Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning approaches to predict Soil Organic Carbon in the North-western Himalayan Forests

Tanay Barman
Indian Council of Forestry Research and Education
Ankur Sharma
ankursharma.fri@gmail.com

Forest Research Institute
Renu Rawal
Soban Singh Jeena University
Gaurav Mishra
Indian Council of Forestry Research and Education
Divesh Pangtey
Indian Council of Forestry Research and Education
Azamat Suleymanov
Ufa University of Science and Technology

Research Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11193577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning approaches to predict Soil Organic Carbon in the North-western Himalayan Forests

Tanay Barman[1], Ankur Sharma [2*], Renu Rawal[3], Gaurav Mishra[1**], Divesh Pangtey[1], and Azamat Suleymanov[4]

[1] Centre of Excellence on Sustainable Land Management, Indian Council of Forestry Research and Education, Dehradun-248006, Uttarakhand, India

[2] Forest Ecology & Climate Change Division, ICFRE-Forest Research Institute, Dehradun-248006, Uttarakhand, India

[3] Department of Botany, S.S.J. University, Almora, Pt. B.D. Pandey Campus, Bageshwar, Uttarakhand 263601, India

[4] Department of Geodesy, Cartography and Geographic Information Systems, Ufa University of Science and Technology, 450076, Ufa, Russia

Corresponding Authors: *ankursharma.fri@gmail.com, **mishragaurav213@gmail.com

Abstract

Soil organic carbon (SOC) is a fundamental indicator of soil quality and ecosystem functioning, particularly in forest ecosystems. Accurate spatial assessment of SOC is essential for understanding its distribution and supporting sustainable forest management. This study aimed to predict and map SOC distribution across the moist temperate *Quercus leucotrichophora* forests of Himachal Pradesh, India, using machine learning (ML) techniques and to identify the major environmental drivers of its spatial variability. A total of 100 composite soil samples (0-15 cm depth) were collected from seven districts, and SOC concentration was determined using the Walkley-Black method. A digital soil mapping framework based on the SCORPAN model incorporated 49 environmental covariates derived from Landsat 8 imagery, SRTM elevation data and WorldClim bioclimatic variables. Following multicollinearity assessment using the variance inflation factor, 28 covariates were retained for model development. Four ML algorithms, namely Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM) and Cubist, were evaluated using R^2, RMSE, MAE and regression slope. Measured SOC concentrations ranged from 1.21% to 9.69%, with comparatively higher values observed at selected sites in Chamba district. RF demonstrated superior predictive performance (R^2 = 0.85; RMSE = 0.83%), outperforming GLM, SVM and Cubist. Variable importance analysis highlighted elevation, multi-resolution index of valley bottom flatness (MRVBF), slope and topographic ruggedness index (TRI) as dominant predictors, emphasising the influence of terrain on SOC variability. The RF-based SOC predictions ranged from 2.82% to 7.71%, while uncertainty was comparatively higher in steep and topographically complex landscapes than in valley floors. Although the outputs were generated at 30 m grid resolution, their effective spatial detail remains constrained by the native resolution of the input covariates. Overall, the findings demonstrate the potential of RF-based digital soil mapping for characterising SOC patterns in heterogeneous Himalayan Forest landscapes and provide a spatial baseline for prioritising forest conservation and soil management interventions.

Keywords: Digital Soil Mapping, Machine Learning, Soil Organic Carbon, Remote Sensing and GIS

1. Introduction

A key component of greenhouse gas dynamics, soil organic carbon (SOC) can behave as a reservoir based on the environment and soil management techniques. The carbon dynamics of soil, the biggest carbon storage system in terrestrial ecosystems, have attracted the attention of scientists worldwide (Schillaci et al. 2019 and Yigini et al. 2016), especially in Indian Himalayan Region (IHR) due to ecological significance (Mishra et al., 2023). As a result, the demand for accurate SOC data is rising globally. Soil maps with high resolution offer the most thorough depiction of the distribution of SOC. It is difficult to create these maps concurrently on a global scale, and SOC data needs to be updated often with data particular to each location. Soil mapping has historically depended on ground-based surveys, which have constraints in precisely capturing the fine-grained spatial distribution of soil parameters (Jena et al. 2022 and Moharana et al. 2022). For this reason, creating a sophisticated and trustworthy approach is crucial to forecasting soil characteristics across various soil types and geographical areas.

Digital soil mapping (DSM), which combines machine learning (ML) methods with remote sensing technologies including hyperspectral, multispectral, and radar imaging, has become a very successful strategy in this respect (Mohrana et al., 2026). For region-specific soil assessments, DSM is highly sought worldwide (Arrouays et al. 2014; Dash et al. 2022; Dharumarajan et al. 2019). It is also essential to accomplishing sustainability objectives, such as land-use planning. Artificial neural networks, decision trees, linear regression models, multivariate adaptive regression splines, regression trees, and support vector machines are some of the key machine learning techniques used in DSM (Pahlavan-Rad and Akbarimoghaddam 2018). These methods have the advantage of being able to create statistical relationships between soil characteristics and environmental factors, satellite images, topography features like digital elevation models, and more (Rossiter 2018). The accuracy and efficiency of DSM are better than those of conventional soil mapping techniques. As a result, investigations examining the regional variability of soil properties based on DSM have increased significantly in recent years.

By combining soil measurements with environmental covariates to produce continuous spatial forecasts, DSM has become a cutting-edge substitute for traditional soil surveys. It offers quantitative, higher-resolution estimations of soil characteristics that are crucial for agricultural planning, land management, and climate research. Limited soil observations, uncertainty in legacy data, and the availability of appropriate variables continue to be obstacles to DSM applications. Recent research demonstrates the increasing significance of remote sensing and machine learning in enhancing the precision and geographical detail of soil prediction. As a result, DSM provides a potential scientific foundation for generating the spatially explicit soil data required for making evidence-based environmental decisions (Nenkam et al., 2024). Similarly, estimated SOC, pH, texture, and bulk density globally using a 250 m resolution DSM (Hengl et al. 2015). A vital platform for combining current and historical regional soil data into a single global framework is the GlobalSoilMap consortium. Applications such as disaster risk assessment, infrastructure development, biodiversity conservation, and policymaking all benefit greatly from this compiled data (Liu et al. 2020).

One of the world's most ecologically sensitive regions is Himalayan Mountain range in northern India, which is distinguished by its high endemism, unique climate, and abundant biodiversity (Mishra et al. 2020). One of the region's distinctive ecosystems, it includes the Lesser Himalayas, also known as Himachal, and the Great Himalayas. The Great Himalayas rise from the Shivalik ranges and move higher into the mid-Himalayas. The

region is known for its snow-capped mountains, rivers, and valleys. The rivers Beas, Ravi, and Chenab originate in Himachal. Notably, this region is home to the valuable trees *Quercus leucotrichophora* in the world, which has great cultural and traditional significance for both carbon sequestration, soil conservation and antioxidant properties. More than 27.73% of the banj oak is present in Himalayan region of Himachal Pradesh with respect to seven districts of Himachal i.e., Mandi, Kullu, Shimla, Solan, Sirmour, Chamba and Kangra.

