beta}$ are the regression coefficients that minimize the sum of the squares. We have reported the $\hat{\beta}'s$ for both the static and dynamic variables in the following tables.

There are some assumptions that need to be satisfied while fitting a linear model.

- **Checking for the Normality of the Residuals**

We are mainly interested in testing the hypotheses and constructing confidence intervals about the model parameters. These procedures require that we make the additional assumption that the model errors ϵ_i are normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean 0 and variance σ^2, that is, homoscedastic. We will check for homoscedasticity later in the following section.

Normal Probability Plot

The normal probability plot is a plot of the ordered standardized residuals versus the so-called normal scores. The normal scores are the cumulative probability

$$P_i = \frac{i}{n}; i = 1, ..., n$$

If the residuals e_1, e_2, ..., e_n are ordered and ranked in increasing order as $e_{(1)} < e_{(n)} < \ldots < e_{(n)}$ and the $e'_{(i)}s$ are plotted against P_i, the plot is called Normal Probability Plot. If the residuals are normally distributed, then the ordered residuals should be approximately the same as the ordered normal scores.

Histogram and Density Plot

In this method, we plot a histogram of residuals and normal density on the same plot to observe if there is any significant difference between the distribution of the errors and theoretical standard normal distribution.

Testing for Normality using Shapiro-Wilk Test

Hypothesis: To test H_0: Residuals follow Normal distribution against H_1: Not H_0.

Test Statistic: The test statistic under H_0 is given by,

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

where $e_{(i)}$ pertains to the ith largest value of the error terms and a_i denotes those values which are calculated using the means, variances and covariances of the e_i. W comes out very small if the sample is not from the normal distribution. Else, it will be large. The test will reject the null hypothesis if the p-value is less than or equal to 0.05. If the normality test fails, it allows us to state with 95% confidence that the residuals are not normally distributed and hence the sample did not come from a Normal distribution. If the normality test is Passed, it will only allow us to state that no significant departure from normality was found.

- **Checking for heteroscedasticity of the residuals**

Originally, heteroscedasticity means "unequal scatter". We usually use the term heteroscedasticity in the context of the residuals or error term.

Testing for Homoscedasticity using Breusch-Pagan Test

To test the hypothesis: H_0: Homoscedasticity is present vs. the alternative H_1: Heteroscedasticity is present

The procedure of the test is as follows: First we have to fit the OLS regression model and calculate the squared residuals of the model. Then, we have to fit a new regression model, using the squared residuals as response variable and calculate the Chi-Square test statistic, as $n * R^2_{new}$, where, n is the total number of observations and R^2_{new} is the R^2 for the new regression model that used the squared residuals as the response values.

3.5.2 Regression Tree

It is a flowchart like formation, where internal nodes is used to represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. It is a supervised learning method.

Working

The algorithm first divide the predictor space, that is, the set of possible values for $X_1, X_2, ..., X_p$ into J distinct and non-overlapping regions, $R_1, R_2, ..., R_J$. The goal is to find boxes $R_1, R_2, ..., R_J$ that will minimize the Residual Sum of Squares(RSS), given by

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{R_j})^2$$

where $\hat{y}_{R_j}$is the mean response for the training observations within the j^{th} box. Unfortunately, it will be computationally infeasible to consider every possible partition of the feature space into J boxes. For this reason, we will take a top-down, greedy approach. It is primarily known as top-down because it begins at the top of the tree (where all observations will belong to a single region) and then successively, it splits the predictor space and each split will be indicated via two new branches further down on the tree. It is known as greedy because at each step of the tree-building process, the best split will be made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step. For performing this, we will first select the predictor X_j and the cut-point s such that splitting the predictor space into the regions $\{X|X_\mathrm{j} < s\}$ and $\{X|X_\mathrm{j} \geq s\}$ leads to the greatest possible reduction in RSS. (The notation $\{X|X_\mathrm{j} < s\}$ means the region of predictor space in which X_j takes on a value less than s.) That is, we consider all predictors $X_1, X_2, ..., X_p$, and all possible values of the cut-point s for each of the predictors, and then choose the predictor and cut-point such that the resulting tree has the lowest RSS. In greater detail, for any j and s, we define the pair of half-planes

$$R_1(j; s) = \{X|X_\mathrm{j} < s\} \quad and \quad R_2(j; s) = \{X|X_\mathrm{j} \geq s\}$$

and we seek the value of j and s that minimize the equation

$$\sum_{i:x_i \epsilon R_1(j;s)} (y_i - \hat{y}_{R_1})^2 \quad + \sum_{i:x_i \epsilon R_2(j;s)} (y_i - \hat{y}_{R_2})^2$$

where $\hat{y}_{R_1}$ is the mean response for the training observations in $R_1(j; s)$, and $\hat{y}_{R_2}$ is the mean response for the training observations in $R_2(j; s)$.

We go on repeating this process and look for the best predictor and best cut-point in order to split the data further so as to minimize the RSS within each of the resulting regions. However, this time, instead of splitting the entire predictor space, we split one of the two previously identified regions and thus We now have three regions. For further minimising the RSS, we look to split

one of these three regions. The process continues until a stopping criterion is reached that we set according to our requirement.

3.5.3 Random Forest Regression

The Random Forest regression algorithm is an ensemble method that collaborates the performance of a large number of decision trees for predict the value of the responses. The decision trees are produced by a process known as **bagging**(the method of drawing a subset of training samples with replacement). It predicts the response by averaging the estimations of every decision tree, that results in estimations with the low-bias property of the decision trees. Furthermore, the variance of the predictions will be low due to the well known Central Limit Theorem:

$$Standard\,Error\ =\ \sigma_{\bar{x}} =\ \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is the standard deviation of the sample, σ the standard deviation of the population, and n is the number of data points.

3.5.4 Generalized Boosting Model

These models are a combination of two techniques: decision tree algorithms and boosting methods. Generalized Boosting Models will recurrently fit many decision trees to upgrade the overall accuracy of the model. A random subset of all the data is selected For every new tree in the model. The input data is weighted in such an algorithm that the data, that was poorly modelled by previous trees has a higher probability of being selected in the new tree which means that, after the first tree is fitted, the model will take into account the error in the prediction of that tree to fit the next tree, and so on. After taking into account the fit of previous trees that were built, the model continuously will try to improve its accuracy. This sequential approach is unique to boosting. Generalized Boosting Models have two important parameters that need to be specified by the user:

- **Interaction depth:** It controls the number of splits in each tree. A value of 1 results in trees with only 1 split, and means that the model does not take into account interactions between environmental variables. A value of 2 results in two splits, etc.
- **Shrinkage:** It determines the contribution of each tree to the growing model. As small shrinkage value results in many trees to be built.

