Using machine learning models for the assessment of permafrost occurrence in the Himalayas

Prashant Baral[1], Mohd Anul Haq[1]

[1]Computer Science and Engineering (GIS) Area, NIIT University, Neemrana, Rajasthan, India

Abstract

We have generated three maps showing distribution of probabilities of permafrost occurrence for a small section in the north-western Himalayan region. Three machine learning models, random forest, support vector machine and logistic regression, were used for determining the distribution of spatial probability of permafrost presence in the area under observation. Two topoclimatic variables, mean annual air temperature (MAAT) and potential incoming solar radiation (PISR), were used as predictor variables to train these three machine learning models. A training data set consisting of initiation line locations of 66 intact rock glaciers and 36 relict rock glaciers were used to train the three models. Results obtained from confusion matrices showed that random forest, support vector machine and logistic regression obtained an overall accuracy of 96%, 78% and 78% respectively, for the training data set. To evaluate the performance of the three models, a test data set consisting of initiation line locations of 38 intact rock glaciers and 22 relict rock glaciers were used to obtain the area under receiver operating characteristic curve (ROC curve) for all three models. The area under ROC curve was found to be 0.66, 0.78 and 0.77 for random forest, support vector machine and logistic regression respectively. The performance of the three models may have been affected by the small sizes of data set that were used to train them and test their performance. Different results for accuracies are also possible if parameters within the three models are attuned differently. Nevertheless, our study demonstrates that machine learning models can be essential to assess permafrost distribution in the Himalayan region.

Background

Understanding the spatial extent and distribution of Himalayan permafrost is necessary to anticipate and alleviate the impacts resulting from extensive thaw of Himalayan permafrost (Gruber et al., 2017). Recent studies have focussed on the application of simple machine learning model such as logistic regression for the estimation of the distributed probability of the occurrence of permafrost (Haq and Baral, 2019; Baral et al., 2019). In this study we have attempted to apply and compare the performance of three machine learning models, logistic regression (Hosmer & Lemeshow, 2013), support vector machine (Vapnik, 1982) and random forest (Brieman, 2001) for evaluating the spatially dispersed probability of permafrost presence in north-western Himalayan region. A detail description of the rock glacier data, MAAT and PISR used in this study can be found in Baral et al. (2019). All three machine learning models were implemented using Scikit-learn package in Python (Pedregosa et al., 2011). Accuracy of the models were evaluated using confusion matrices and area under ROC curve.

Study area

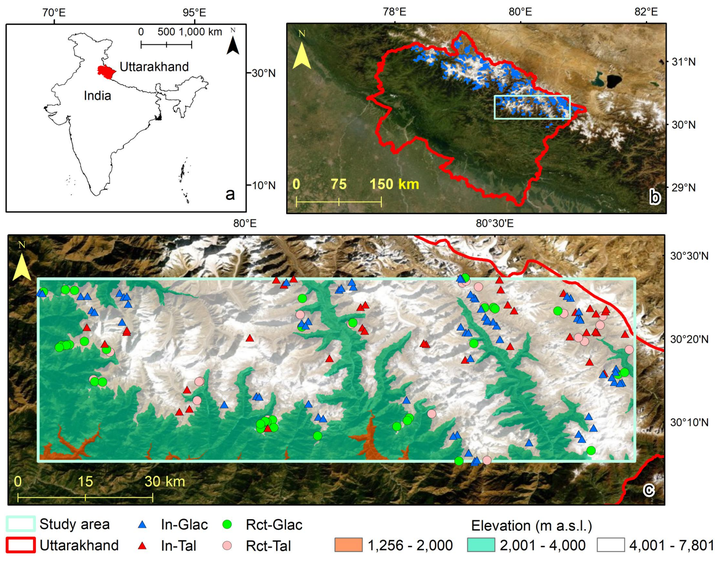

Figure 1. Study area (a) Map of India (b) Uttarakhand with rectangular section over Basemap imagery from ArcMap (ESRI, 2015) showing study area. (c) Study area with points indicating initiation line locations of four types of rock glaciers. Elevation of the study area is indicated using coloured legends.