However, widespread land-use changes, deforestation, and abandoned fallow areas are all contributing to the drop in SOC levels in the Lesser Himalayas and the Great Himalayas (Rawal et al., 2026). Assessing changes in the carbon pool across forest and agricultural environments requires an understanding of SOC dynamics (Singh et al., 2026). Compared to conventional grid-based mapping techniques, which are frequently linked to increased uncertainty, machine learning provides a potent way to estimate SOC at a landscape scale with less physical inputs. Given this, 100 number of SOC records from the soils of *Quercus leucotrichophora* forest in Himachal study were used in the current study. The study's objectives are to: (i) produce a digital SOC map using a variety of machine learning methods; (ii) evaluate the effectiveness of various models; and (iii) identify the main environmental factors affecting the distribution of SOC.

2. Materials and Methods

2.1. Study area

The northern Indian state Himachal Pradesh is in the Western Himalaya located between 32° 5' 3.1416" N-77° 34' 16.2012" E. Union Territory of Jammu & Kashmir and Ladakh are in the north west; Punjab, Haryana and Uttarakhand sharing borders to it in the south west, south and south east; Tibet in the north east. The state has 12 districts and covers an area of 55,673 square kilometres is a part of Indian Himalayan Region. Despite being one of the least urbanized states in the nation, Himachal Pradesh has had one of the fastest-growing economies in recent decades. The main pillars of the state's economy are tourism, hydropower, and agriculture. The state has tropical, temperate and alpine forest ecosystem. With notable variations based on altitude the southern and western regions of Himachal Pradesh experience higher temperatures than the northern and eastern regions.

The state's mean annual rainfall is approximately 1,251 millimetres, and its mean annual temperature is approximately 17.13°C. In the moist temperate high hill zones, the average temperature ranges from 15.4°C to 20.5°C, while in the higher altitude regions, it is often lower from south to north. Because of the varied altitude soil temperatures vary greatly throughout it which subsequently influence the SOC. The predominant soil types in Himachal Pradesh are red loamy, sub-mountain, mountain meadow, and brown hill soils. Altitude, vegetation, slope, and structure all affect the type of soil. Soils of Himachal Pradesh show wide variability from light yellowish-brown in low hills to dark brown in forested and high-altitude regions which reflects the variance in organic matter and parent material. Soil structure is weak to moderately developed, while high-elevation soils are shallow, coarse-textured and skeletal. Entisols and Inceptisols dominate the Himachal Pradesh with localized presence of Alfisols and Mollisols in valleys and mid-hill zones (Rao et al. 2024).

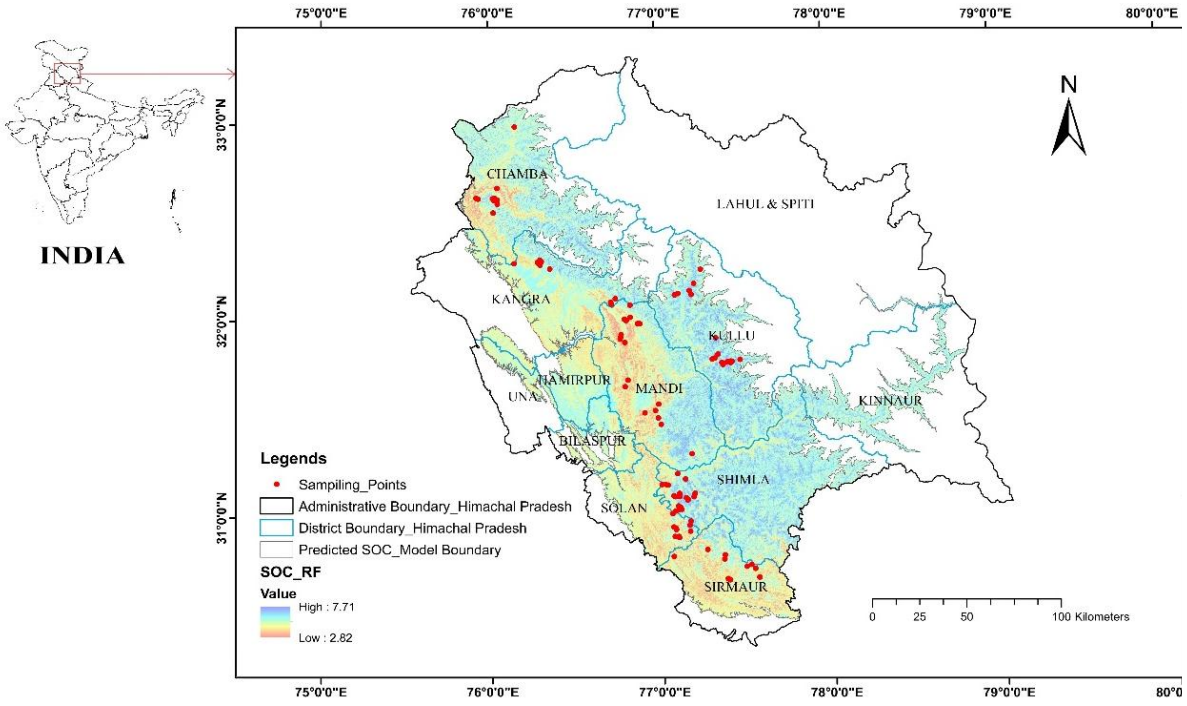


Fig. 1 Map of the study area showing sampling points

2.2. Soil Sampling and Analysis

Using a V-notch and systematic randomized sampling soil samples were collected from each studied plot of 50 × 50 m. The soil profile at each point was cored to a depth of 30 cm to characterise the profile; the surface 0-15 cm layer was retained for compositing and SOC determination, consistent with standard topsoil SOC sampling protocols. Five soil samples, four from the corners and one from the centre of each plot were collected, pooled and mixed properly to make a composite sample of 500 gram. The 100 sampling locations were distributed across the seven districts in proportion to the areal extent of *Quercus leucotrichophora* forest present in each, spanning the elevational and climatic gradient of the moist temperate zone; this density is broadly comparable to other regional digital soil mapping studies in the Indian Himalaya, although a sample-to-covariate ratio of ~3.6:1 (100 samples to 28 retained covariates) is acknowledged as a constraint on model stability in the Limitations section below. In total, 100 samples were collected from 100 locations of *Quercus leucotrichophora* forests. The air-dried soil samples were assessed for further tests and analysis. SOC was analyzed as described in Walkley and Black method (Walkley & Black, 1934).

2.3. Soil Mapping Framework

The soil mapping framework was prepared using SCORPAN (soil, climate, organisms, relief, parent material, age and spatial position) model, which utilizes environmental attributes such as soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a) and geographic position (n) to indirectly infer soil classification and characteristics of a specific location. This model establishes a quantitative (empirical) relationship between soil properties and the variables influencing spatial variability. The SCORPAN model is based on the principles of Jenny and Dokuchaev. It can incorporate various input data sources, including existing soil maps, point data

formats, or remotely sensed spectral characteristics. Environmental covariates are digitized and represented as geospatial raster data, processed using a geographic information system (GIS). This model simplifies the quantification of relationships between spatially explicit digital environmental variables and predicts soil attributes. Compared with traditional methods, it facilitates a more straightforward estimation of uncertainty.