3.5.5 Artificial Neural Network

Artificial Neural Networks (ANNs) are a type of machine learning algorithm inspired by the structure and function of the human brain. The theory behind ANNs is based on the idea that information processing in the brain is carried out by networks of interconnected neurons that communicate with each other.The basic building block of an ANN is the artificial neuron, which receives input from one or more sources, applies a mathematical function to the input, and produces an output

that is passed to other neurons in the network. The output of a neuron is determined by the sum of the weighted inputs plus a bias term, which is passed through an activation function. The activation function determines the output of the neuron based on the input it receives, and is usually a non-linear function such as the sigmoid function. ANNs are typically organized into layers, with each layer consisting of a number of neurons that perform a specific computation. The first layer of the network is the input layer, which receives the raw input data. The final layer of the network is the output layer, which produces the final output of the network. In between the input and output layers, there can be one or more hidden layers that perform intermediate computations. During training, ANNs learn to perform a specific task by adjusting the weights and biases of the neurons in the network to minimize a loss function that measures the difference between the predicted output and the true output. The weights and biases are adjusted using an optimization algorithm such as stochastic gradient descent, which takes small steps in the direction of the negative gradient of the loss function with respect to the weights and biases.

Let us define a neural network by taking an input vector of p variables $X = (X_1, X_2, ..., X_p)$ and build a nonlinear function $f(X)$ to predict the response Y. Each of the inputs from the input layer feeds into each of the K hidden units. The neural network model has the form

$$\begin{aligned} f(X) =& \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X) \\ =& \beta_0 + \sum_{k=1}^{K} \beta_k g(w_{k0} + \sum_{j=1}^{P} w_{kj} X_j) \end{aligned}$$

where $g(z)$ is a nonlinear activation function such that

$$g(z) = \frac{e^z}{1 + e^z}.$$

All the parameters $\beta_0, ..., \beta_K$ and $w_{10}, ..., w_{Kp}$ need to be estimated from data. For a quantitative response, typically squared-error loss is used, so that the parameters are chosen to minimize

$$\sum_{i=1}^{n} (y_i - f(x_i))^2.$$

3.5.6 Reason to try out these selected ML Models

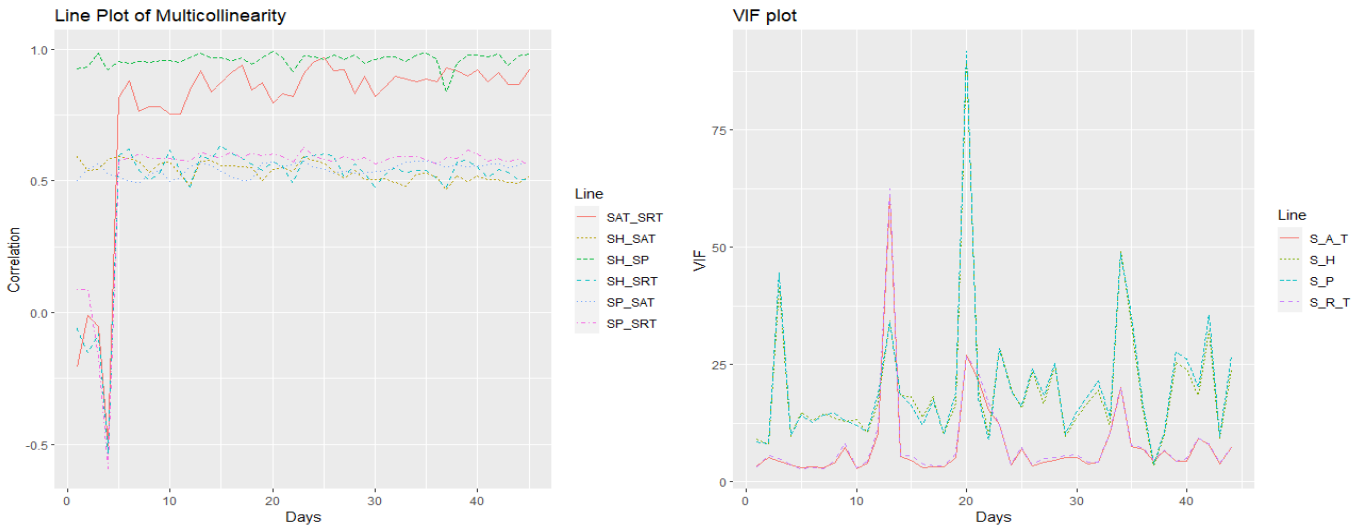
Non-linear machine learning models are often better than linear regression models because they can help us capture more complex relationships between the predictor variables and the response variable. Basic assumption of the Linear regression model is that the relationship between the predictor variables and the response variable is linear, which means that the effect of a change in one predictor variable on the response variable will be constant, regardless of the value of the other predictor variables and hence, we have done the Regression Tree. But, then, regression tree is non-robust which means that a small change in the data can cause a large change in the final

estimated tree, that is, it can cause overfitting and hence we have done Random Forest Regression. Boosting trees can be sometimes more accurate than random forests as we train them to correct each other's errors and hence they're capable of capturing complex patterns in the data. ANN model is used since it increases the accuracy for prediction.

4 Results

4.1 Statistical Inter-relationships

First, we are going to check the inter-relationship in between dynamic variables by taking the whole data at once.



(a) Multicollinearity Plot

(b) VIF Plot

Figure 2: Dynamic Variables

In the first figure, we have taken the data as a whole and plotted Multicollinearity for each pair of dynamic variables over time and found out that, **Surface Pressure-Specific Humidity** pair has the highest correlation. In the second figure, we have plotted the Variance Inflation Factor(VIF) for each of the variables over time and found out that **Surface Pressure** has the highest VIF. **We have also checked multicolliearity at different levels of elevation by partially taking the dataset and all are having similar results. Hence, we remove Surface Pressure from our analysis.**

Next, we are going to check the inter-relationship in between static variables by taking the whole data at once.

