Leveraging Machine Learning and Google Earth Engine for Snowline Altitude Analysis: Insights from the Parbati Basin, India

Ray, A.,[1] Raavi, S.,[2] Gaddam, V. K.,[3*] Prasad, S. K.,[4] Ranjan, R.[5] and Gangadhar, K.[2]
[1]Department of Geoinformatics, Symbiosis Institute of Geoinformatics, Shivaji Nagar, Pune, India – 412115
[2]Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India-521137
[3]Department of Civil Engineering, VR Siddhartha Engineering Collee, Siddhartha Academy of Higher Education, Vijayawada, Andhra Pradesh, India- 520010
E-mail: vinaygaddam@vrsiddhartha.ac.in*
[4]Department of Computer Science, SRM University, Vijayawada, Andhra Pradesh-522240
[5]DST Excellence of Excellence, Sikkim University, 6th mile, Samdur, P. O.Tadong, Gangtok, Sikkim India- -737102
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v20i9.3545

Abstract
Glaciers are highly responsive to climate variations, yet monitoring them in the rugged Himalaya region poses significant challenges. This study explores the effectiveness and cost-efficiency of using machine learning models integrated with remote sensing data from Google Earth Engine (GEE) to map glacier accumulation (snow) zones in the Pārbati Valley. We tested various machine learning algorithms, including Otsu (image segmentation), K-means and cascade K-means (unsupervised classification), and random forest, minimum distance, smile CART, naive Bayes, robust tree, and support vector machine (supervised classification). Our analysis shows that the Otsu method, along with K-means, cascade K-means, and all supervised classification methods except smile CART and naive Bayes, perform similarly in mapping snowlines. Notably, the Otsu method achieved a maximum predictable error of 57 meters, which is a substantial improvement over traditional methods and indicates higher accuracy in snowline mapping. The study reveals that the regional snowline in the Pārbati Valley ranged between 5048 meters and 5113 meters during the study period. Given its superior performance, the Otsu method is recommended for identifying snowline altitudes across a wide range of glaciers in the Himalayas.

Keywords: Snowline Altitude, Supervised Classification, Unsupervised Classification, Image Segmentation, Parbati Basin

1. Introduction

The Himalayas, with their extensive ice reserves including glaciers, snow, and permafrost play a crucial role in the Earth's climate system. These ice masses regulate the radiation balance and provide essential meltwater to downstream populations, supporting agriculture, irrigation, hydropower, and aquaculture, especially during dry periods. Recent climate change has accelerated glacier retreat in the region, causing significant ice loss and increasing the risk of disasters such as glacial lake outburst floods, rockslides, and landslides. These events have led to substantial damage to infrastructure and communities, resulting in loss of life and property. Monitoring these cryospheric components is essential due to the global implications of such climate-driven changes.

The Pārbati Valley is particularly important for study due to its sensitivity to climate change and its impact on local and global water resources. This valley hosts several hydropower projects that rely on glacier-fed rivers for energy generation. Meltwater from these glaciers is critical for maintaining river flows necessary for hydropower plants, a key component of the region's energy infrastructure. Moreover, the glaciers in the Pārbati Valley significantly influence the regional climate system by reflecting solar radiation and helping regulate temperature.

Their ice and snow cover affect local weather patterns and help maintain the radiation balance, which can impact broader climatic conditions. The glaciers are also associated with potential hazards, such as glacial lake outburst floods (GLOFs) and landslides, making it crucial to monitor and understand glacier dynamics to manage and mitigate these risks. Additionally, the rivers and streams fed by glacial melt support diverse ecosystems, providing habitats for various species and maintaining ecological balance.

Numerous studies have been conducted to investigate glacier variations in terms of geomorphological changes, mass loss, and Equilibrium-Line Altitude (ELA) [1], where annual accumulation equals ablation [2][3] and [4]. For most Himalayan glaciers, where superimposed ice is absent, the Snow Line Altitude (SLA) at the end of the ablation period is considered as the ELA, which is rigorously monitored as it is crucial for evaluating mass balance [5][6][7] and [8]. Typically, the ELA represents a zero mass balance, where the total accumulated snow in a hydrological year equals the amount of ice lost from the glacier. A regression equation developed using field-based mass balance and satellite-based Accumulation Area Ratio is used to evaluate ELA and mass balance [9]. Studies suggest that a glacier's response to climate change is better assessed by ELA rather than changes in its area or length. This approach enhances understanding of the relationship between glaciers and climate at both glacier and regional scales. The former focuses on specific features and processes within individual glaciers, while the latter examines broader environmental factors influencing multiple glaciers over a larger area [10] and [11]. ELA is also a key parameter in studying the ablation gradient. Since ground-based mass balance observations in the Himalayas are limited to a few accessible glaciers, constraints like harsh weather, rugged topography, limited manpower, and insufficient financial support restrict field investigations.

Indirect methods for ELA estimation are based on morphological and parametric approaches. Morphological methods include cirque-floor altitudes (C-F), Maximum Altitude of Lateral Moraine (M-A-L-M), the toe-to-headwall altitude ratio (T-H-A-R), and mean glacier elevation (M-G-E) [12] and [13]. The parametric approach includes the Accumulation Area Ratio (AAR) and Accumulation Area Balance Ratio (AABR) methods [14]. The C-F method uses the altitude of the cirque's bottom to assess glacier ELAs. The M-A-L-M method considers glacier flow from the center in the accumulation zone to the edge in the ablation zone, with moraine deposition occurring only under the ELA, making the ELA higher than the M-A-L-M. The T-H-A-R method calculates glacier ELA by maintaining a constant ratio between the terminus and highest glacier altitude. These methods are simple to apply but are affected by geomorphological preservation and lack accuracy, as they do not consider glacier area or altitude information.

In contrast, the AAR and AABR methods are based on assumed glacier mass-balance gradients and can overcome the limitations of morphological methods. The AAR method provides a constant ratio of accumulation and ablation areas when a glacier is in a steady state. AAR can produce similar values for glaciers of different sizes. The AABR method considers both mass-balance gradient and hypsometry, developing a balanced ratio for ELA estimation. These methods are more accurate and reliable than morphological approaches, but each has its own benefits and limitations, and their use depends on data availability and the scale of investigation [15][16] and [17]. While these methods are applicable to Himalayan glaciers, the lack of field observations limits their large-scale applicability. As a result, remote sensing-based ELA studies have become an alternative for estimating mass balance in Himalayan glaciers.

Remote sensing is a rapidly evolving technology that monitors changes in surface objects by providing data with high spatio-temporal, radiometric, and spectral resolutions. In this context, remote sensing is particularly effective for mapping the Snow Line Altitude (SLA) of glaciers on a large spatial scale. To determine SLA from remote sensing data, geospatial analysts extract glacier areas, delineate snow-ice boundaries using spectral characteristics, and use a Digital Elevation Model (DEM) to estimate snowline altitudes. Various data processing methods are employed, including band ratio techniques (e.g., NDSI [18]), automated spectral combination analyses (e.g., Principal Component Analysis, Fuzzy Logic [19]), and automated threshold methods (e.g., Otsu algorithm [20]). These methods utilize the spectral reflectance properties of snow and ice to differentiate them from other surface features like rock, soil, or vegetation [20] and [21]. Traditional snowline detection methods have primarily relied on the Near-Infrared (NIR) wavelength region [22] and [23], using threshold values to separate snow from ice. However, few studies have focused on SLA estimation using machine learning methods via the Google Earth Engine (GEE) interface.

Machine learning methods are increasingly preferred for their ability to identify trends and patterns in datasets, operate without human intervention, handle multi-dimensional and multi-criteria datasets, and be applied to a wide range of applications. The main drawback of machine learning methods is the time-consuming process of data collection. GEE is favored as it is a free, open-source platform with a vast repository of geospatial data that facilitates the visualization and analysis of spatial information of surface objects. This study focuses on using machine learning algorithms such as Otsu, K-means, cascade K-means, Random Forest, Naive Bayes, Smile-CART, and Support Vector Machine through the GEE interface to analyze the Equilibrium Line Altitudes (ELAs) of glaciers and their changes in the Parbati basin from 2013 to 2023, covering a decade-long period. The accuracy of the results is assessed using standard deviation and Mean Absolute Error (MAE), and the findings are compared with manual delineation to highlight the advantages of machine learning methods over traditional techniques.

2. Study Area

This study aims to evaluate the snowline altitudes of glaciers located in the Parbati basin, a part of the Beas River system in the Kullu district of Himachal Pradesh, India. Within the total basin area of 1,774 km², 71 glaciers cover an area of 346.20 ± 3.6 km². These glaciers range in elevation from 4,235 to 5,600 meters above sea level. The Parbati Glacier is one of the most prominent and largest glaciers in the Parbati River basin. Figure 1 shows the study area, including the locations of these glaciers, along with their altitude distribution, areal distribution, and drainage patterns.