2.4. Model Validation Approach

Four models (Random Forest, Generalized Linear Model, Support Vector Machine, and Cubist) were compared for SOC prediction by contrasting observed SOC with SOC predicted for samples not used in model fitting, using R^2, RMSE, MAE and the regression slope. The 100 soil samples were randomly partitioned into a calibration (training) set of (n = 80 samples; 80%) and an independent validation (testing) set of (n = 20 samples; 20%). The split was random and not stratified by district or elevation. During model development, model hyper-parameters were tuned on the calibration set only, using (10-fold cross-validation repeated 3 times), so that the validation samples were never used for model fitting or tuning; Random Forest was additionally assessed using out-of-bag (OOB) error. The accuracy measures reported in Table 2 and Fig. 7.

2.4.1 Regression Analysis: Observed versus Predicted

Predicted SOC values were shown against field-measured SOC readings for each model. The dependent variable, observed SOC, and the independent variable, projected SOC, were fitted using a straightforward linear regression. Model biasness and consistency with measured data are directly evaluated by this method. To visually evaluate how well the predictions match the observed data, a 1:1 reference line (slope = 1, intercept = 0) was added to each figure. Variations from this line show that the models either overestimated or underestimated.

2.4.2. Determination Coefficient (R^2)

The percentage of variance in observed SOC that each model could account for was measured using the coefficient of determination (R^2). Stronger agreement between predicted and actual SOC levels and improved model performance are indicated by higher R^2 values.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where yi is observed SOC, y^i is predicted SOC, and $\bar{y}$ is the mean observed SOC.

2.3.3. RMSE (Root Mean Square Error)

Root Mean Square Error (RMSE), which calculates the average magnitude of prediction errors in the same units as SOC, was used to quantify model prediction error. Higher prediction accuracy is shown by lower RMSE values.

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

RMSE provides an intuitive measure of how far predictions deviate from observed values.

2.4.4. MAE (Mean Absolute Error)

Mean Absolute Error (MAE) is the mean of the absolute differences between observed and predicted SOC, expressed in the same units as SOC (%). For a given set of residuals, MAE cannot exceed RMSE (MAE ≤ RMSE), and this relationship was used as a consistency check on the reported error metrics.

2.4.5. Regression Slope

Systematic bias was evaluated using the slope of the regression line between expected and observed SOC. Unbiased forecasts are shown by a slope around 1.0, but overestimation and underestimation are indicated by slopes larger or smaller than 1.

2.5 Selection of Environmental Covariates

Environmental covariates (ECs) associated with soil-forming processes were collected from multiple sources, including satellite-derived products, digital elevation models (DEMs), and climatic datasets. Initially, 49 environmental covariates were compiled and classified into seven categories based on the SCORPAN conceptual framework: climate, organisms, parent material, position, relief, remote sensing, and soil. Landsat 8 Operational Land Imager (OLI) imagery was used to derive spectral predictors because of its 30 m spatial resolution and 16-day revisit period (Moharana et al. 2022). Landsat 8 surface reflectance data were obtained from the United States Geological Survey (USGS; https://www.usgs.gov; accessed on 21 May 2024). Spectral bands (1–11) and derived indices, including the Normalized Difference Vegetation Index (NDVI), were considered as predictor variables for SOC modelling.

Topographic covariates were derived from the Shuttle Radar Topography Mission (SRTM) DEM (30 m resolution) to characterize terrain conditions influencing soil formation and SOC distribution. Sixteen terrain attributes were generated using SAGA GIS version 6.3.0: Aspect (Asp), Elevation (Elev), Slope (Slp), Relative Slope Position (RSP), Topographic Wetness Index (TWI), Multi-Resolution Valley Bottom Flatness Index (MRVBF), Valley Depth (VD), Convergence Index (CI), Channel Network Base Level (CNBL), Channel Network Distance (CND), LS-factor, Plan Curvature (PlaC), Profile Curvature (PrC), Total Catchment Area (TCA), Topographic Position Index (TPI), and Topographic Ruggedness Index (TRI). Additionally, all 19 bioclimatic variables (BIO1–BIO19) were obtained from the WorldClim version 2 dataset (http://www.worldclim.com/version2; accessed on 25 May 2024) to represent temperature and precipitation conditions.

All covariates were harmonized to a common 30 × 30 m prediction grid using average or bilinear resampling, as appropriate to the data type. Sampling coordinates were transformed into the WGS 84 Universal Transverse Mercator (UTM) coordinate system, Zone 46N, to ensure spatial consistency between soil observations and environmental predictors. All datasets were maintained in raster format for subsequent modelling.

It is important to distinguish the 30 m output grid resolution from the native spatial resolution of the input covariates. While Landsat 8 imagery and SRTM-derived terrain attributes provide information at 30 m resolution, other environmental predictors may originate from coarser-resolution datasets. In particular, the spatial resolution of vegetation products and WorldClim bioclimatic variables depends on their original source products. Resampling these datasets to 30 m facilitates spatial alignment but does not increase their original information content or resolve additional fine-scale variability. Consequently, although SOC predictions are generated on a 30 m grid, their effective spatial detail is influenced by the native resolution of the contributing predictors. The resulting maps are therefore more appropriate for landscape- to district-scale assessments, identifying priority

areas for soil carbon conservation and restoration, and supporting regional land-management planning rather than plot-level decisions without field verification.

To reduce predictor redundancy and potential multicollinearity, Variance Inflation Factor (VIF) analysis was performed using IBM SPSS Statistics version 20.0. VIF quantifies the inflation in the variance of regression coefficients caused by correlations among explanatory variables. An iterative backward-elimination procedure was applied, removing the variable with the highest VIF at each step and recalculating VIF values until all retained variables had values below 5. Based on this criterion, 36 environmental covariates were retained from the initial 49 predictors (Table 1) for subsequent SOC prediction modelling.