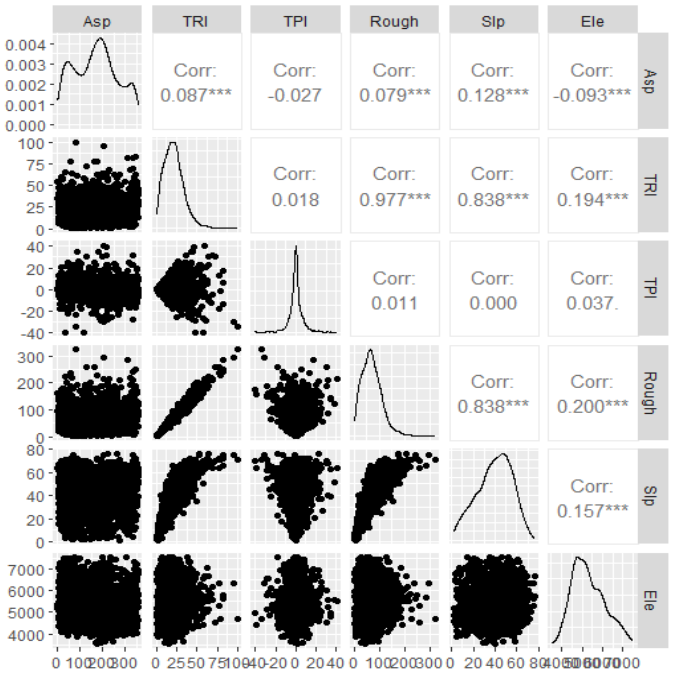


Figure 3: Scatterplot of Static Variables

Interpretation: From the above plot, we can see that the variable TRI, Roughness and Slope has very high multicollinearity among themselves.

A checking of the VIF reveals that the value of VIF for both TRI and Roughness is pretty high. But once we remove TRI, all other variables are having very less VIF. **Hence, it indicates that TRI was causing the multicollinearity problem. We have also checked multicolliearity at different levels of elevation by partially taking the dataset and all are having similar results. Hence, we remove TRI from our analysis.**

4.2 Snow Melting during Winter: A discussion

We have done our data analysis intensively for both summer and winter season but we will report mainly the analysis of the summer season. The reason is that, in winter, we found out that there is no significant melting in a small interval of time, that can lead to disaster. Also, not much events regarding GLOF has been observed in this region during winter.

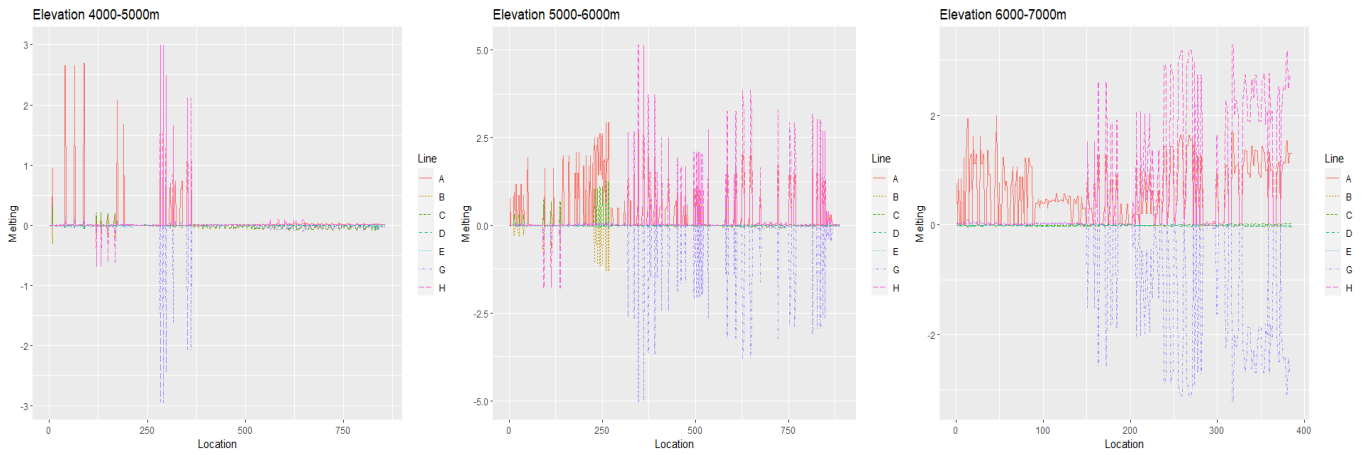


Figure 4: Graphs depicting Elevation-Wise Snow Melting during Winter

In the above graphs, we have plotted **change in snow depth** for 7 different days denoted by A, B, C, D, E, G and H for different elevations.

Interpretation: From the graphs, we can visually conclude that no significant melting has occurred in a small interval of time.

4.3 Cross-correlation

We have checked lags up to order 7 for different elevations for each of the dynamic variables.

Elevation 3000-4000m, 4000-5000m, 5000-6000m, 6000-7000m and 7000-8000m has 30, 856, 886, 385 and 59 spatial points respectively.