Methods

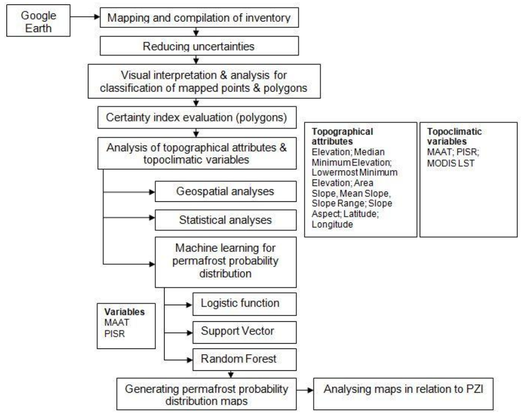

Figure 2 . Schematic representation of methods applied in Baral et al. (2019).

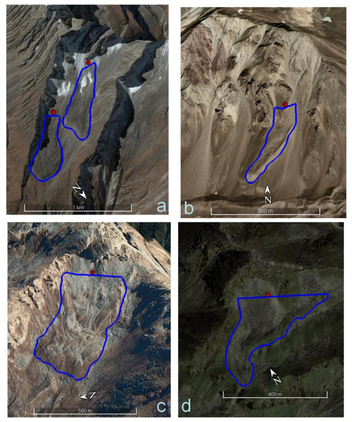

Figure 3 . Four classes of rock glaciers with initiation points. (a) In-Glac (b) In-Tal (c) Rct-Glac (d) Rct-Tal.

Results

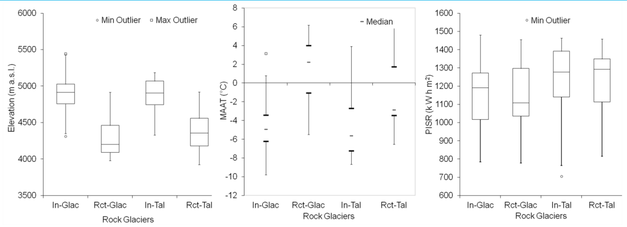

Figure 4. Box plots showing distribution of (a) elevation, (b) MAAT and (c) PISR of four types of rock glaciers located within the study area. Circles represent maximum and minimum outliers.

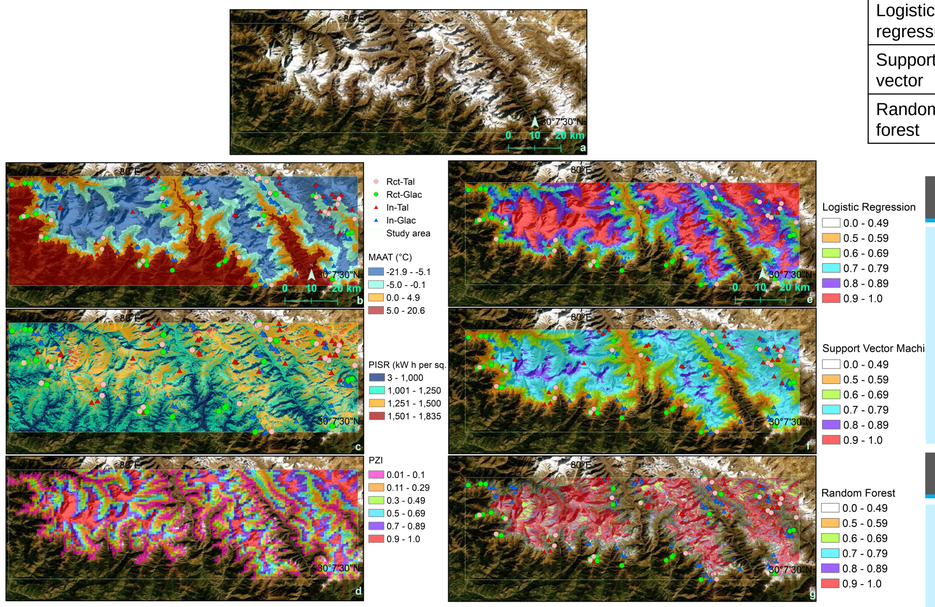

Figure 5. (a) Study area shown using a rectangular box over Basemap imagery from Arcmap. Distribution of MAAT (b) and PISR (c) within the study area. (d) PZI map (Gruber, 2012) within the study area. Permafrost probability distribution within the study area obtained using logistic regression function (e), support vector machine (f) and random forest classification (g). Values below 0.5 in probability distribution maps e, f and g is shown without any colour legend.