Table 1 Different environmental covariates used in the models for prediction of SOC

Soil forming factor	**Predictor**	**Abbreviation**	**Resolution**
Relief	Elevation (m)	Elev	90 m
	Slope	Slp	90 m
	Aspect	Asp	90 m
	Topographic Positioning Index	TPI	90 m
	Topographic Ruggedness Index	TRI	90 m
	Topographic Wetness Index	TWI	90 m
	LS-factor	LSF	90 m
	Channel Network Base Level	CNBL	90 m
	Multi-resolution Index of Valley Bottom Flatness	MRVBF	90 m
	Multi-resolution Ridge Top Flatness	MRRTF	90 m
	Relative Slope Position	RSP	90 m
	Valley Depth	VD	90 m
	Plan Curvature	PlaC	90 m
	Profile Curvature	PrC	90 m
	Total Catchment Area	TCA	90 m
Vegetation	Normalized Difference Vegetation Index	NDVI	250 m _16 days
	Enhanced Vegetation Index	EVI	250 m _16 days
Climate	Annual Mean Temperature (◦C)	BIO1	1 km
	Mean Diurnal Range (Mean of monthly (max temp - min temp)) (◦C)	BIO2	1 km
	Isothermality (Bio_2/Bio_7) (* 100)	BIO3	1 km
	Temperature Seasonality (standard deviation *100) (◦C)	BIO4	1 km
	Max Temperature of Warmest Month (◦C)	BIO5	1 km
	Min Temperature of Coldest Month (◦C)	BIO6	1 km

Soil forming factor	Predictor	Abbreviation	Resolution
	Temperature Annual Range (Bio_5-Bio_6) (◦C)	BIO7	1 km
	Mean Temperature of Wettest Quarter (◦C)	BIO8	1 km
	Mean Temperature of Driest Quarter (◦C)	BIO9	1 km
	Mean Temperature of Warmest Quarter (◦C)	BIO10	1 km
	Mean Temperature of Coldest Quarter (◦C)	BIO11	1 km
	Annual Precipitation (mm)	BIO12	1 km
	Precipitation of Wettest Month (mm)	BIO13	1 km
	Precipitation of Driest Month (mm)	BIO14	1 km
	Precipitation Seasonality (Coefficient of Variation) (mm)	BIO15	1 km
	Precipitation of Wettest Quarter (mm)	BIO16	1 km
	Precipitation of Driest Quarter (mm)	BIO17	1 km
	Precipitation of Warmest Quarter (mm)	BIO18	1 km
	Precipitation of Coldest Quarter (mm)	BIO19	1 km

2.6 Machine Learning Techniques

In this study, SOC content was predicted using four ensemble models: Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM), and Cubist. Several learners and algorithms are integrated to create these ensemble models. The ensemble learning/method, which combines several algorithms, created the ensemble model. R 4.1.2 software was used to do the corresponding trend analysis (https://www.R-project.org/), and ArcGIS 10.3 software was used to replicate the maps.

Random Forest (RF)

An effective ensemble learning method is random forest, which is an extension of the regression tree model. It uses two levels of randomization for every tree in the ensemble, combining several regression and classification trees (Breiman 2001). Because of its capacity to reduce model overfitting, increase prediction accuracy, and efficiently manage missing data, this model has become very popular. Its ability to operate with both quantitative and categorical data makes it flexible (Mishra et al., 2026). The following important elements affect how well the RF model performs: i) number of trees (n-tree): this parameter indicates how many trees are in the ensemble and is crucial to the model's performance. The number of predictors used for tree fitting (mtry) controls the selection of predictors taken into consideration at each split in the tree-building process. ii) minimum number of samples at terminal nodes (n-min): figuring out the minimum number of samples needed in terminal nodes of individual trees is essential for controlling model complexity. By examining how frequently they are used in nodes during tree development, the model finds important predictors. The mean square error (MSEOOB), which is calculated using the aggregated out-of-bag (OOB) predictions, is used to assess the performance of the RF model.

$$MSE_{OOB} \quad n^{-1} \sum_{i=1}^{n} \left(\text{zi} - \widehat{zi}^{OOB} \right)^2$$

Where n = number of observations, zi = average prediction of the ith observation, and ẑi^{OOB} = average prediction for the ith observation derived from all trees in RF ensemble for which the observation was out-of-bag (OOB).

Support Vector Machine (SVM)

SVM is a reliable technique that uses support vectors (seen data points) to try to determine a maximal margin, also known as the margin of tolerance. The margin, which stands for the distance from the decision surface, facilitates linear data fitting or separation. According to (Kovačević et al. 2010), SVM is well known for its ability to generalize effectively to unknown data, which increases the application of its results. SVM uses kernel functions to map vectors into a high-dimensional space to address non-linear problems. As can be seen in the SVM equation, SVM creates an ideal hyperplane within this space to efficiently separate classes or forecast target variables.

$$f(x) = [w \bullet x] + b$$

In the equation, [w • x] represents the dot product between a weighting vector w and an input vector x, while 'b' signifies the bias term.

Generalized Linear Model

A statistical framework known as a Generalized Linear Model expands on conventional linear regression to account for response variables with distinct distributions, such as normal, Poisson, or binomial. It is composed of three parts: a link function that connects the response mean to the linear predictor, a systematic component that represents a linear combination of predictor variables, and a random component that specifies the distribution of the dependent variable. GLMs are very flexible for real-world applications in domains like healthcare, finance, and the social sciences since they are frequently used to model a variety of data types, such as continuous variables, count data (Poisson regression), and binary outcomes (log istic regression). In this study, the GLM was fitted to SOC (%) on its original scale with Gaussian error distribution and identity link function, no bounds were imposed on the predicted values.

Cubist

In recent years, digital soil mappers have been increasingly interested in Cubist, a rule-based system (Miller et al. 2015). The following is a summary of its methodology: First, Cubist uses a collection of given variables to build a tree structure. This tree illustrates many paths for forecasting. By improving training, it condenses the tree's routes to produce rules. Every rule outline particular circumstance under which it can be applied. To forecast the target variable under its specified conditions, each rule uses an MLR model. These models essentially depict how input variables and the target variable relate to one another in different situations. A collection of MLR models, each designed to predict values under certain circumstances, are represented by the tree's terminal nodes. Cubist improves prediction accuracy by using ensemble learning supported by boosting. Models developed using several iterations of the same method are combined in ensemble learning, which usually produces better prediction performance than individual models. Regarding the precise representations of

the Cubist linear models, they may be shown as follows:

$$\zeta par = (1 - a) \times \zeta(p) + a \times \zeta(c)$$

Here, ζ(c) represents the prediction made by the current model, while ζ(p) represents the prediction made by the parent model located higher up in the tree.