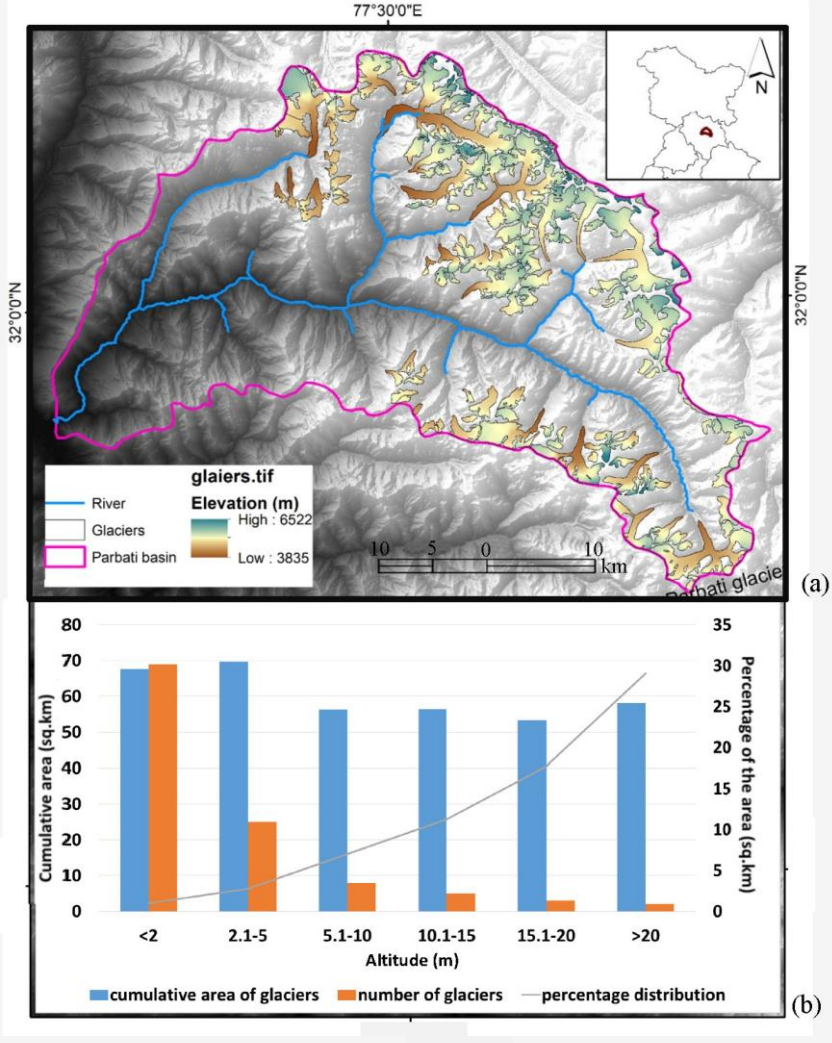


Figure 1: (a) Himachal Pradesh, India (b) glacia altitude and cumulative area

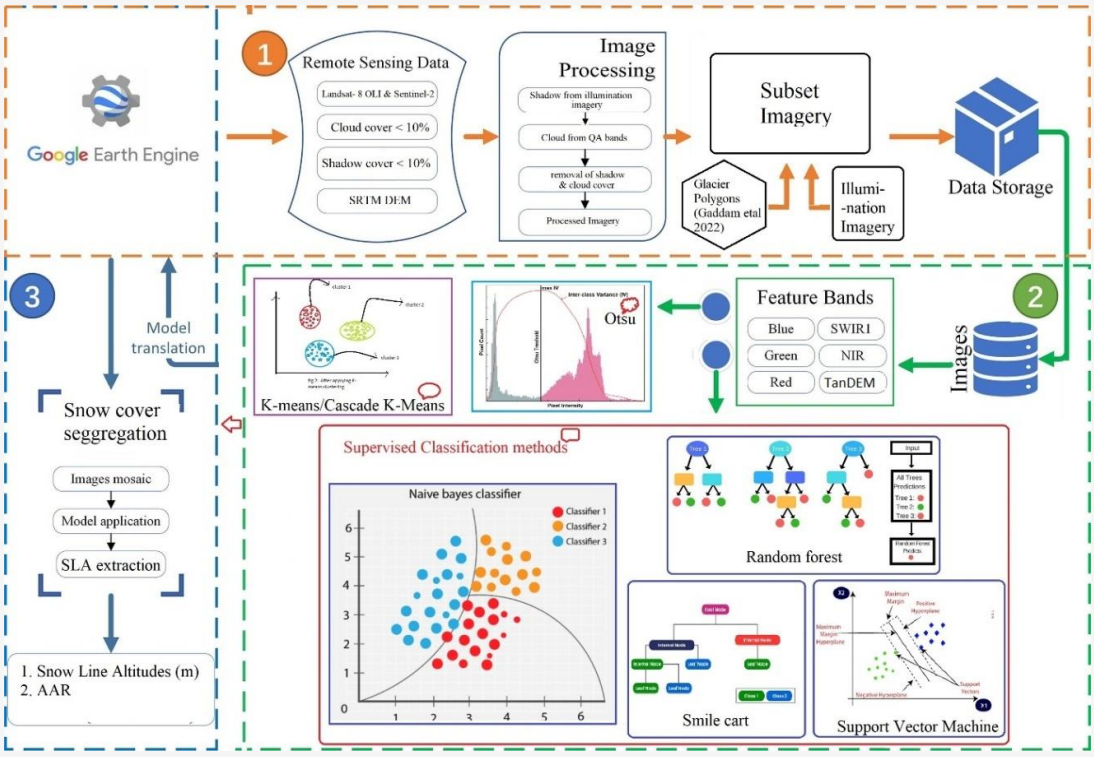


Figure 2: Study workflow executed through Google Earth Engine for snow line altitude estimation for Parbati basin using various machine learning algorithms

3. Data Source and Methodology

Google Earth Engine (GEE) was chosen for this analysis because it is a cloud-based platform that provides unrestricted access to extensive satellite imagery databases, allowing for the efficient retrieval, visualization, and analysis of large datasets. GEE offers several advantages: it is cost-effective due to its open-source nature, user-friendly with intuitive computational tools, and highly efficient at generating high-quality visualizations through advanced processing methods and algorithms. These features help achieve optimal results with minimal errors. The detailed analysis of Snow Line Altitude (SLA) interpretation in this study is illustrated in Figure 2, involving image segmentation, supervised, and unsupervised algorithms applied to stacked imagery of the Parbati basin for hydrological years from 2012-13 to 2021-22.

The imagery data utilized from GEE includes Landsat 8 OLI, Sentinel-2A, Tandem elevation datasets, and glacier extents imported from the Randolph Glacier Inventory (RGI). The initial step involves sorting the satellite imagery based on the acquisition date. During this step, overlapping images within the study area on the same acquisition date are identified. For further processing, only single-tile images with clear contrast for each acquisition date are selected.

Images partially obscured by shadows, which may result from cloud cover or the surrounding glacier topography, are also identified. To address this, morphological filters and local illumination angle images are created by calculating the local illumination angle using satellite and topographical data. A threshold value of 54175 is set to remove cloud-covered pixels, ensuring a clearer dataset. The refined subset of satellite data, free from clouds and shadows and accurately outlining glacier boundaries, is then used to separate snow and ice and extract SLAs using machine learning frameworks through GEE.

3.1 OTSU (Image Segmentation) Algorithm

OTSU is a method used for adaptive thresholding in image segmentation and binarization [24] and [25]. It determines the optimal threshold value for an input image by evaluating all possible thresholds to minimize the intra-class variance. The result is a single intensity threshold image that separates pixel values into two categories: foreground (snow) and background (ice and other features). This algorithm has been successfully applied to differentiate snow from ice and to delineate snowline altitudes in the Chandra River basin, yielding satisfactory results [24]. Equation 1 illustrates the mathematical formulation of the Otsu algorithm.

$$\sigma_{\omega}^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t)$$

Equation 1

Where, ω_0 and ω_1 are the probabilities of two defined classes separated with a threshold t and σ_0^2, σ_1^2 as variances.

3.2 K-means Algorithm

The K-means clustering algorithm organizes an unlabeled dataset into distinct clusters by computing the distance between each data point and the cluster centers. It then assigns each data point to the cluster center with the minimum distance, as determined among all the cluster centers (see Equation 2).

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\left\|x_i - v_j\right\|\right)^2$$

Equation 2

Where:

$\|x_i - v_j\|^2$ is the Euclidean distance between x_i and v_j

x_i is the number of data points in i^{th} cluster

c_i is the number of cluster centers

3.3 Cascade K-means Algorithm

The cascade K-means algorithm identifies clusters of points in multidimensional Euclidean space by incorporating both agglomerative and divisive approaches, allowing it to discern spatial relationships among the points. The computational formulation of the cascade K-means algorithm is provided in Equation 3.

$$K\text{-mean} = (SSB/(K-1))/(SSW/(n-K))$$

Equation 3

Where n is the number of data points, K is the number of clusters, *SSW* is the sum of squares within the clusters while *SSB* is the sum of squares among the clusters.

3.4 Random Forest Algorithm

The Random Forest supervised algorithm combines the results from multiple decision trees to produce a single output, using a technique known as "bootstrap aggregating" (Equation 4). This method is both straightforward and flexible, effectively handling both classification and regression tasks. By increasing the number of trees, the algorithm improves accuracy and reduces overfitting. It also has the advantage of training data more quickly compared to other algorithms and maintains high accuracy even when a substantial portion of the data is missing.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_0(x')$$

Equation 4

Where, Training dataset x is represented by $x_1,\ldots,x_n$, with responses as $Y= y_1,\ldots,y_n$ bagging repeatedly for b = 1,…., B, and x' is the predictions for unseen samples.