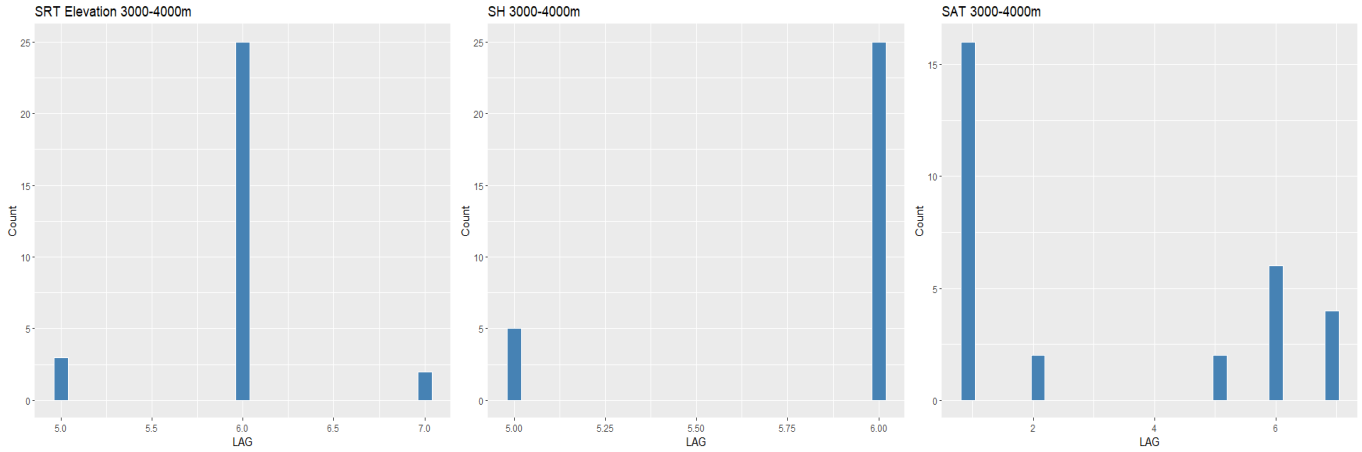


Figure 5: Lags of Dynamic Variables at Elevation 3000-4000m

Table 1: Lags of Dynamic Variables at Elevation 3000-4000m

Lags	1	2	3	4	5	6	7
SRT	0	0	0	0	3	25	2
SH	0	0	0	0	5	25	0
SAT	16	2	0	0	2	6	4

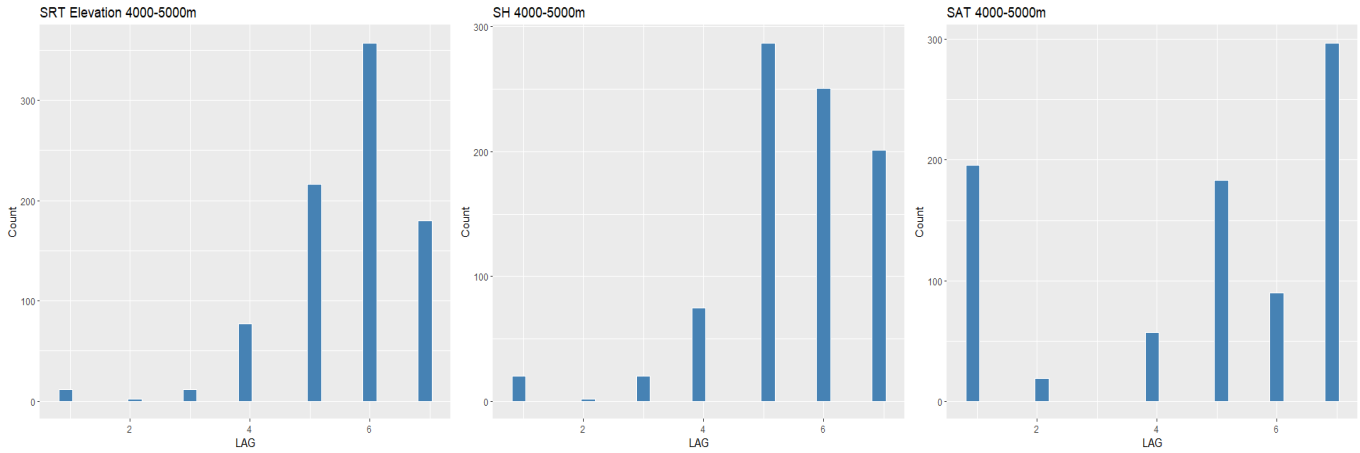


Figure 6: Lags of Dynamic Variables at Elevation 4000-5000m

Table 2: Lags of Dynamic Variables at Elevation 4000-5000m

Lags	1	2	3	4	5	6	7
SRT	12	2	12	77	216	357	180
SH	20	2	20	75	287	251	201
SAT	196	19	0	57	183	90	297

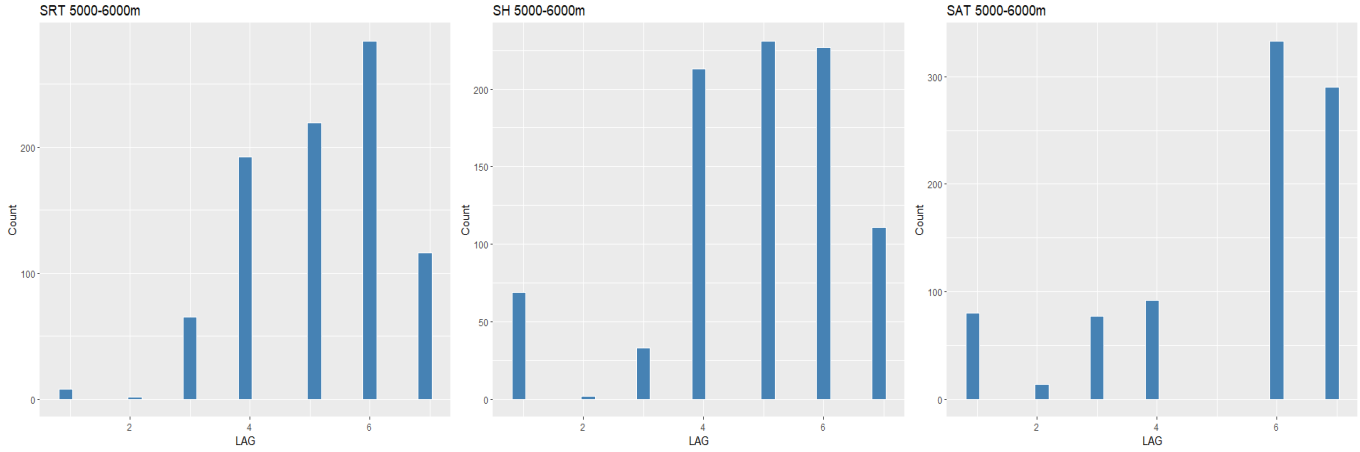


Figure 7: Lags of Dynamic Variables at Elevation 5000-6000m

Table 3: Lags of Dynamic Variables at Elevation 5000-6000m

Lags	1	2	3	4	5	6	7
SRT	8	2	65	192	219	284	116
SH	69	0	33	207	227	221	99
SAT	80	14	77	92	0	333	290

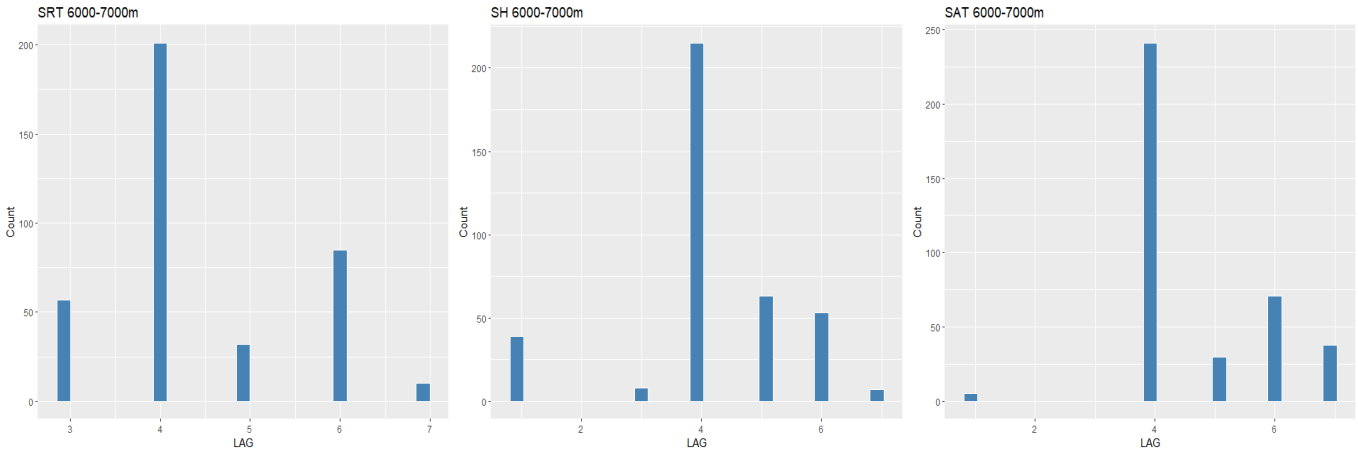


Figure 8: Lags of Dynamic Variables at Elevation 6000-7000m

Table 4: Lags of Dynamic Variables at Elevation 6000-7000m

Lags	1	2	3	4	5	6	7
SRT	0	0	57	201	32	85	10
SH	39	0	8	215	63	53	7
SAT	5	0	0	241	30	71	38