3. Results

3.1. Soil Organic Carbon

Total SOC ranged from 1.21 to 9.69% across the 100 survey sites in the seven districts (namely Kullu, Mandi, Kangra, Shimla, Sirmour, Solan and Chamba). SOC values fell mostly within a narrower 1.21 - 4.98% band across six of the seven districts, with markedly higher values recorded at a small number of sites in Chamba district (up to 9.69%; see below). In district Kullu, SOC ranged between 1.44 - 4.91 %. Maximum SOC was recorded in the Great Himalayan National Park V population (4.91 %), followed by Great Himalayan National Park II (4.23 %), Great Himalayan National Park I (3.99 %), Great Himalayan National Park IV (3.61 %), and Sainj (3.41 %) populations. The remaining populations had < 3.40 % SOC. The minimum was recorded in Allain Duhangan Hydro Power Project (1.44 %) population. In Mandi district, SOC ranged between 1.57 - 4.54 %. Maximum SOC was recorded in the Badu population (4.54 %), followed by Nag Nala (4.52 %), Makriri (3.87 %), Chail Chowk (3.55 %), and Chaprot I (2.79 %) populations. The remaining populations had < 2.78 % SOC. The minimum was recorded in Tindunal (1.57 %) population. In Kangra district, SOC ranged between 1.97 - 4.38 %. Maximum SOC was recorded in the Dharamkot population (4.38 %), followed by Hanuman Ka Tibba (4.38 %), Macleodganj Cantonment (4.26 %), Bhagsu Nag (3.83 %), and Macleodganj II (3.39 %) populations. The remaining populations had < 3.38 % SOC. The minimum was recorded in Naddi (1.97 %) population. In Shimla district, SOC ranged between 1.37 - 4.52 %. Maximum SOC was recorded in the Shoghi II population (4.52 %), followed by Shoghi III (4.23 %), Jakhu I (4.21 %), Jakhu II (4.19 %), and Navbahar (4.17 %) populations. The remaining populations had < 4.16 % SOC. The minimum was recorded in Ghanahatti (1.37 %) population. In Sirmour district, SOC ranged between 2.11 - 4.17 %. Maximum SOC was recorded in the Nohradhar II population (4.17 %), followed by Sangrah II (3.85 %), Nohradhar I (3.71 %), Haripurdhar II (3.54 %), and Sangrah I (3.32 %) populations. The remaining populations had < 3.31 % SOC. The minimum was recorded in Rajgarh (2.11 %) population. In Solan district, SOC ranged between 1.21 - 4.98 %. Maximum SOC was recorded in the Kandaghat I population (4.98 %), followed by Chail Wildlife Sanctuary (4.66 %), Jauna Ji (4.23 %), Bajdol (3.39 %), and Kandaghat II (3.25 %) populations. The remaining populations had < 3.24 % SOC. The minimum was recorded in Janedghat (1.21 %) population. In district Chamba, SOC ranged between 2.51 - 9.69 %. The maximum SOC was recorded in the Banikhet population (9.69 %), followed by KKWS-I (8.25 %), KKWS-V (8.12 %), KKWS-III (7.89 %), and Mankot (6.58 %) populations. Moderate SOC content was observed in Khajjiar-II (5.45 %), Hatli-Dramman (5.23 %), and Khajjiar-I (4.65 %) populations. The remaining populations had < 4.00% SOC. The minimum SOC was recorded in the Dumas-Tissa population (2.51 %).

Thus, the measured SOC distribution is strongly right-skewed: most sites in six of the seven districts fall within 1.21 - 4.98%, whereas only a few sites in Chamba district exceed this range (up to 9.69%). This distinction between the overall range (1.21 - 9.69%) and the range found in most districts (1.21 - 4.98%) is maintained throughout the interpretation of model performance.

3.2. Variable importance

Comparative relative variable significance plots in Figure 2 demonstrate that RF selected more variables than the SVM, Cubist, and GLM models. BIO 15 > Network Base Level > DEM > BIO 11 > Band 10 > BIO 8 > Band 11 > Shadow Max > BIO 4 > BIO 9 was the order of the most important parameters in the RF model for forecasting the SOC. Slp, TRI, and Band 4 were the most important factors for using a Cubist model to forecast the variance of the SOC. The SVM and GLM models' most crucial components were the AP and LS. For both the RF and Cubist models, the elevation and slope had the greatest impact on the SOC forecast, followed by the other topographical and vegetation parameters. Landslides and floods in the upper Ganga Valley region of north-western India cause soil sediments to be transported from higher elevations and deposited in lower-lying areas. It is worth noting that Boruta and the model-specific importance measures used here rank covariates by their predictive contribution within the fitted models; a high importance score indicates that a variable helped the model discriminate SOC values, but does not by itself establish that the variable is a causal driver of SOC formation or loss several of the top-ranked covariates (e.g. elevation, slope) are likely proxies for unmeasured processes such as microclimate, erosion history, and land management. The importance of elevation, MRVBF, slope and TRI are interpreted in terms of the processes they represent: elevation integrates temperature and moisture gradients that regulate litter input and decomposition; slope and TRI describe the susceptibility of a site to erosion and runoff and to shallow, well-drained soils; and MRVBF identifies flat, low-lying depositional positions where soil material and organic matter accumulate and where moisture is higher.

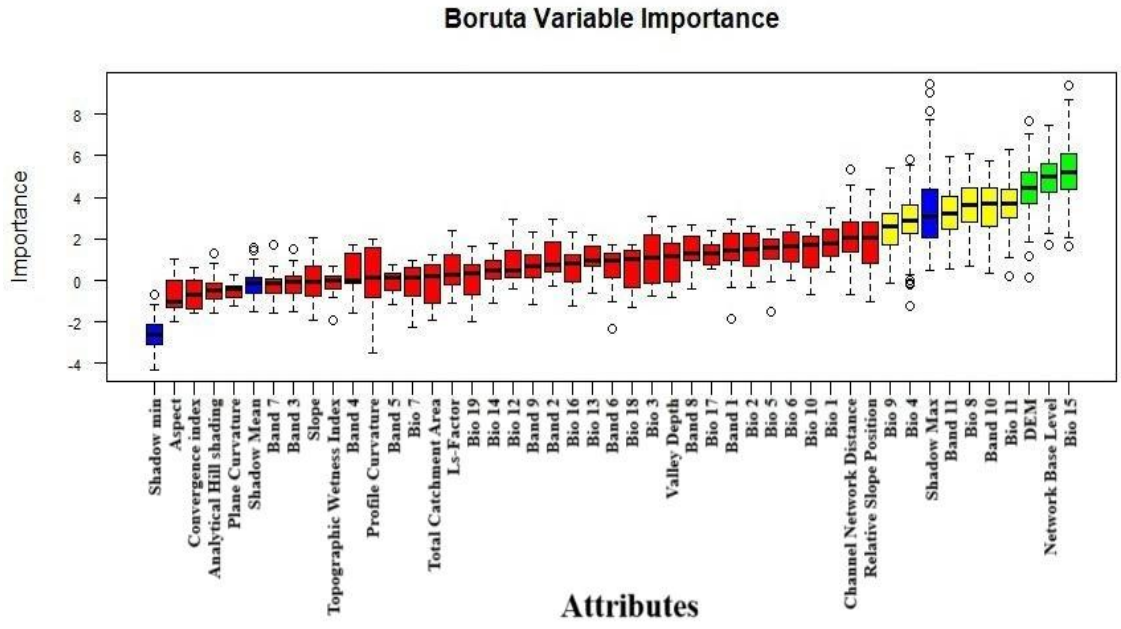


Fig. 2 Boruta variable importance plot illustrating the relative contribution of terrain, spectral, hydrological, and environmental covariates used in the RF model. Boxplots represent the distribution of importance scores across iterations, with confirmed important variables highlighted, marginal features shown in yellow, and rejected variables in red, indicating their respective influence on the model performance.