3.5 Smile CART Algorithm

The Statistical Machine Intelligence and Learning Engine (SMILE) includes a suite of classification algorithms that feature the Classification and Regression Tree (CART) approach. CART is a predictive method that uses a binary decision tree classifier to make straightforward decisions based on logical if-then questions. Initially, the classifier evaluates the input variables to identify the one with the most information, determining the node splits at each level. The input data is randomly divided into multiple subsets and trees, with one subset held out for validation. The pruned tree is then adjusted to minimize deviance, a concept known as Gini impurity, which is mathematically defined in Equation 5.

$$Gini = 1 - \sum_{i=1}^{n}(pi)^2$$

Equation 5

Where pi is the probability of an object being classified into a particular class.

3.6 Support Vector Machine Algorithm

This algorithm is considered as a binary linear classifier that selects a hyperplane representing. The SVM algorithm seeks to achieve the most significant separation between two distinct classes. In cases where a hyperplane cannot be found, the algorithm attempts to separate positive and negative instances as effectively as possible. Nonlinear SVMs are created by applying a kernel function to maximum-margin hyperplanes. When a hyperplane is found, it is called the maximum-margin hyperplane, and the corresponding classifier is known as the maximum-margin classifier. Both linear and nonlinear SVMs produce similar results as they map data into a feature space. This mapping can be nonlinear, leading to a high-dimensional space. The various types of SVMs are detailed in Equations 6 to 9, with further description provided by [24].

Linear: $K(\omega,b) = \omega^T x + b$

Equation 6

Polynomial: $K(\omega,x) = (\gamma\omega^T x + b)^N$

Equation 7

Gausian RBF: $K(\omega,x) = exp(-\gamma||x_i - x_j||^n)$
Equation 8

Sigmoid: $K(x_i, x_j) = \tanh(\alpha x_i^T x_j + b)$
Equation 9

3.7 Naïve Byes Method

The Naïve Bayes algorithm is a straightforward method for predictive modeling, using two key approaches to compute outputs from training data: a) calculating the probability of each class and b) determining the conditional probability for a given class based on a specific "x" value. Once these probabilities are defined, the model uses Bayes' theorem to classify new data. For data with real values, probabilities are often modeled using a Gaussian distribution (bell curve). A key advantage of the Naïve Bayes theorem is its assumption that each input variable is independent. The Bayes method (Equation 10) is particularly effective for handling data with a wide range of values, classifying data by maximizing (P(O|C_i)P(C_i)), where (O) represents the object and (i) is the class index. The Bayes theorem finds the category of class having higher posterior probability:

$$PC_iO = \frac{POC_iPC_i}{PO}$$

only if, $P(C_i|\ O) > P\ (C_j|\ O)$ *for* $1 \leq j \leq n$, *where* $j \neq i$
Equation 10

It includes the training set of objects, associated labels and classes as “D”, “O” and “C”. Each object is represented with an n-dimensional vector as shown below:

Objects *(O) = (O1,O2,…,On),*
Attributes = *A1, A2, A3,…, An*
Classes *(m) = (C1, C2,….., Cn),* with a given tuple *O.*

4. Results and Discussion

4.1 Labels Acquisition from Image Data and Model Training

Seven machine learning models were implemented using Google Earth Engine (GEE) to analyze Landsat and Sentinel imagery of the Parbati River basin. The goal was to interpret glacier facies, map snowline altitudes (SLA), and estimate SLA variations at both local and regional scales. Initially, index labels for model training were generated from the imagery. To cover an entire hydrological year for the Parbati basin, approximately 80 to 85 satellite scenes are needed. However, during the accumulation season (November to April), heavy snow and extensive cloud cover hinder SLA extraction. Consequently, the investigation period is limited to the ablation season, from June 15th to October 1st [25], reducing the usable imagery to between 28 and 33 scenes. These processed and de-clouded images serve as input data for training the models to classify snow and non-snow pixels. The training accounts for various glacier scenarios, including different surface features such as rivers, rocks, and vegetation [25] [26] and [27]. The deep learning approach minimizes redundant operations and prevents duplication in overlapping areas, thus improving efficiency. To analyze hypsometric information and SLAs, the images were re-projected to match the coordinates of Tandem-X data, which provides 12-meter spatial resolution and 5-meter vertical accuracy (obtained from tandemx-science@dlr.de).

4.2 Model(s) Prediction

The seven machine learning models were initialized using 358 images with spectral combinations including Green, NIR, SWIR2, slope, and DEM. The image data were transformed into approximately 1,433 labels to create 52,83,840 training samples, with an input format of 8 × 16 × 4 (number of images × temporal availability × number of bands) and an output format of 1 × 1 × 3 × 6 (one-hot encoding of labeled data). Of the total 52.83 million pixel samples, 28.08 million were non-snow samples, resulting in a balanced ratio of 0.53:0.47. The models were implemented using the GEE code editor, with each model iterated approximately five times to achieve optimal accuracy. All models exhibited similar training iteration curves, reaching maximum accuracy after the fifth iteration. Figure 3(a) depicts the accuracy and loss of pixel values per image, while Figure 3(b) shows the F1 score across all iterations. Accuracy improved consistently across models, stabilizing at 93.66% after the fifth iteration, indicating that the five features used were effective in training the model to extract snow cover pixels. The F1 score, a measure of accuracy that accounts for precision and classification quality, peaked at 0.85 after the fifth iteration. Consequently, models trained up to the fifth iteration were selected for snow cover pixel classification. A 1024 × 1024-pixel image set, including TM/OLI and S-2 spectral bands and corresponding glacier masks, was collected at a 0.16° interval (approximately every 10 minutes), yielding a total of 52.83 million pixel values. The final glacier dataset was noise-corrected to obtain standard snow and non-snow pixel values.

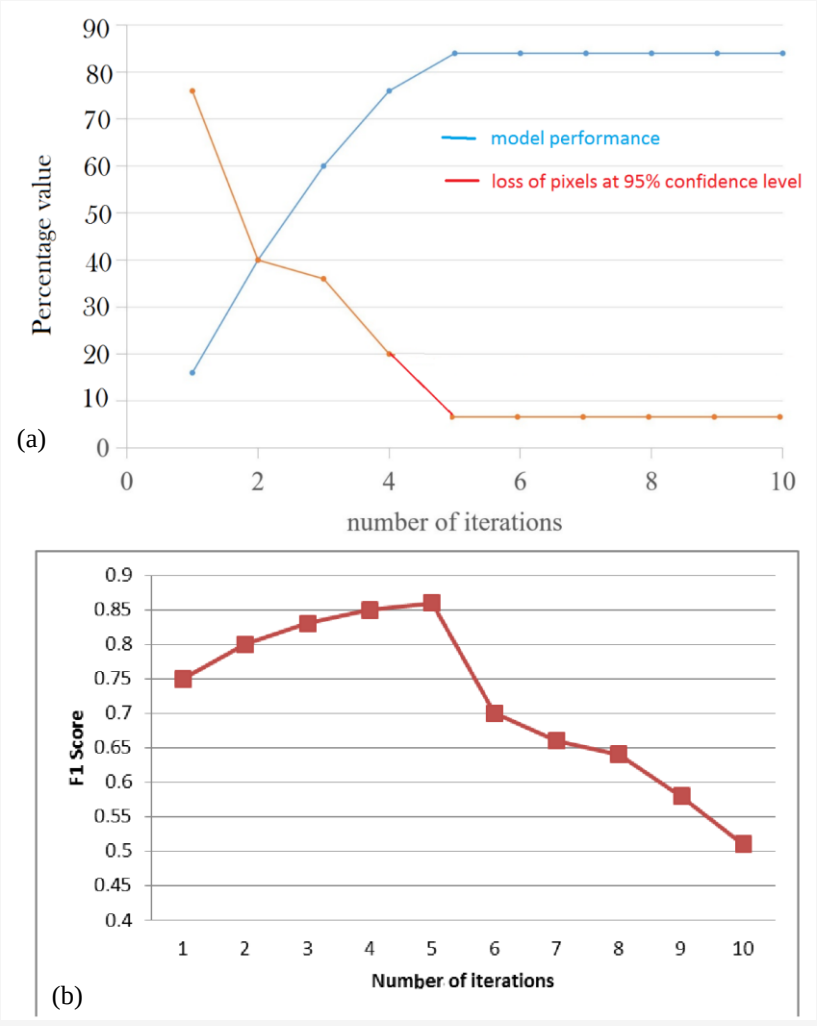


Figure 3: The model performance evaluation (a) loss and accuracy (b) F1 score

Figure 3(a) illustrates that model performance was highest after the fifth iteration, with minimal pixel loss during interpolation. Figure 3(b) represents the F1 score, reflecting the model's best performance across iterations, confirmed that the parameters tuned in the fifth iteration were optimal for the study.

4.3 Snowline Extraction Results and Accuracy

The trained models effectively identified snow and non-snow pixels on glaciers at a regional scale by classifying glacier facies. Figures 5(a)- 5(g) display the accumulation areas identified and interpreted by each machine learning algorithm. Additionally, these interpreted accumulation area pixels were correlated with elevation values to assess the spatial distribution of snow and firn in the glaciers' accumulation zones. The analysis showed that the seven machine learning methods detected a total glacial area of 346.20 km². However, each method classified the glaciated area differently, as detailed in Table 1 and Figure 5. These variations in snow cover pixel interpretation are due to the distinct characteristics and predefined procedures of each model.