Results

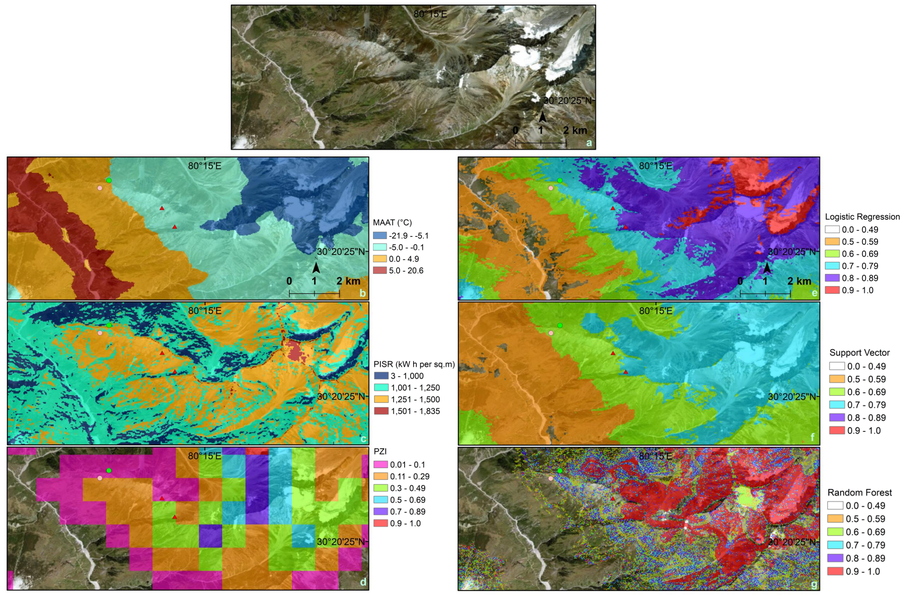

Figure 6. (a) A location within the study area. Distribution of MAAT (b) and PISR (c) within the location. PZI map (d) in the location. Permafrost probability distribution within the location obtained using logistic regression function (e), support vector (f) and random forest classification (g). Values below 0.5 in permafrost probability distribution maps e, f and g are shown without any colour legend.

Table 2. Accuracy assessment of three machine learning algorithms and areal distribution of permafrost probability obtained from three algorithms.

Machine learning algorithms	Overall acuracy (training data set)	Area under ROC curve (test data set)	Area with probability >0.5 (in km²)	Area with probability <0.5 (in sq. km²)
Logistic regression	0.78	0.77	3332.093	1441.107
Support vector	0.78	0.78	3413.178	1360.021
Random forest	0.96	0.66	2542.574	2230.625

Conclusion

Machine learning models can be essential for estimating the spatial extent and distribution of probability of existence of permafrost in remote Himalayan terrain. Due to small data size, random forest could have suffered from overfitting. Although area under ROC curve provides reliable measurement of accuracy of the performance of the machine learning models, field validations using ground surface and near surface air temperature measurements would be crucial to justify the observations from these models. Without proper field verifications, the results from models are prone to several uncertainties arising from remote sensing observations. WorldClim 2.0 data used in this study present the conditions of permafrost existence for the years 1970-2000.

References

Baral, P., Haq, M. A., & Yaragal, S. (2019). Assessment of rock glaciers and permafrost distribution in Uttarakhand, India. Permafrost and Periglacial Processes.
Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
ESRI (Environmental Systems Research Institute). (2015). ArcGIS 10.3. 1 for desktop.
Gruber, S. (2012). Derivation and analysis of a high-resolution estimate of global permafrost zonation. The Cryosphere, 6(1), 221.
Gruber, S., Fleiner, R., Guegan, E., Panday, P., Schmid, M. O., Stumm, D., ... & Zhao, L. (2017). Inferring permafrost and permafrost thaw in the mountains of the Hindu Kush Himalaya region.
Haq, M. A., & Baral, P. (2019). Study of permafrost distribution in Sikkim Himalayas using Sentinel-2 satellite images and logistic regression modelling. Geomorphology, 333, 123-136.
Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.
Vapnik, V. (1982). Estimation of Dependences Based on Empirical Data. Springer Verlag.