3.3. Mapping of SOC

The spatial predictions of the SOC provided by the different models, demonstrates notable differences in the projections. RF's predictions more closely matched the geographical distribution of the SOC we expected in the study region, even though the cross-validation accuracy scores of the four models were similar. The SVM forecast seems to underestimate the SOC in the study region. The prediction models' maps all showed both sudden and slow changes over the study area. Accordingly, the projected OC in the RF, SVM, Cubist, and GLM models varied from 2.82 to 7.71% (Fig. 4), 2.89 to 8.11% (Fig. 3), 0.0 to 14.06% (Fig. 5), and -576.75 to 132.91% (Fig. 6) respectively. The negative and >100% GLM values, and the Cubist upper value (14.06%) exceeding the highest value ever measured in the field (9.69%), are not physically plausible: a standard GLM with an identity link is unbounded, and Cubist's rule-based linear sub-models can extrapolate sharply beyond the range of the training data, so both models can generate physically impossible values when applied to covariate combinations that lie outside, or at the edge of, the envelope defined by the 100 calibration samples. These extreme values are best interpreted as a diagnostic of model extrapolation rather than as credible SOC estimates. Without independently verifying these predictions, it is hard to determine whether a model is the most accurate on its own; nevertheless, we selected RF as the most suitable model primarily on the basis of its quantitatively superior validation performance (highest R^2, lowest RMSE and MAE among the four models; Table 2) RF's tree-based structure also constrains predictions to within the range of the training data, avoiding the implausible extrapolation seen for GLM and Cubist. The visual plausibility of its spatial pattern, discussed next, was used only as a secondary, corroborating check rather than a selection criterion. With high values in the study area's North-western region which is primarily forested and covered in dense vegetation and low values in its western regions which contain soils that are heavily eroded and used for agricultural cultivation the spatial patterns of the SOC in all models were realistic. The R^2 of the RF model was the highest, consistent with the detailed quantitative comparison presented in Section 3.4 (Fig. 7). In all the calibration and validation sets, the RF model performed at its best R^2 = 0.85, (RMSE = 0.83% and Slope = 1.5)

Implementation of the GLM and Cubist predictions: SOC (%) was modelled on its original scale, without transformation of the response variable. The GLM was fitted with Gaussian error distribution and identity link function, and the Cubist model was fitted as a rule-based model with linear sub-models. Both models were fitted to the calibration data and then applied to the stack of 36 covariate raster's to obtain pixel-wise predictions over the entire mapped area, with no truncation of the output to the observed SOC range. The ranges quoted above (-576.75 to 132.91% for GLM and 0 - 14.06% for Cubist) therefore describe all pixels of the output maps and not predictions at the 100 sampled locations. Values below 0%, above 100% (GLM), or above the maximum measured SOC of 9.69% (Cubist) are physically implausible, are extrapolation artefacts rather than valid SOC estimates, and are flagged as such in the captions of Figs. 5 and 6; they were not used to interpret the spatial pattern of SOC.

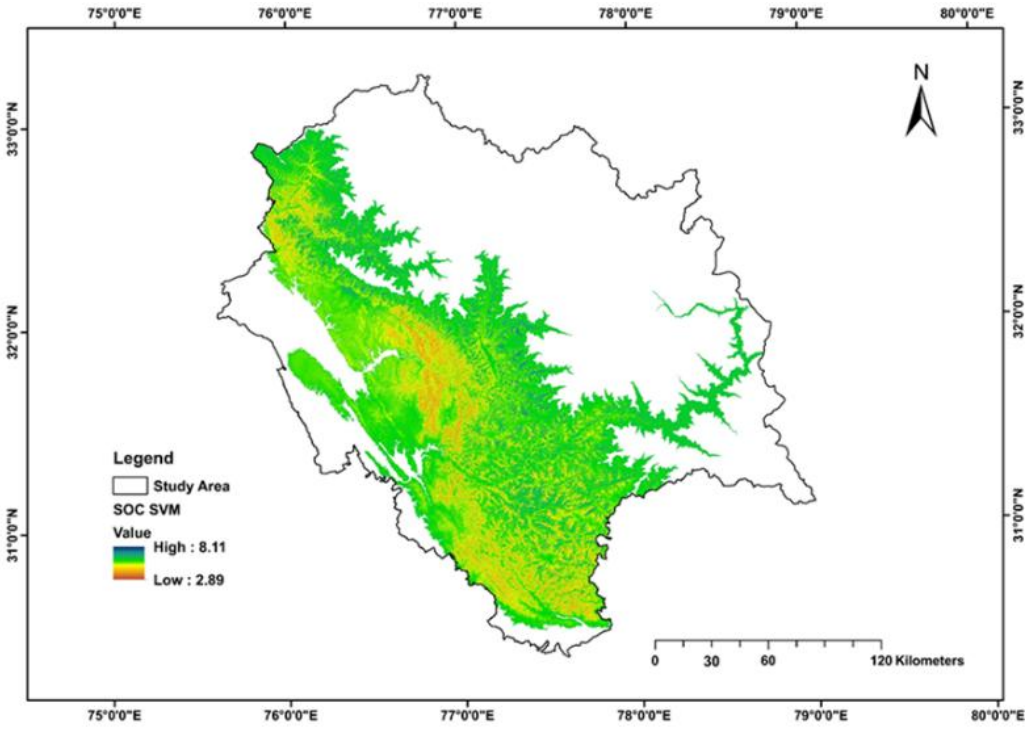


Figure. 3 Spatial distributions of SOC content across the study area in Himachal Pradesh, India, predicted using the SVM model. The map depicts SOC values ranging from ~2.89 to ~8.11, highlighting spatial variability in relation to topography and landscape heterogeneity within the study area boundary.

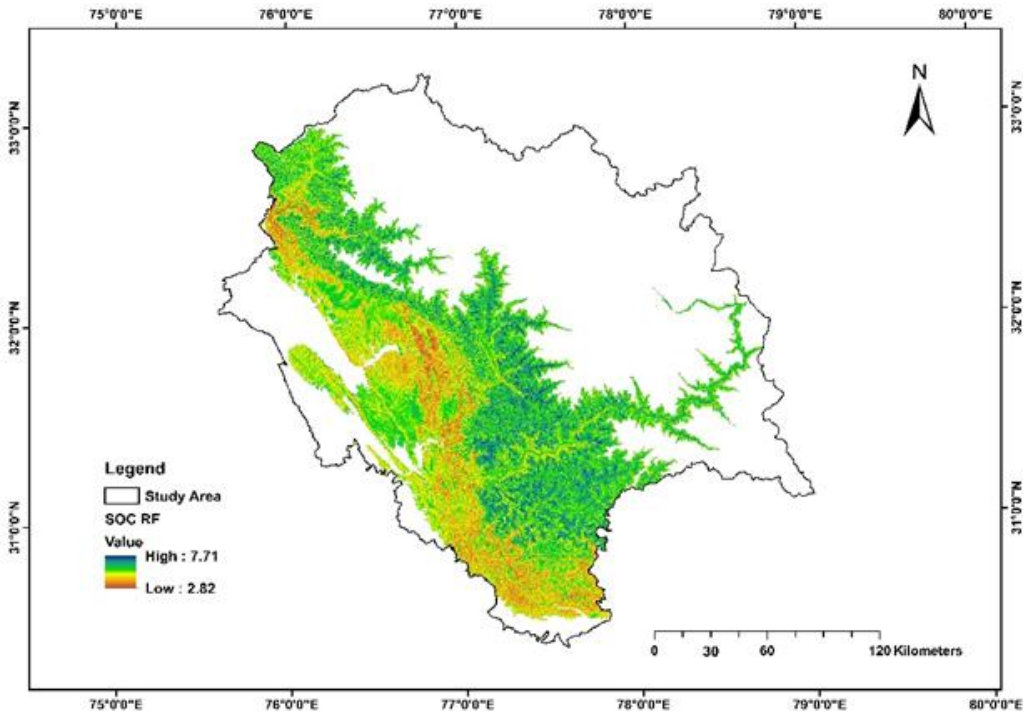


Figure. 4 Spatial distributions of SOC content across the study area in Himachal Pradesh, India, predicted using the RF model. The map depicts SOC values ranging from ~2.82 to ~7.71, highlighting spatial variability in relation to topography and landscape heterogeneity within the study area boundary.