The results indicate that image segmentation (Otsu), unsupervised methods (K-means and Cascade K-means), and Support Vector Machine (SVM) exhibited similar patterns in snow cover extraction when compared to supervised classification methods like Smile CART and Naïve Bayes, at both the glacier and regional scales (Figures 4(a)-4(g) and Figures 6(a)-6(g)). The supervised classification methods, which include a sub-level pixel classification mechanism, were effective in detecting saturated and mixed area pixels. Overall, the Otsu image segmentation algorithm demonstrated superior accuracy compared to the supervised methods, based on the robustness of calculations and training parameters (Table 1 and Figure 5). Similar results were reported by [24] for snowline monitoring in the Chandra basin using the Otsu method.

Table 1: Information on classification of snow cover pixels, error in mapping accumulation area and mean regional snowline altitude by each method for Parbati basin between 2013 and 2023

Method	Error in mapping snow cover area (km^2) at 95% confidence interval	Mean absolute error (%)	Mean snowline altitude of the basin
Otsu	5.19	1.50	5082
K-means	5.19	1.50	5062
Weka cascade K-means	5.19	1.50	5062
Random forest	3.18	0.92	5072
SVM	3.98	1.15	5089
Smile cart	4.77	1.38	5113
Naive byes	4.62	1.33	5056
Average	4.59	1.33	5077

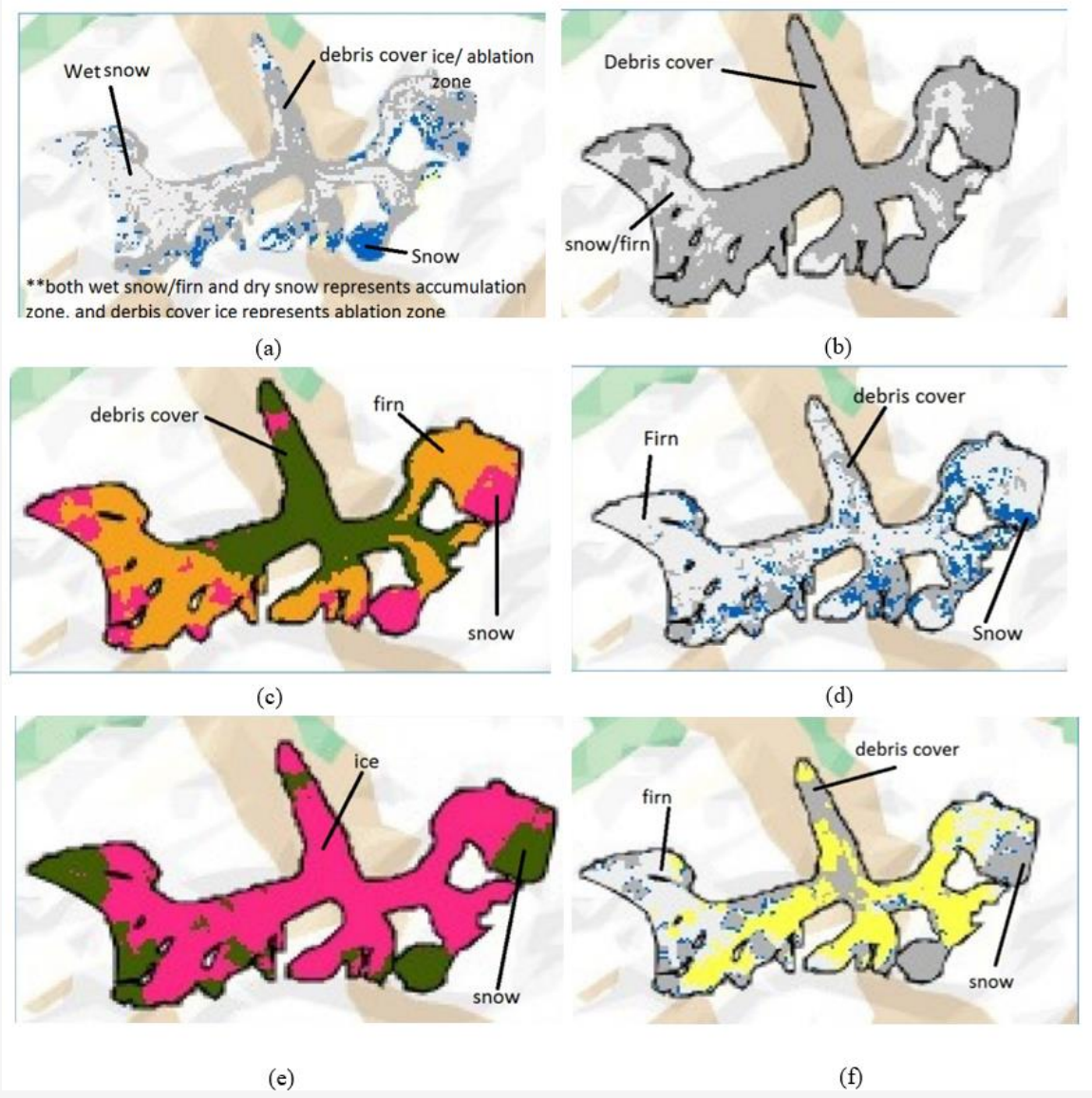


Figure 4: Accumulation and ablation area identification using ML algorithms: (a) SVM (b) Smile CART (c) K-means and Cascade K- means (d) Random Forest (e) OTSU and (f) Naïve Byes

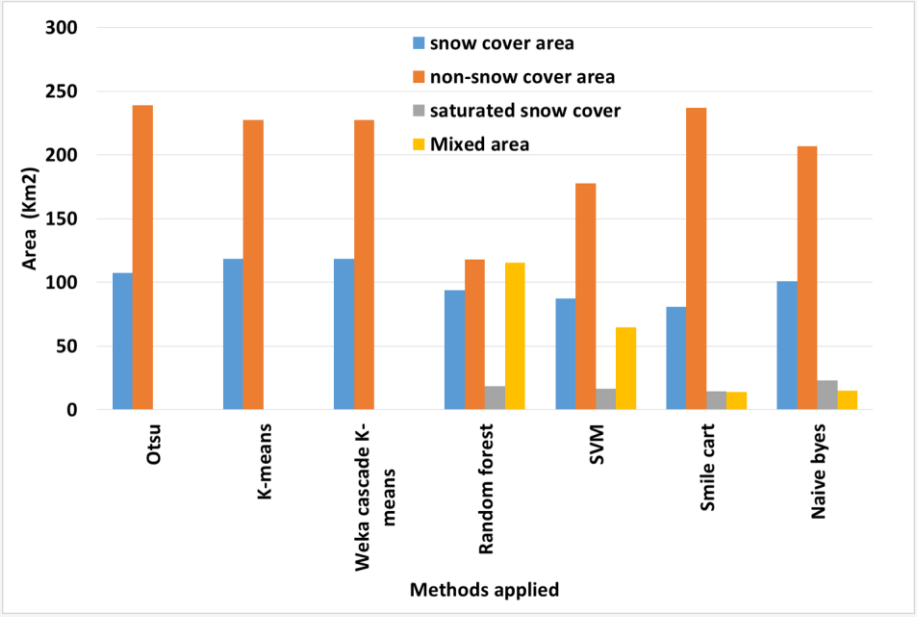


Figure 5: Classified areas of snow cover, non-snow cover, saturated and mixed pixels by machine learning algorithms

Table 2: Assessment of regional scale ELA of the basin in the study period using Otsu algorithm

Time of investigation	Satellite	Snow area (km^2)	Non-snow area (km^2)	2 standard deviation	Error in mapping snow cover area at 95% confidence interval	Mean absolute error (%)	Mean snowline altitude of the basin (Otsu)	Mean snowline altitude of the basin (manual)	AAR
18-Sep-13	Landsat 8 OLI	178.55	167.65	1.54	1.00	1.54	5057	5029	0.52
21-Sep-14	Landsat 8 OLI	155.88	190.32	4.87	3.17	4.88	5049	5025	0.45
08-Sep-15	Landsat 8 OLI	143.62	202.58	8.34	5.42	8.35	5047	5024	0.41
17-Sep-16	Landsat 8 OLI	111.73	234.47	17.36	11.28	17.39	5073	5037	0.32
13-Sep-17	Landsat 8 OLI	160.04	186.16	3.70	2.40	3.70	5050	5025	0.46
16-Sep-18	Landsat 8 OLI	161.52	184.68	3.28	2.13	3.28	5050	5025	0.47
12-Sep-19	Landsat 8 OLI	137.70	208.50	10.01	6.51	10.03	5048	5024	0.40
19-Sep-19	Sentinel-2A	140.66	205.54	9.17	5.96	9.19	5047	5024	0.41
21-Sep-20	Landsat 8 OLI	111.45	234.75	17.44	11.33	17.47	5074	5037	0.32
21-Sep-20	Sentinel-2A	101.14	245.06	20.35	13.23	20.39	5113	5057	0.27
06-Sep-21	Sentinel-2A	138.96	207.24	9.66	6.28	9.67	5048	5024	0.40
10-Oct-21	Landsat 8 OLI	137.89	208.31	9.96	6.47	9.98	5048	5024	0.40
10-Sep-22	Landsat 8 OLI	159.37	186.83	3.88	2.52	3.89	5050	5025	0.46
11-Sep-22	Sentinel-2A	107.32	238.88	18.60	12.09	18.64	5082	5041	0.31
06-Sep-23	Sentinel-2A	153.95	192.25	5.42	3.52	5.43	5048	5024	0.44
08-Oct-23	Landsat 8 OLI	158.88	187.32	4.02	2.61	4.03	5049	5025	0.46