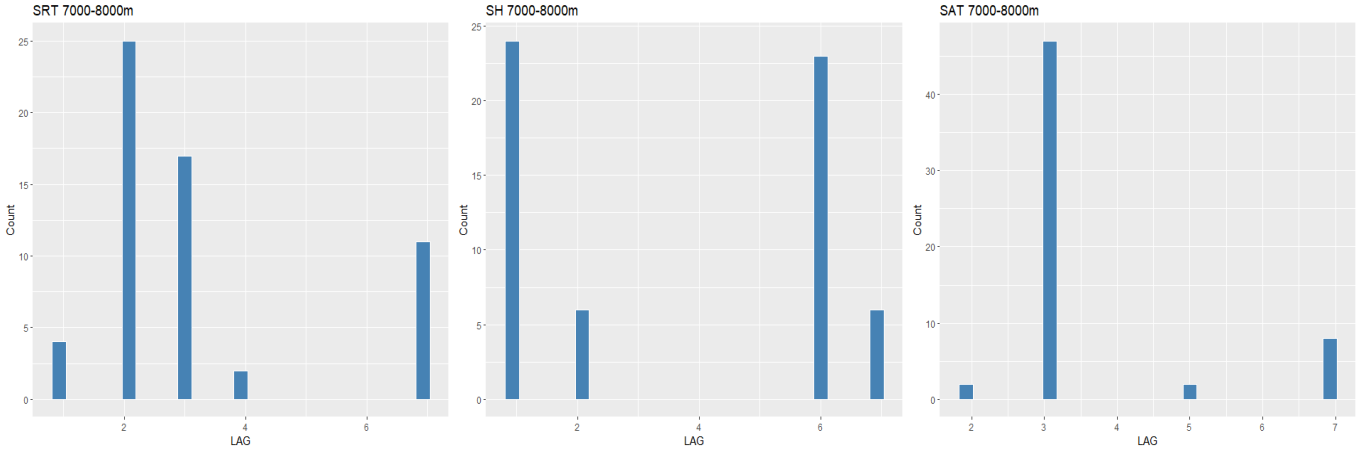


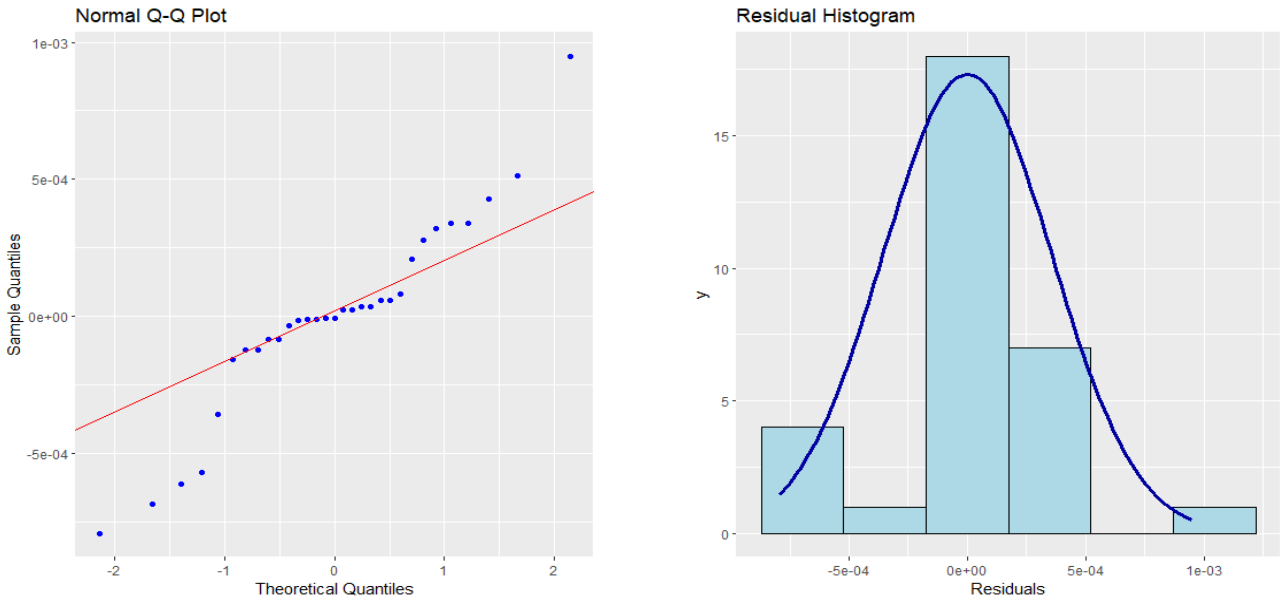
Figure 9: Lags of Dynamic Variables at Elevation 7000-8000m