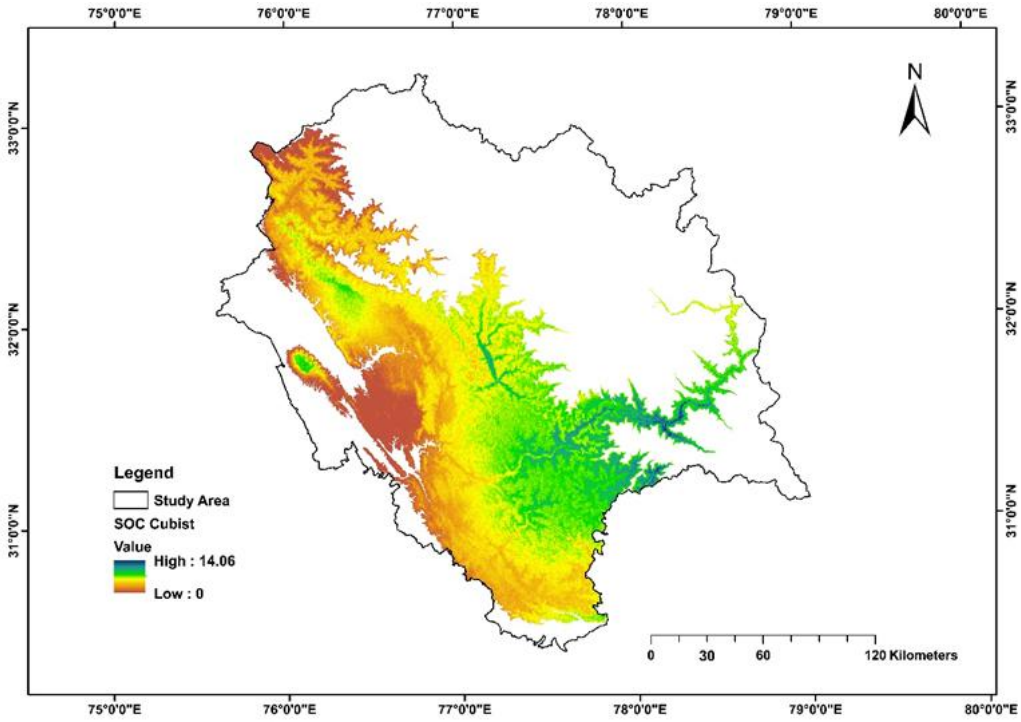


Figure. 5 Spatial distributions of SOC content across the study area in Himachal Pradesh, India, predicted using the Cubist model. The map depicts SOC values ranging from ~0 to ~14.06, highlighting spatial variability in relation to topography and landscape heterogeneity within the study area boundary. Values above ~9.69%, the maximum measured SOC, are physically implausible extrapolation artefacts and should not be read as valid SOC estimates.

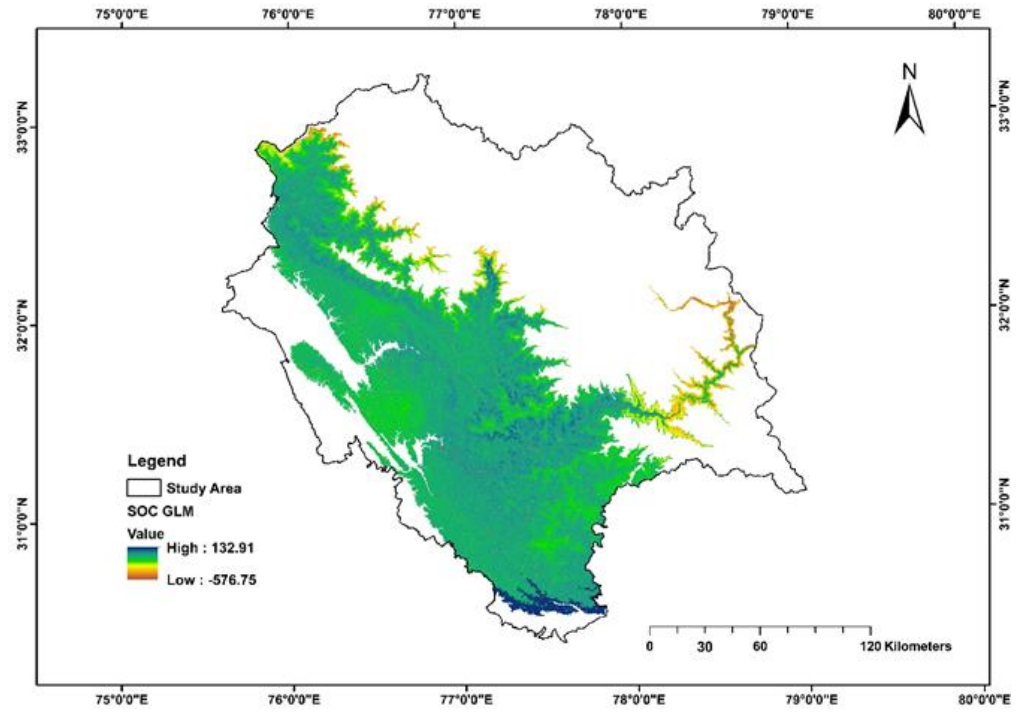


Figure. 6 Spatial distributions of SOC content across the study area in Himachal Pradesh, India, predicted using the GLM model. The map depicts SOC values ranging from -576.75 to ~132.91, highlighting spatial variability in relation to topography and landscape heterogeneity within the study area boundary. Negative

values and values >100% are physically impossible extrapolation artefacts; this map is shown for diagnostic purposes only and does not represent valid SOC estimates.