Table 1 provides classified values for snow and non-snow cover regions, along with the corresponding Equilibrium Line Altitudes (ELAs) and Accumulation Area Ratios (AAR) from 2013 to 2023, including 2-sigma (standard deviation) and Mean Absolute Error (MAE). Supplementary Table 1 lists the imagery used and the dates when the maximum snowline altitude was observed. A maximum deviation of 57 meters in snowline altitude predictions was observed across all methods at the regional scale (Figure 7). As a result, the Otsu image segmentation method was selected to interpret snow and non-snow cover areas at the regional scale for the study period (Table 2), showing snowline variations between 5048 and 5113 meters (Figure 8). Manual interpretation was also carried out for selected glaciers to validate the results, with differences between manual and Otsu methods ranging from 10 to 29 meters, as shown in Figure 8. These findings suggest that integrating machine learning algorithms with remote sensing data provides a high level of accuracy in delineating snow cover areas and estimating SLAs on a large spatial scale in the challenging terrain of the Himalayas.

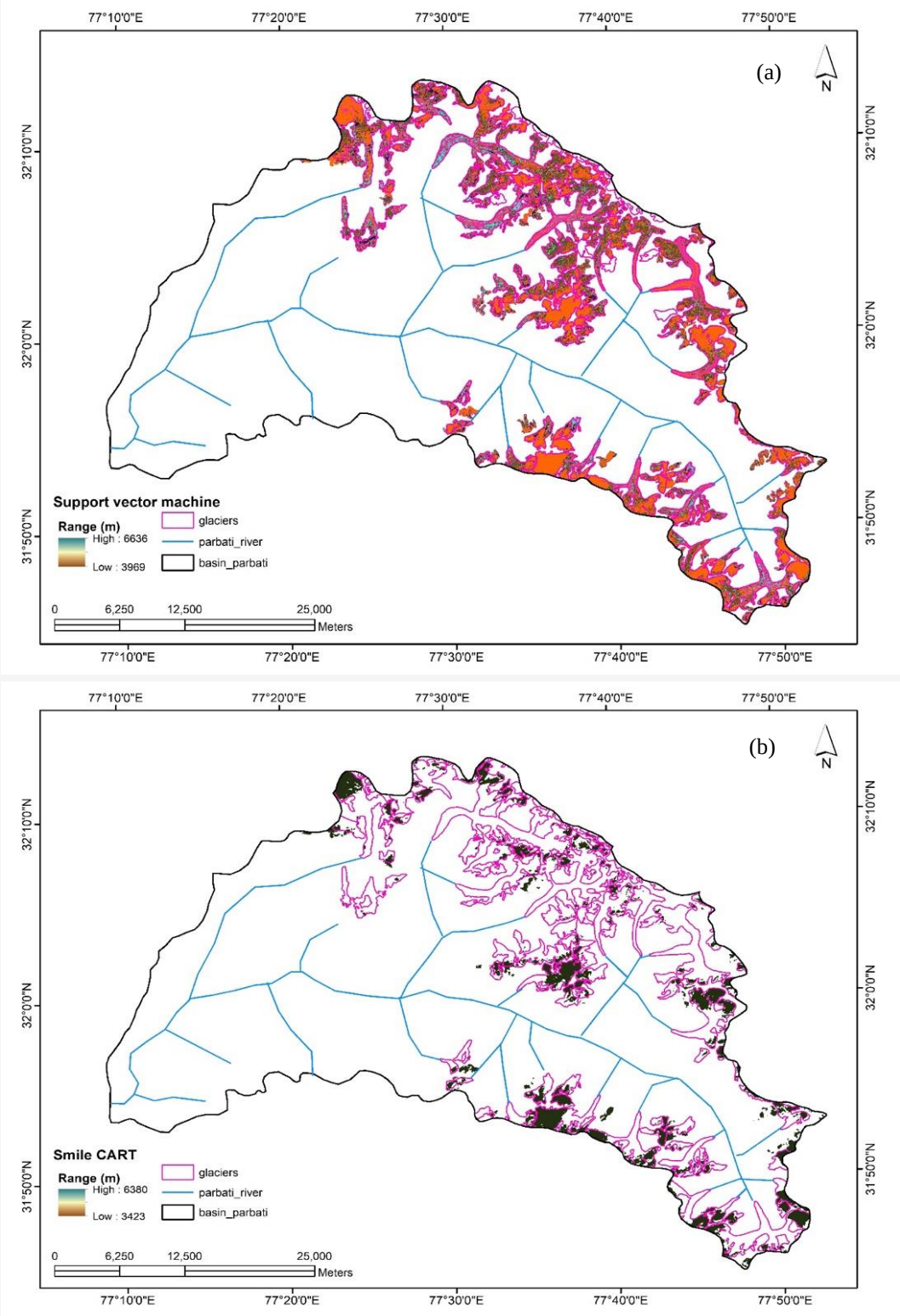
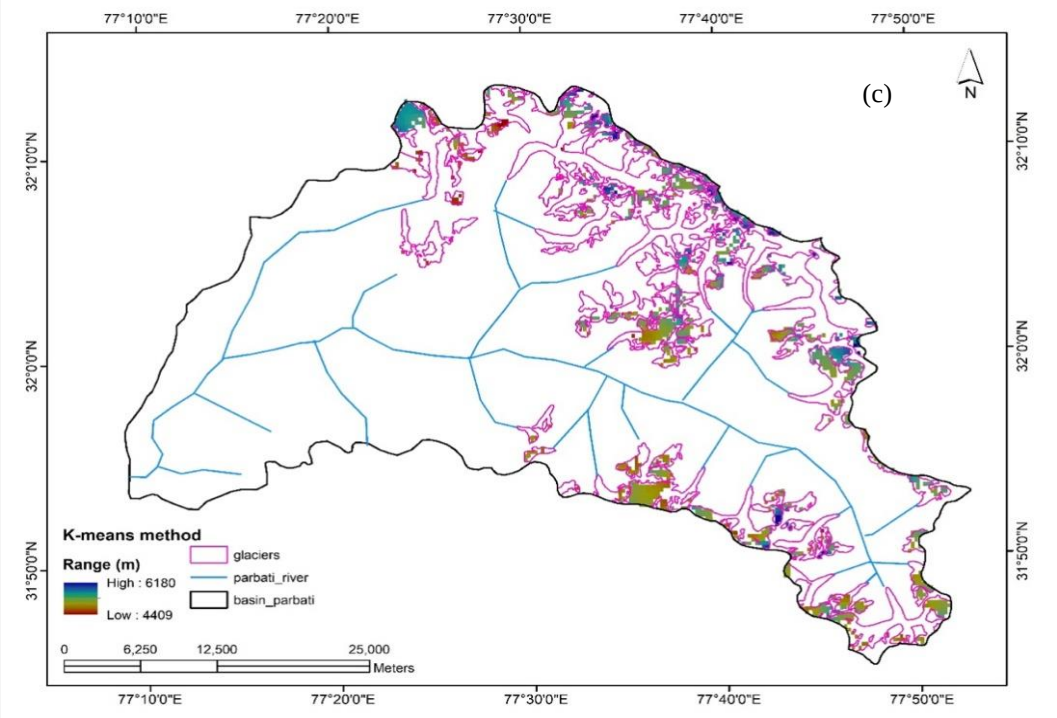


Figure 6: Snow altitude derived from various ML algorithms:
(a) SVM (b) Smile CART