Table 5: Lags of Dynamic Variables at Elevation 7000-8000m

Lags	1	2	3	4	5	6	7
SRT	4	25	17	2	0	0	11
SH	24	6	0	0	0	23	6
SAT	0	2	47	0	2	0	8

4.4 Discarding Linear Model

4.4.1 Checking for Normality



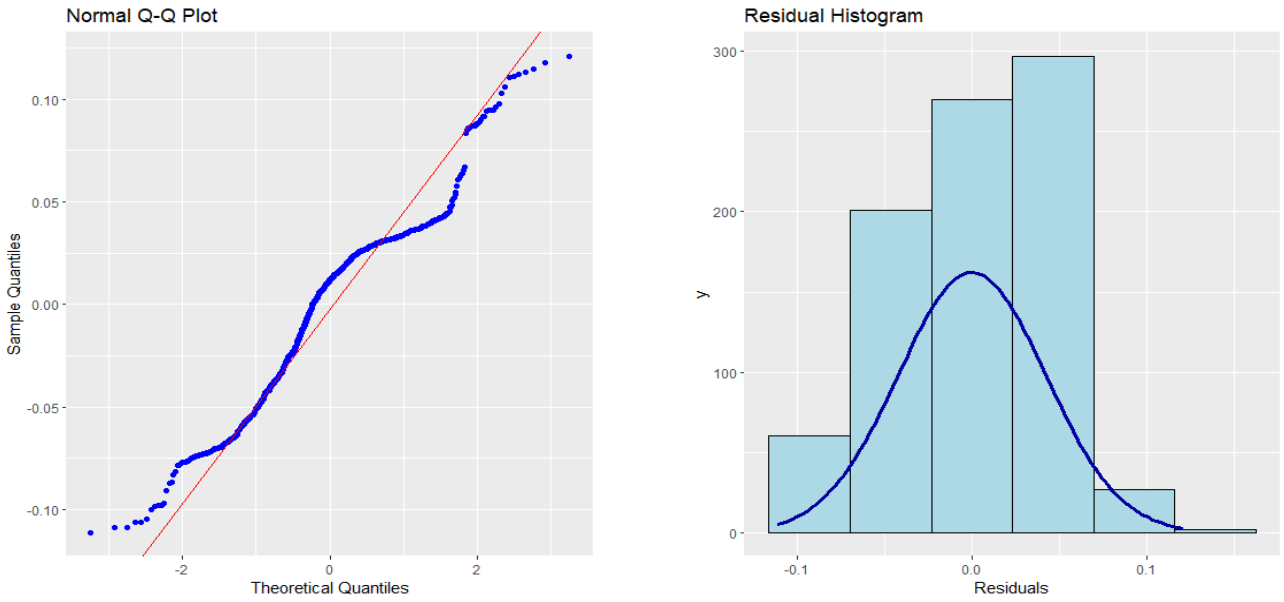
(a) QQ Plot

(b) Histogram Plot

Figure 10: Dynamic Variables

From the above plots, we can visually conclude that the normality assumption does not hold.

Also, the p-value of Shapiro-Wilk test came out to be 0.04 and hence, we can say that the residuals are not distributed normally in the model.

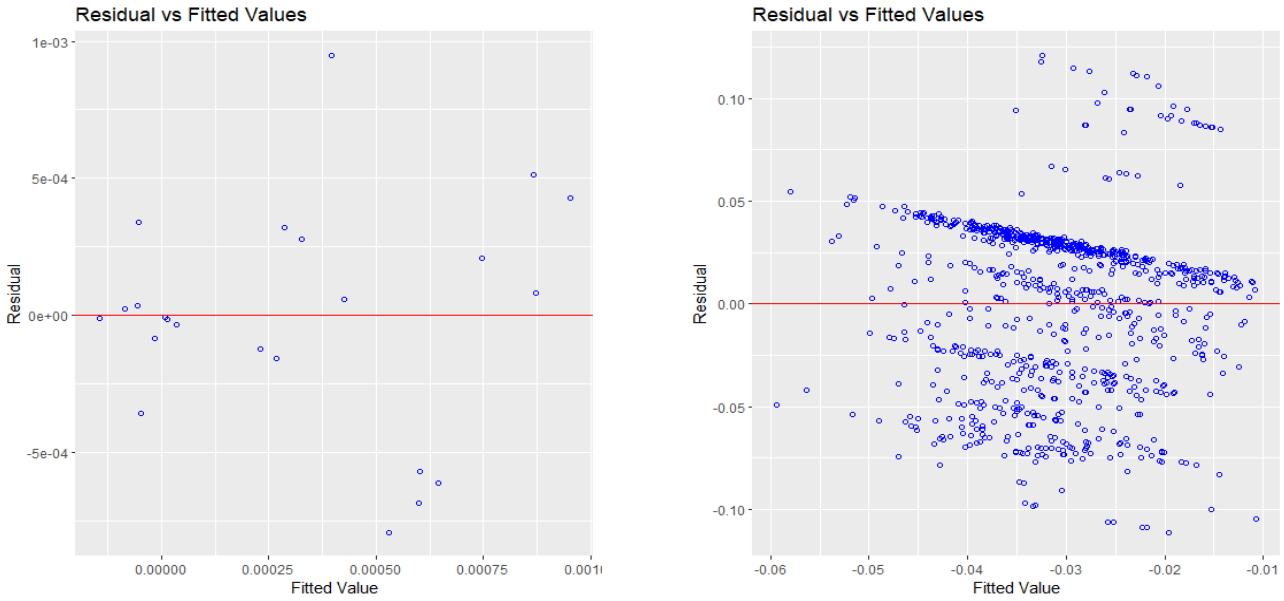


(a) QQ Plot (b) Histogram Plot

Figure 11: Static Variables

From the above plots, we can visually conclude that the normality assumption does not hold. Also, the p-value of Shapiro-Wilk test came out to be $6.25 * 10^{-16}$ and hence, we can say that the residuals are not distributed normally in the model.

4.4.2 Checking for Heteroscedasticity



(a) Dynamic Variables (b) Static Variables

Figure 12: Residual Plot

For dynamic variables, heteroscedasticity seems to be present as the plot is more or less not random and the p-value of BP test came out to be 0.00767; and hence, we can say that the residuals are not homoscedastic. For static variables, heteroscedasticity seems to be present as the plot is not random and the p-value of BP test came out to be 0.0001794; and hence, we can say that the residuals are not homoscedastic.

Hence, we will not use Linear Models for modelling our response variable.

4.5 Variable Selection using Machine Learning (ML) Models

For the static variables, we have done this feature selection at a single time point only but with varying locations.

Table 6: Static Variables

Elevation	**ML model**	**Significant Variables**
	Regression Tree	Roughness,Elevation,Aspect
3000-4000m	Random Forest	Aspect,Elevation,TPI
	Generalised Boosting	Roughness,Slope,Elevation,Aspect
	Regression Tree	Elevation,Aspect,TPI,Roughness
4000-5000m	Random Forest	Slope,Elevation,Aspect,TPI
	Generalised Boosting	Roughness,Slope,Elevation,Aspect
	Regression Tree	Roughness,Slope,Elevation,Aspect
5000-6000m	Random Forest	Roughness,Slope,Elevation,Aspect
	Generalised Boosting	Roughness,Slope,Elevation,Aspect
	Regression Tree	Roughness,Slope,Elevation,Aspect
6000-7000m	Random Forest	Slope,Elevation,Aspect
	Generalised Boosting	Roughness,Slope,Elevation,Aspect
	Regression Tree	Roughness,Elevation,Aspect
7000-8000m	Random Forest	Roughness,TPI,Elevation,Aspect
	Generalised Boosting	Roughness,TPI,Elevation,Aspect

From the table, we can interpret that Roughness, Slope, Elevation and Aspect are the most significant static variables.

Notation

SRT_i denote Surface Radiative Temperature at lag i and similarly, we have named the other dynamic variables with their corresponding lags.