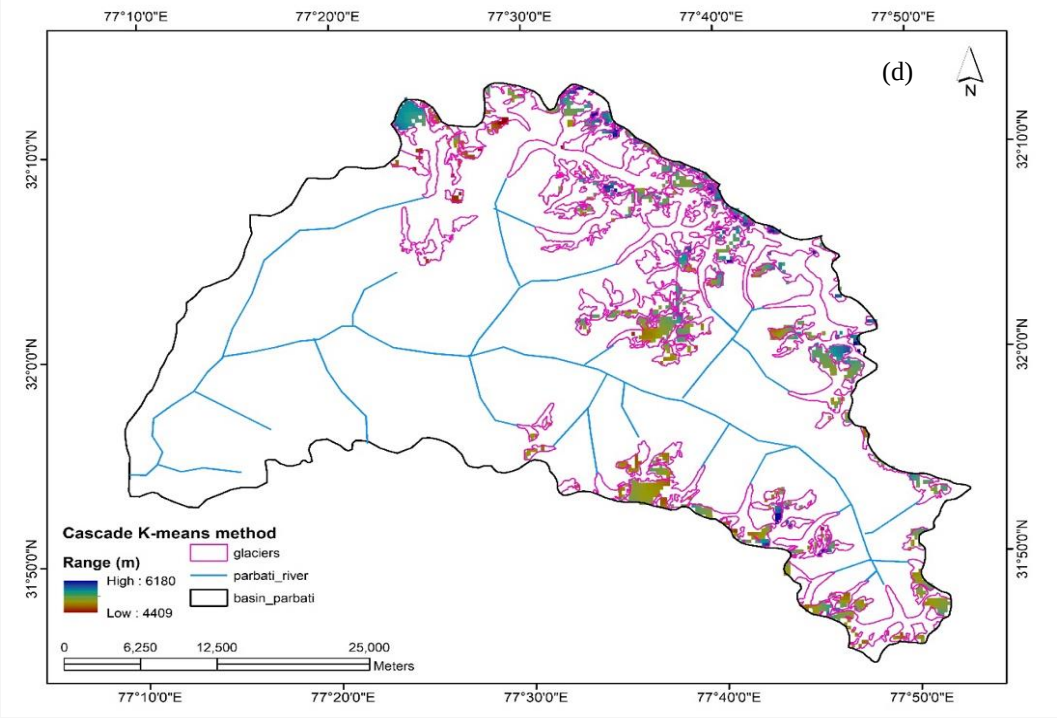
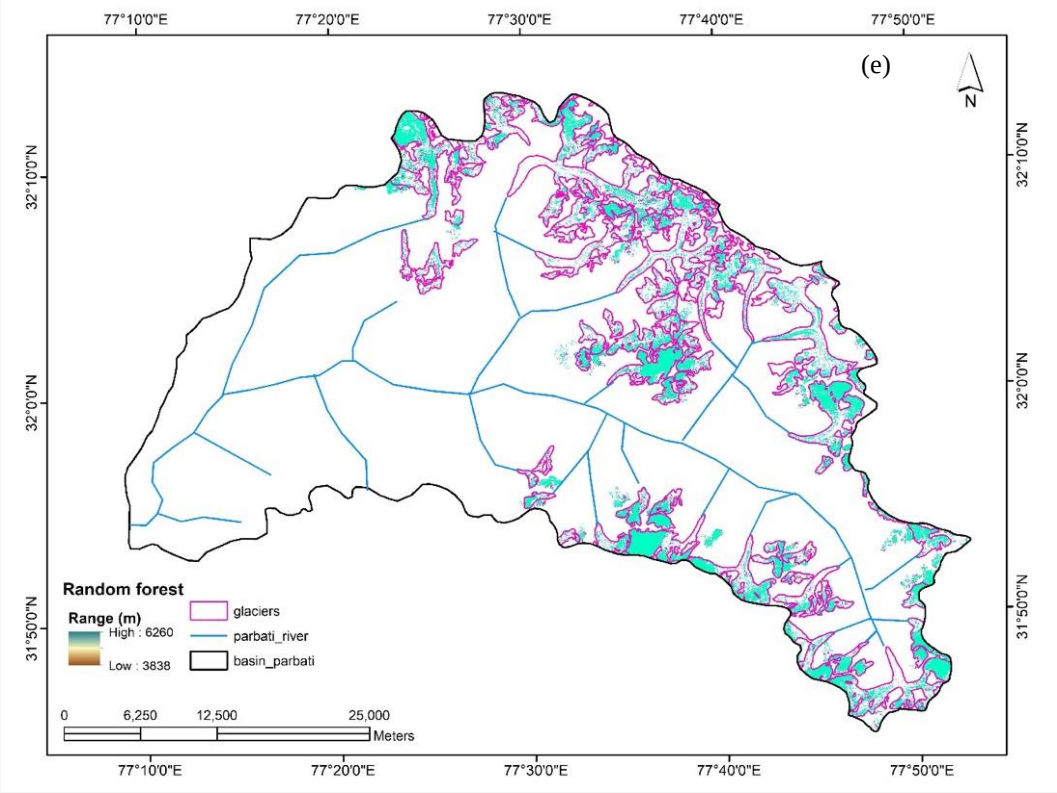


Figure 6: Snow altitude derived from various ML algorithms:
(c) K-means (d) Cascade K- means



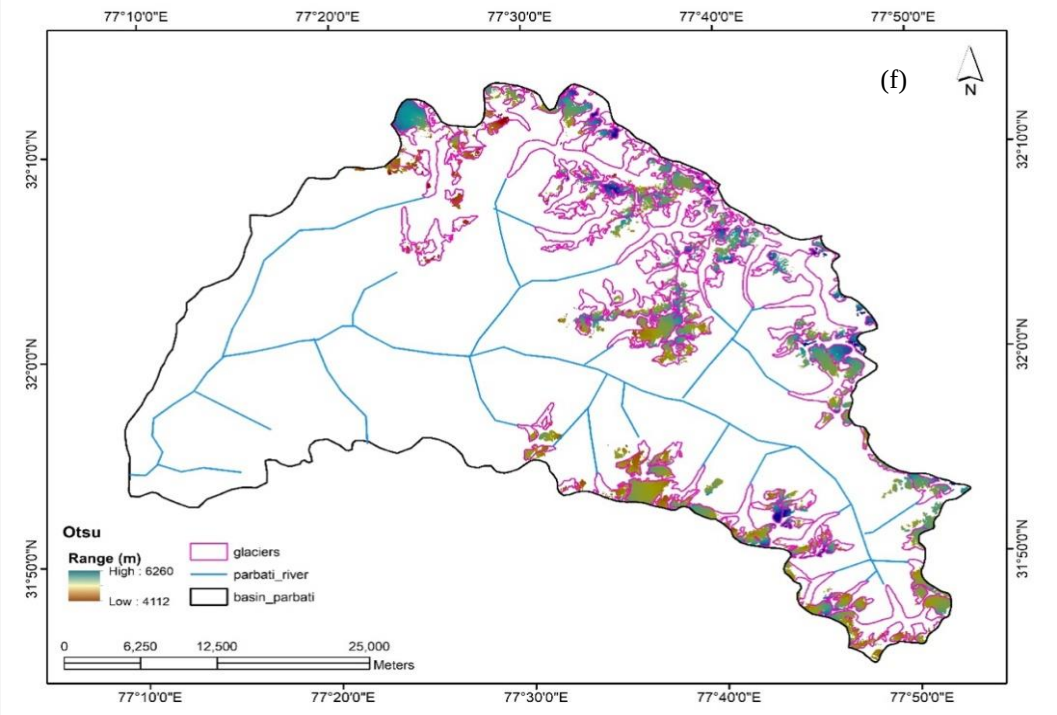


Figure 6: Snow altitude derived from various ML algorithms:
(e) Random forest (f) OTSU

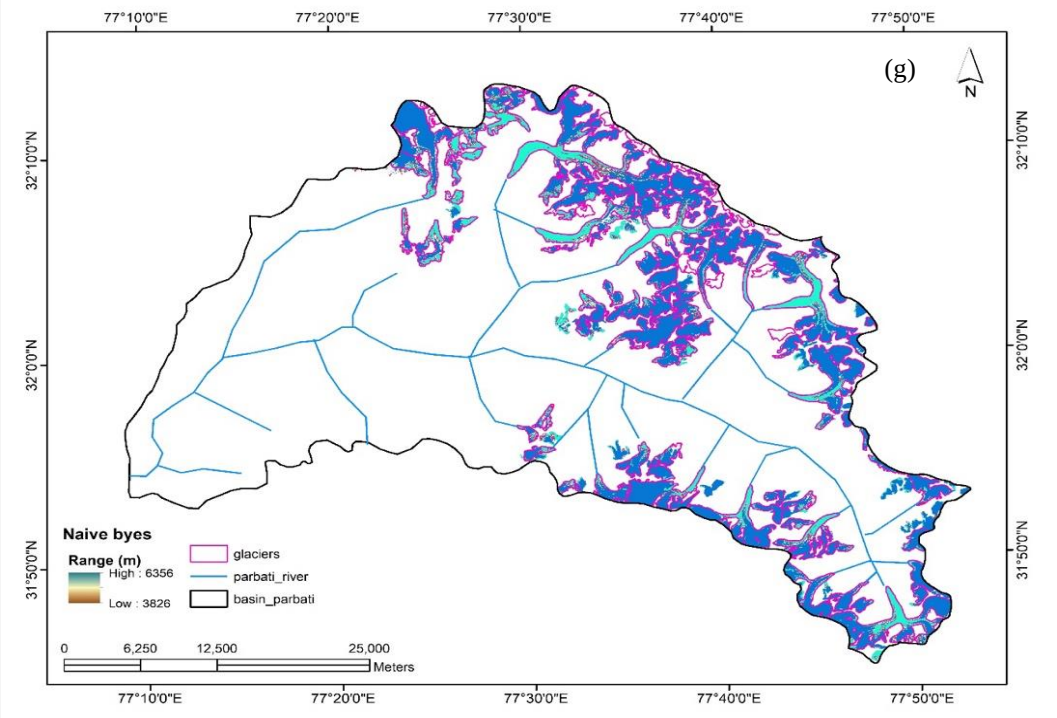


Figure 6: Snow altitude derived from various ML algorithms: (g) Naïve Byes

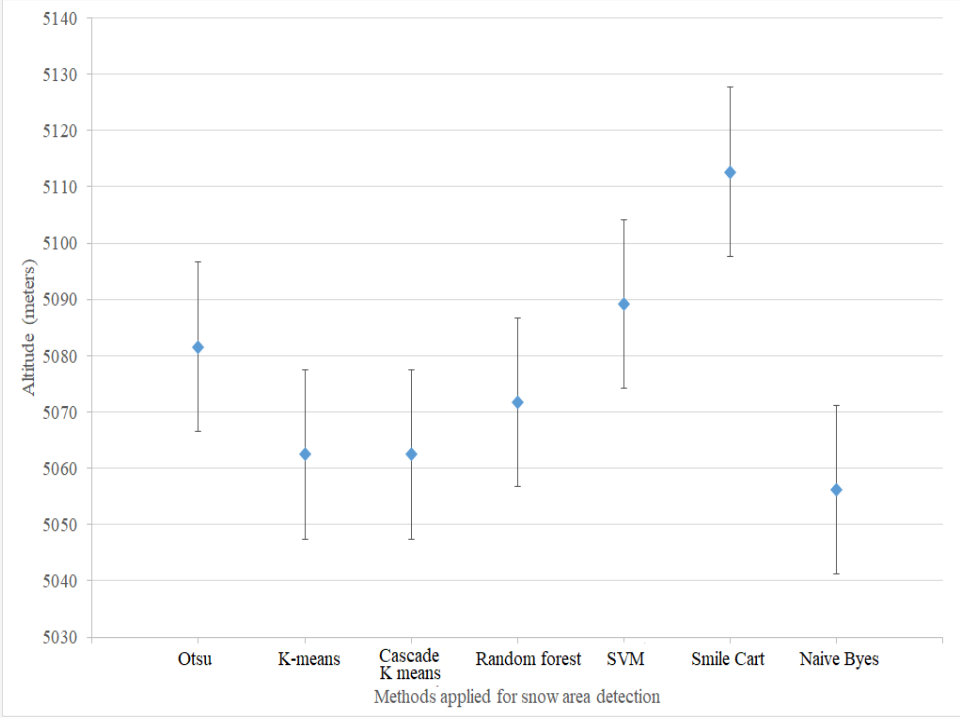


Figure 7: Identification of ELA's at regional scale in the Parbati basin using different machine learning algorithms

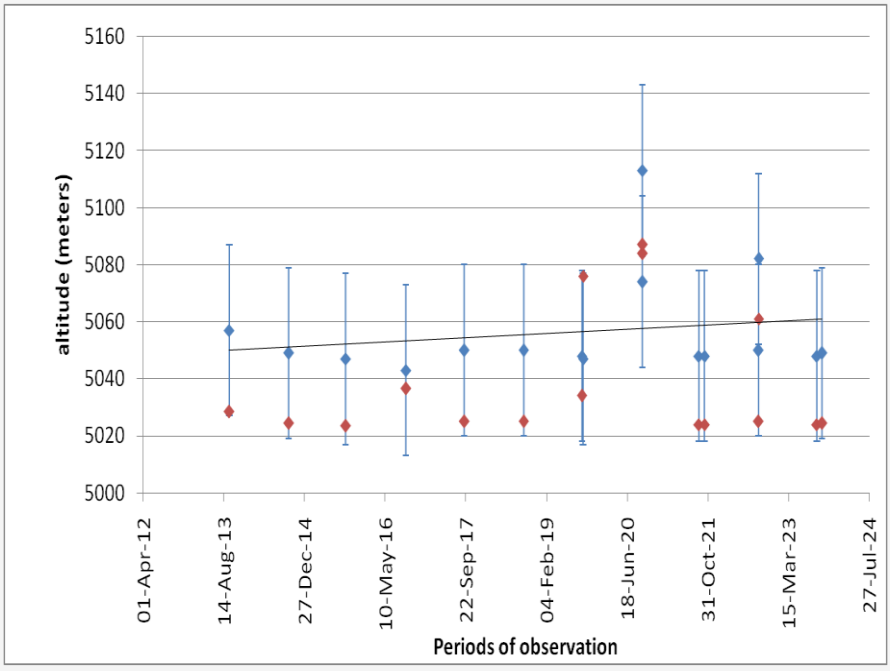


Figure 8: Snowline altitudes estimated using Otsu and manual delineation methods at regional scale in Parbati basin during the study period

4.4 Limitations of the Machine Learning Methods

Data-driven methods such as Support Vector Machine (SVM), Smile CART, K-means/Cascade K-means, Random Forest, Otsu, and Naïve Bayes are powerful tools in remote sensing and other fields for classification and prediction tasks. However, each methods have its own limitations such as:

4.4.1 Support Vector Machine (SVM)

1. High Computational Cost: SVMs, especially with large datasets and high-dimensional spaces, can be computationally expensive, requiring significant time and resources.
2. Choice of Kernel: The performance of SVM heavily depends on the choice of the kernel function. Selecting the appropriate kernel and its parameters can be challenging and may require domain expertise.
3. Less Effective with Noisy Data: SVMs can be sensitive to noise and overlapping classes. Outliers can affect the decision boundary, leading to less accurate classifications.
4. Binary Classification Limitation: SVMs are inherently binary classifiers and require extensions to handle multi-class problems, which can add complexity.

4.4.2 Smile CART (Classification and Regression Trees)

1. Prone to Overfitting: Decision trees like Smile CART can easily overfit the data, especially when the tree depth is not properly controlled or when working with small datasets.
2. Instability: Small changes in the training data can result in significantly different trees, making them less stable than traditional methods.
3. High Variance: CART models can have high variance if not pruned properly, leading to poor generalization on unseen data.
4. Poor Performance with Imbalanced Data: These methods can perform poorly on highly imbalanced datasets, as they may end up biased towards the majority class.

4.4.3 K-means / Cascade K-means

1. Requires Predefined Number of Clusters: The K-means algorithm requires the number of clusters (K) to be specified in advance, which can be difficult to determine without prior knowledge.
2. Sensitive to Initialization: The algorithm's results depend on the initial selection of centroids, which can lead to different outputs with different runs.
3. Assumes Spherical Distribution: K-means assumes that clusters are spherical and equally sized, which may not be true for many real-world data distributions.

4. Not Suitable for Non-Convex Clusters: K-means struggles with identifying clusters of non-convex shapes or varying densities.

4.4.4 Random Forest

1. Complexity and Interpretability: Random Forest models, being an ensemble of many decision trees, can become complex and less interpretable compared to simpler, traditional models.
2. Computational Cost: Training a large number of decision trees and aggregating their results can be computationally expensive and time-consuming.
3. Overfitting Risk: While Random Forest reduces overfitting compared to a single decision tree, it can still overfit in cases where too many trees are built or where trees are very deep.
4. Sensitivity to Noisy Data: Although more robust than single-tree methods, Random Forest can still be influenced by noisy data, especially if not tuned properly.

4.4.5 Otsu's method

1. Assumes Bimodal Distribution: Otsu's method assumes a bimodal histogram of the pixel intensities, which may not be the case in many real-world scenarios, leading to suboptimal thresholding.
2. Not Suitable for Multimodal Distributions: If the data has more than two classes or multimodal distributions, Otsu's method might fail to provide accurate segmentation.
3. Global Thresholding: It uses a global threshold, which may not work well for images with varying illumination or local differences.
4. Limited to Single Feature: Otsu's method typically considers only grayscale intensity, making it less versatile compared to multi-feature methods.

4.4.6 Naïve Bayes

1. Strong Independence Assumption: Naïve Bayes assumes that all features are independent of each other given the class label, which is often unrealistic in real-world data, leading to suboptimal results.
2. Limited Handling of Complex Relationships: It cannot model interactions between features, making it less effective when complex dependencies exist among the features.
3. Performance on Large Datasets: While computationally efficient, Naïve Bayes may perform poorly on large datasets where the independence assumption does not hold.
4. Continuous Data Requires Discretization: Naïve Bayes often requires continuous data to be discretized, which can lead to a loss of information and affect model performance.

Despite these limitations, data-driven methods are powerful tools that offer advantages like automation, scalability, and the ability to handle complex patterns in large datasets, making them valuable for modern remote sensing and geospatial analysis.

5. Conclusions

This study highlights a) the significance of SLA as a crucial parameter for understanding glacier mass balance b) the applications of machine learning algorithms (Otsu, K-means, cascade K-means, random forest, naïve byes, smile cart, and support vector machine) in extracting the snowlines, and SLA variations. Findings reveal a diverse range of SLAs, fluctuating between 5048 and approximately 5113 ± 57 meters. The glaciers exhibit complete snow coverage from November to April, with snowline retreat starting at 3900 m.a.s.l and reaching a peak of 5113 ± 57 m.a.s.l by the ablation period's end. Despite similar performance across most methods, disparities emerge with smile CART and naïve Byes approaches. This is because of detailed sub pixel categorization of glacial features and lead to over/under estimation. The estimated uncertainties are confined within ±57 m. The regional SLA is notably higher in the hydrological year 2019-2020 (5098 meters). These insights into seasonal snow extents and snowline altitudes offer valuable data for reconstructing the mass balance and hydrological budget of the Parbati River basin.

Acknowledgements

The corresponding authors would like to thank the inform of Seed Grant support in phase II provided by VRSEC, Siddhartha Academy of General and Technical Education, Vijayawada. Special thanks to Dr. M. Ravichandran, Secretary Ministry of Earth Sciences, Prof. Helgi Bjornsson, University of Iceland, Dr Thamban Meloth, Director NCPOR and DST Center of Excellence, University of Sikkim for the support being provided for technical support and completion of the work. Sincere thanks from all the authors for the data support being provided by Tandem X Team, DLR, USGS, BBMB Chandigarh, Planet team and RGI Inventory.