Table 7: Dynamic Variables for April 10 and April 20

		April 10	April 20
Elevation	**ML model**	**Significant Variables**	**Significant Variables**
	Regression Tree	SRT, SAT, SRT_6	SRT_6, SRT
3000-4000m	Random Forest	SAT_1, SAT, SRT	SH, SRT_6, SAT_1, SRT
	Generalised Boosting	SH, SAT, SH_6	SRT, SRT_6, SH, SAT_1
	Regression Tree	SRT, SH_6, SRT_5, SRT_6	SRT, SRT_6, SRT_5
4000-5000m	Random Forest	SRT, SAT, SRT_5, SRT_6	SRT, SRT_6, SRT_5
	Generalised Boosting	SAT, SRT_6, SAT, SRT	SRT, SRT_6, SRT_5
	Regression Tree	SAT_6, SH_6, SAT_7, SRT	SAT_6, SRT_6, SAT, SRT
5000-6000m	Random Forest	SRT, SRT_6, SH_6, SAT_7	SAT_6, SRT_6, SAT_7, SRT
	Generalised Boosting	SRT_6, SH_6, SAT_7, SRT	SRT_6, SH_6, SAT_7, SRT
	Regression Tree	SRT_4, SRT, SH	SH_4, SAT, SH
6000-7000m	Random Forest	SRT_4, SAT, SAT_4	SRT, SH_4, SAT, SRT_4
	Generalised Boosting	SAT_4, SH_4, SH, SRT_4	SAT_4, SH_4, SH, SRT_4
	Regression Tree	SRT_2, SRT	SRT, SRT_2, SAT
7000-8000m	Random Forest	SRT, SRT_2	SRT, SRT_2, SAT
	Generalised Boosting	SRT_2, SAT, SRT, SH	SRT_2, SRT, SAT

Table 8: Dynamic Variables for April 30 and May 10

		April 30	May 10
Elevation	**ML model**	**Significant Variables**	**Significant Variables**
	Regression Tree	SRT, SAT_1	SRT_6, SAT, SH
3000-4000m	Random Forest	SAT, SRT, SRT_6	SAT, SRT, SH
	Generalised Boosting	SH, SH_6, SRT, SAT	SRT, SAT, SH, SRT_6
	Regression Tree	SH, SAT, SRT_5	SRT, SAT, SRT_6, SAT_7
4000-5000m	Random Forest	SRT_6, SAT, SAT_7, SRT_5	SRT_6, SAT_7, SRT, SRT_5
	Generalised Boosting	SH, SRT, SAT, SAT_7	SAT_7, SRT_6, SRT
	Regression Tree	SH, SRT_6, SAT, SH_6	SH_5, SRT, SAT_7
5000-6000m	Random Forest	SRT_6, SAT	SRT, SAT_7, SAT, SRT_5
	Generalised Boosting	SAT, SRT_6, SH	SAT_7, SRT, SRT_6, SH_5
	Regression Tree	SAT_4, SRT_4, SH_4	SAT, SRT, SRT_4
6000-7000m	Random Forest	SAT_4, SAT, SRT_4	SRT, SRT_4
	Generalised Boosting	SRT_4, SAT, SRT	SRT, SH, SH_4
	Regression Tree	SRT, SRT_2, SAT	SRT, SRT_2
7000-8000m	Random Forest	SRT, SRT_2, SRT_3	SRT, SRT_2
	Generalised Boosting	SH, SRT, SRT_2, SAT	SRT, SH, SAT

Here, we have chosen 4 days namely April 10, 20, 30 and March 10 for our analysis. We have divided the data elevation-wise since elevation is one of the most significant static variable. So, for each elevation, we have obtained different set of significant predictors and their significant lagged

versions. Now, using the overall significant predictors obtained from these 4 days, we will try to predict our response elevation-wise.

One thing to note that, we will not use Regression tree for our prediction since it is non-robust which means that a small change in the data can cause a large change in the final estimated tree, that is, it can cause overfitting. Hence, in place of that, we are going to use a more reputed model known as the Artificial Neural Network.

4.6 Evaluation of the accuracy of Predictions of Machine Learning Models

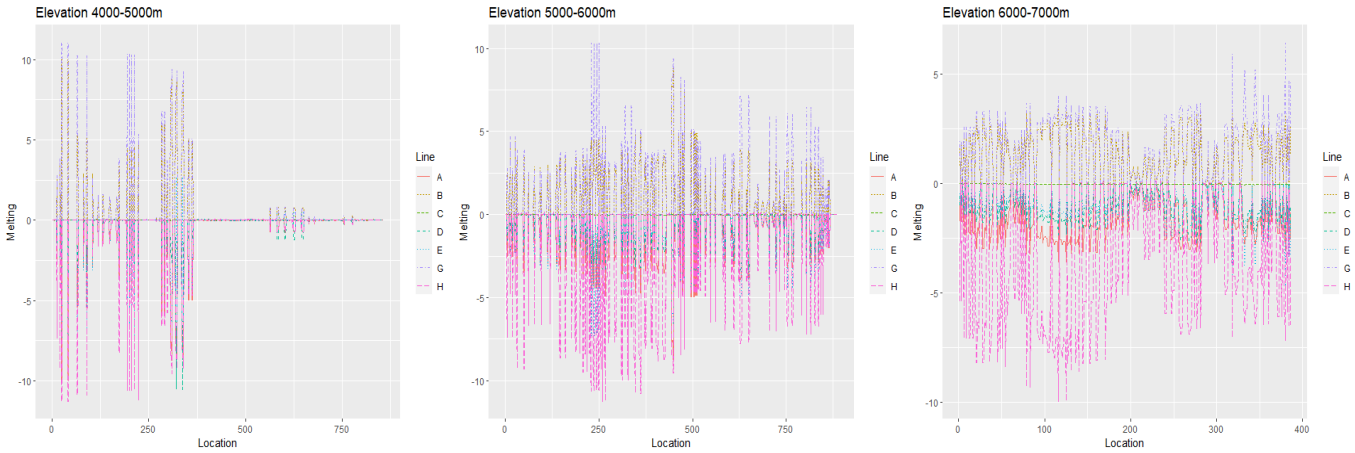


Figure 13: Graphs depicting Elevation-Wise Snow Melting during Summer

In the above graphs, we have plotted **change in snow depth** for 7 different days denoted by A, B, C, D, E, G and H for different elevations.