Author(s) Contributions

All the authors contributed to the study conception, design and execution. Material preparation, data collection and analysis were performed by Dr. Vinay Kumar Gaddam, Sindhura Raavi, Aishwarya Ray and Sai Prasad. The first draft of the manuscript was written by Vinay Kumar Gaddam, corrections and technical suggestions were given by Dr. Rakesh Ranjan and Dr Gangadhar Kancherla. All authors read and approved the final manuscript.

References

[1] Tang, Z., Wang, X., Deng, G., Wang, X., Jiang, Z. and Sang, G., (2020). Spatiotemporal Variation of Snowline Altitude at the End of Melting Season Across High Mountain Asia, Using MODIS Snow Cover Product. *Advances in Space Research*, Vol. 66(11), 2629-2645.

[2] Singh, A. T., Rahaman, W., Sharma, P., Laluraj, C. M., Patel, L. K., Pratap, B., Gaddam, V. K. and Thamban, M., (2019). Moisture Sources for Precipitation and Hydrograph Components of the Sutri Dhaka Glacier Basin, Western Himalayas. *Water*, Vol. 11(11). https://doi.org/10.3390/w11112242.

[3] Cuffey, K. M. and Paterson, W. S. B., (2010). The Physics of Glaciers. Academic Press.

[4] Bolch, T., Kulkarni, A., Kääb, A., Huggel, C., Paul, F., Cogley, J. G., Frey, H., Kargel, J. S., Fujita, K., Scheel, M., Bajracharya, S. and Stoffel, M., (2012). The State and Fate of Himalayan Glaciers. *Science*, Vol. 336(6079), 310-314. https://doi.org/10.1126/science.1215828.

[5] Lei, L., Zeng, Z. and Zhang, B., (2012). Method for Detecting Snow Lines from MODIS Data and Assessment of Changes in the Nianqingtanglha Mountains of the Tibet Plateau. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 5(3), 769-776. https://doi.org/10.1109/JSTARS.2012.2200654.

[6] Tang, G., Li, S., Yang, M., Xu, Z., Liu, Y. and Gu, H., (2019). Streamflow Response to Snow Regime Shift Associated with Climate Variability in Four Mountain Watersheds in the US Great Basin. *Journal of Hydrology*, Vol. 573, 255-266. https://doi.org/10.1016/j.jhydrol.2019.03.021.

[7] Rastner, P., Prinz, R., Notarnicola, C., Nicholson, L., Sailer, R., Schwaizer, G. and Paul, F., (2019). On the Automated Mapping of Snow Cover on Glaciers and Calculation of Snow Line Altitudes from Multi-Temporal Landsat Data. *Remote Sensing*, Vol. 11(12). https://doi.org/10.3390/rs11121410.

[8] Racoviteanu, A. E., Rittger, K. and Armstrong, R., (2019). An Automated Approach for Estimating Snowline Altitudes in the Karakoram and Eastern Himalaya from Remote Sensing. *Frontiers in Earth Science*, Vol. 7. https://doi.org/10.3389/feart.2019.00220.

[9] Kulkarni, A. V., (1992). Mass balance of Himalayan Glaciers Using AAR and ELA Methods. *Journal of Glaciology*, Vol. 38(128), 101-104.

[10] Keeler, D. G., Rupper, S. and Schaefer, J. M., (2021). A First-Order Flexible ELA Model Based on Geomorphic Constraints. *MethodsX*, Vol. 8. https://doi.org/10.1016/j.mex.2020.101173.

[11] Millan, R., Mouginot, J., Rabatel, A., Jeong, S., Cusicanqui, D., Derkacheva, A. and Chekki, M., (2019). Mapping Surface Flow Velocity of Glaciers at Regional Scale Using a Multiple Sensors Approach. *Remote Sensing*, Vol. 11(21). https://doi.org/10.3390/rs11212498.

[12] Kurowski, L., (1890). Die Höhe der Schneegrenze: mit besonderer Berücksichtigung der Finsteraarhorn-Gruppe. na. (in portugese).

[13] Meierding, T. C., (1982). Late Pleistocene Glacial Equilibrium-line Altitudes in the Colorado Front Range: A Comparison of Methods. *Quaternary Research*, Vol. 18(3), 289-310.

[14] Osmaston, H., (2005). Estimates of Glacier Equilibrium Line Altitudes by the Area× Altitude, the Area× Altitude Balance Ratio and the Area× Altitude Balance Index methods and their validation. *Quaternary International*, Vol. 138, 22-31. https://doi.org/10.1016/j.quaint.2005.02.004.

[15] Žebre, M., Colucci, R. R., Giorgi, F., Glasser, N. F., Racoviteanu, A. E. and Del Gobbo, C., (2021). 200 Years of Equilibrium-Line Altitude Variability Across the European Alps (1901−2100). *Climate Dynamics*, Vol. 56, 1183-1201. https://doi.org/10.1007/s00382-020-05525-7.

[16] Gaddam, V. K., Kulkarni, A. V. and Gupta, A. K., (2020). Assessment of the Baspa Basin Glaciers Mass Budget Using Different Remote Sensing Methods and Modeling Techniques. *Geocarto International*, 35(3), 296-316. https://doi.org/10.1080/10106049.2018.1516247.

[17] Kulkarni, A. V. and Karyakarte, Y., (2014). Observed Changes in Himalayan Glaciers. *Current Science*, Vol. 106(2). 237-244. https://www.jstor.org/stable/24099804.

[18] Kulkarni, A. V., Singh, S. K., Mathur, P. and Mishra, V. D., (2006). Algorithm to Monitor Snow Cover Using AWiFS Data of RESOURCESAT-1 for the Himalayan Region. *International Journal of Remote Sensing*, Vol. 27(12), 2449-2457. https://doi.org/10.1080/01431160500497820

[19] Zhang, P., Qiao, Y., Schineider, M., Chang, J., Mutzner, R., Fluixá-Sanmartín, J., Yang, Z., Fu, R., Chen, X., Cai, L. and Lu, J., (2019) (2019). Using a Hierarchical Model Framework to Assess Climate Change and Hydropower Operation Impacts on the Habitat of an Imperilled Fish in the Jinsha River, China. *Science of the Total Environment*, Vol. 646, 1624-1638. https://doi.org/10.1016/j.scitotenv.2018.07.318.

[20] Mir, R. A., Jain, S. K., Saraf, A. K. and Goswami, A., (2014). Detection of changes in Glacier Mass Balance Using Satellite and Meteorological Data in Tirungkhad Basin Located in Western Himalaya. *Journal of the Indian Society of Remote Sensing*, Vol. 42, 91-105. https://doi.org/10.1007/s12524-013-0303-2.

[21] Paul, F., Winsvold, S. H., Kääb, A., Nagler, T. and Schwaizer, G., (2016). Glacier Remote Sensing Using Sentinel-2. Part II: Mapping Glacier Extents and Surface Facies, and Comparison to Landsat 8. *Remote Sensing*, Vol. 8(7). https://doi.org/10.3390/rs8070575.

[22] Rabatel, G., Gorretta, N. and Labbé, S., (2014). Getting Simultaneous Red and Near-Infrared Band Data from a Single Digital Camera for Plant Monitoring Applications: Theoretical and Practical Study. *Biosystems Engineering*, Vol. 117, 2-14. https://doi.org/10.1016/j.biosystemseng.2013.06.008.

[23] Shea, J. M., Menounos, B., Moore, R. D. and Tennant, C., (2013). An Approach to Derive Regional Snow Lines and Glacier Mass Change from MODIS Imagery, Western North America. *The Cryosphere*, Vol. 7(2), 667-680. https://doi.org/10.5194/tc-7-667-2013.

[24] Gaddam, V. K., Boddapati, R., Kumar, T., Kulkarni, A. V. and Bjornsson, H., (2022). Application of "OTSU"—An Image Segmentation Method for Differentiation of Snow and Ice Regions of Glaciers and Assessment of Mass Budget in Chandra Basin, Western Himalaya using Remote Sensing and GIS Techniques. *Environmental Monitoring and Assessment*, Vol. 194(5). https://doi.org/10.1007/s10661-022-09945-2.

[25] Otsu, Y., Furuhama, Y., Hoshina, S., & Ito, S. (1979). Propagation measurements and TV-reception tests with the Japanese broadcasting satellite for experimental purposes. *IEEE transactions on Broadcasting*, (4), 113-120.

[26] Chandrasekharan, A. and Ramsankaran, R. A. A. J., (2023). Reconstructing 32 Years (1989–2020) of Annual Glacier Surface Mass Balance in Chandra Basin, Western Himalayas, India. *Regional Environmental Change*, Vol. 23(4). https://doi.org/10.1007/s10113-023-02112-4.

[27] Singh, H., Varade, D., de Vries, M. V. W., Adhikari, K., Rawat, M., Awasthi, S. and Rawat, D., (2023). Assessment of Potential Present and Future Glacial Lake Outburst Flood Hazard in the Hunza Valley: A Case Study of Shisper and Mochowar Glacier. *Science of the Total Environment*, Vol. 868. https://doi.org/10.1016/j.scitotenv.2023.161717.