Table 9: Accuracy Measures of ML Models done for 3 independent locations

		Location 1		Location 2		Location 3	
Elevation	**ML model**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
	Artificial Neural Network	7.7638	8.6808	7.7639	8.6808	6.1872	7.0781
4000-5000m	Random Forest	7.9126	8.8070	7.7305	8.5914	6.1302	7.0969
	Generalised Boosting	6.8807	7.9883	6.8741	7.9956	5.4666	6.6290
	Artificial Neural Network	5.9731	6.9198	4.2587	5.2378	5.8267	6.5331
5000-6000m	Random Forest	5.1393	6.4090	4.0629	5.2919	5.7216	6.4897
	Generalised Boosting	4.9307	6.1316	3.5305	4.9382	5.3707	6.1238
	Artificial Neural Network	3.4169	3.9000	3.9973	4.5794	3.0487	3.5308
6000-7000m	Random Forest	2.7580	3.4055	3.2859	3.9894	2.4278	3.0601
	Generalised Boosting	2.8942	3.7714	3.0410	3.7448	2.4921	3.3369

From the table, we can interpret that all of the models are performing very well because the test errors of the accuracy measures is very small.

5 Discussion

5.1 Important features: Static versus Dynamic Variables

Out of the six static variables Elevation, Slope, Aspect, Roughness, TRI and TPI, TPI is the least significant one and we have omitted TRI from our analysis due to it's multicollinearity constraint. Similarly, out of the five dynamic variables Surface Radiative Temperature, Surface Air Temperature, Precipitation, Specific Humidity and Surface Pressure, Precipitation is the least significant one and hence, we have removed precipitation from our analysis. Along with that, we have also omitted Surface Pressure due to the multicollinearity constraint.

5.2 Elevation dependence of lags

We have used this lag table for model building of dynamic variables.

Table 10: Elevation-wise usage of Lag Values in ML models

Variables	**3000-4000m**	**4000-5000m**	**5000-6000m**	**6000-7000m**	**7000-8000m**
SRT	**6**	**5,6**	**5,6**	**4**	**2**
SH	**6**	**5,6**	**5,6**	**4**	**1**
SAT	**1**	**7**	**6,7**	**4**	**3**

5.3 ML Model Performance

Table 11: Most Significant Variables in Prediction

Elevation	**ML model**	**Overall Significant Variables**
	Regression Tree	SRT, SRT_6, SAT
3000-4000m	Random Forest	SAT, SRT, SAT_1
	Generalised Boosting	SH, SRT, SAT
	Regression Tree	SAT, SRT, SRT_5, SRT_6
4000-5000m	Random Forest	SAT, SRT, SRT_5, SRT_6
	Generalised Boosting	SAT, SRT, SRT_5, SRT_6
	Regression Tree	SRT, SRT_6, SAT, SH_6
5000-6000m	Random Forest	SRT,SRT_6,SAT_7
	Generalised Boosting	SRT, SRT_6, SAT_7
	Regression Tree	SH, SH_4, SRT_4
6000-7000m	Random Forest	SAT, SRT_4, SAT_4
	Generalised Boosting	SRT_4, SH, SH_4
	Regression Tree	SRT, SRT_2, SAT
7000-8000m	Random Forest	SRT, SRT_2
	Generalised Boosting	SH, SRT, SRT_2, SAT

Note

We have also checked by taking both the static and dynamic variables together in a model for feature selection and found out that only the dynamic variables are coming out significant among them which is quite natural for us to expect.

6 Conclusions and Scope of Study

Among the static variables, Elevation, Slope, Roughness and Aspect came out to be the most significant. One interesting thing we have observed is that in case of dynamic variables, if we divide the data elevation-wise, we get different set of significant predictors for our response, and it has varied in between models also. We have performed the prediction when actually GLOF has occurred in Shisper region and the models' accuracy rate has come out to be very high and thus, we can confidently say that our models are performing very good in predicting the required response. The most important advantage is that they can be used for prediction in other locations too, if the elevations and the predictors are kept constant.

For further scope of study, one can surely try out the traditional multivariate time series models for predicting the change in snow depth and can compare those models with the ML models that we have suggested. One can try out working with other significant dynamic and static variables other than the ones we have used in this project.

7 Acknowledgement

I would like to thank my supervisor Dr. Sarbani Palit for giving me the chance to do this project and for her constant support throughout the project. I would also like to thank Ayoti Banerjee (SRF GSU ISI, Kolkata) for helping me during collection of data and also helping me understand how QGIS worked without which this project was not possible .

8 Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

9 References

- Hock, R.,(2003). Temperature index melt modelling in mountain areas. J. Hydrol. 282, 104–115.
- Steiner, D., Walter, A. & ZumbÃŒhl, H. J.,(2005). The application of a non-linear back propagation neural network to study the mass balance of Grosse Aletschgletscher, Switzerland. J. Glaciol. 51, 313–323.

- Clarke, G. K. C., Berthier, E., Schoof, C. G. & Jarosch, A. H.,(2009). Neural networks applied to estimating subglacial topography and glacier volume. J. Clim. 22, 2146–2160.
- Maussion, F. et al.,(2019) The Open Global Glacier Model (OGGM) v1.1. Geosci. Model Dev. 12, 909–931.
- Zekollari, H., Huss, M. & Farinotti, D.,(2019). Modelling the future evolution of glaciers in the European Alps under the EURO-CORDEX RCM ensemble. The Cryosphere 13, 1125–1146.
- Anshuman Misraa, Amit Kumara, Rakesh Bhambria, Umesh K. Haritashyab, Akshaya Vermaa, Dwarika P. Dobhala, Anil K. Guptac, Gaurav Guptad, Rajeev Upadhyaye,(2020). Topographic and climatic influence on seasonal snow cover: Implications for the hydrology of ungauged Himalayan basins, India,Journal of Hydrology.
- Lenin Campozano, Leandro Robaina, Luis Felipe Gualco, Luis Maisincho, Marcos VillacÃs, Thomas Condom, Daniela Ballari, Carlos PÃ¡ez,(2021). Parsimonious Models of Precipitation Phase Derived from Random Forest Knowledge: Intercomparing Logistic Models, Neural Networks, and Random Forest Models.
- Zhihua He, Doris Duethmann, Fuqiang Tian,(2021). A meta-analysis based review of quantifying the contributions of runoff components to streamflow in glacierized basins.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2021). An introduction to Statistical Learning with applications in R. 2nd edition: Springer Science+Business Media, LLC, part of Springer Nature 2021.
- David Hartmann, Philip Kraaijenbrink, Walter Immerzeel,(2022). Impacts on glacier mass balance in High Mountain Asia assessed using machine learning.
- C. Bouchayer, J. M. Aiken, K. Thogersen, F. Renard, T. V. Schuler,(2022). A Machine Learning Framework to Automate the Classification of Surge-Type Glaciers in Svalbard.
- Jordi Bolibar, Antoine Rabatel, Isabelle Gouttevin, Harry Zekollari, Clovis Galiez,(2022). Nonlinear sensitivity of glacier mass balance to future climate change unveiled by deep learning.