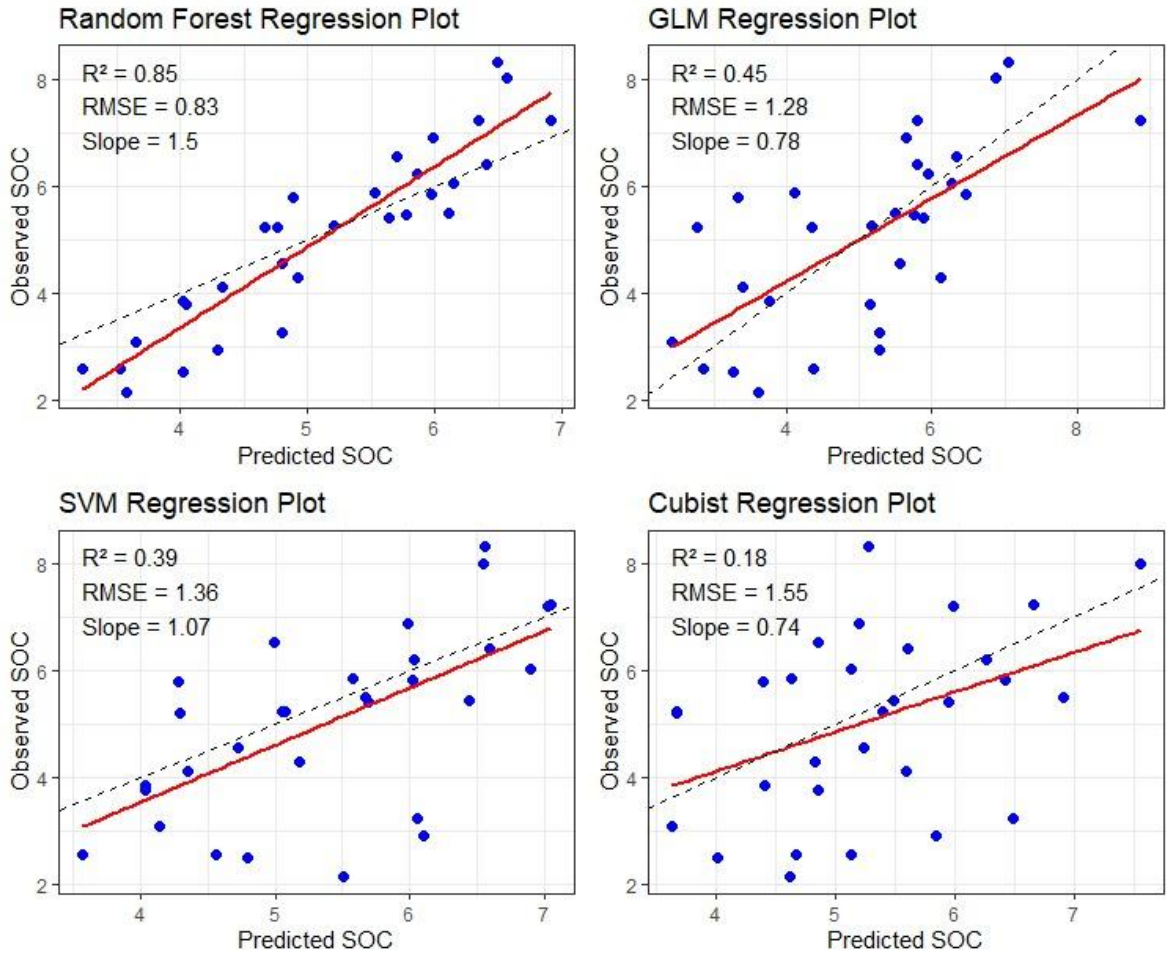


Fig. 7 Predicted vs. observed soil organic carbon by various models in the study area.

3.4. Accuracy of Models

The efficacy of four regression models to predict SOC varied significantly, according to the performance evaluation. With a high coefficient of determination (R^2 = 0.85), which indicates that the model accounted for 85% of the variation in observed SOC values, the RF model demonstrated the most robust predictive performance. Among the models, the model with the lowest RMSE value (0.83%) showed the least amount of prediction error. Although there was considerable overestimation at higher SOC levels, the fitted line's slope (1.50, i.e. >1) indicated that RF successfully tracked fluctuations over the SOC gradient; a slope above 1 in the observed-vs-predicted fit means predicted values increase more steeply than observed values across the gradient, a pattern often linked, for RF, to the relatively small number of high-SOC training samples (the small number of very high-SOC sites recorded in Chamba district) relative to the covariate space the model must generalise across this slope should therefore be interpreted jointly with R^2 and RMSE rather than in isolation. The over-steepening of the RF slope is a known artefact of tree-based bagging ensembles under skewed target distributions, which is why RMSE and R^2 remain the primary evaluative criteria here. Strong agreement between the expected and observed data was shown by the scatter points continued tight clustering around the 1:1 reference line.

The measured SOC range must be kept in mind when interpreting these metrics. Measured SOC spans 1.21 - 9.69% overall, but most sites lie within 1.21 - 4.98% and only a few Chamba sites exceed this. The R^2 and slope in Table 2 are computed over the full range and are therefore strongly influenced by the few high-SOC Chamba samples, which exert high leverage on the regression; for the majority of sites (1.21 - 4.98%), an RMSE of 0.83% is a substantial fraction of the observed range, and performance there may be lower than the overall R^2 of 0.85 suggests. Model performance should therefore not be generalised to high-SOC conditions that are represented by only a handful of samples.

With an R^2 score of 0.45 and an RMSE of 1.28%, the GLM model demonstrated a decent level of predictive power. The slope (0.78) showed persistent underestimate, particularly at higher SOC values, even if the model was able to represent the overall trend of SOC fluctuation. In comparison to the RF model, the scatter plot displayed a greater dispersion of points around the identity line, indicating GLM's diminished capacity to represent the nonlinear interactions that frequently define SOC spatial patterns.

The SVM model was like that from the GLM model, however, they were marginally worse. Its predictive ability was comparatively low, with an R^2 of 0.39 and an RMSE of 1.36%. The scatter distribution showed considerable variability and divergence from the identical line throughout both low and high SOC ranges, despite the slope (1.07) being near to 1, indicating proportionality between observed and anticipated SOC. The SVM regression line showed uneven predictive power, underscoring difficulties in capturing the intricate relationships between SOC and environmental factors.

Out of the four algorithms, the Cubist model performed the worst. Its RMSE (1.55%) was the highest and its R^2 value was just 0.18, indicating significant prediction error and little explanatory power. A consistent underestimating of SOC, especially in mid- to high-value circumstances, was further highlighted by the slope (0.74). Cubist's insufficient capacity to depict SOC dynamics for the given dataset was demonstrated by the scattered dots, which were widely distributed and showed no discernible clustering around the 1:1 reference line.

In terms of explanatory power, error reduction, and visual conformance with the 1:1 line, the RF model fared significantly better overall than the other methods. The gradual deterioration in model performance from RF to GLM and SVM, and ultimately Cubist, emphasizes how much model structure and the capacity to represent nonlinear behaviour affect SOC forecast accuracy shown in (Table 2).

Table 2: Model Validation and Comparison of performance of different models for prediction of soil organic carbon in the study area.

Model	R^2	RMSE, %	MAE, %	SLOPE
RF	0.85	0.83	1.54	1.5
GLM	0.45	1.28	2.09	0.78
SVM	0.39	1.36	1.53	1.07
Cubist	0.18	1.55	1.58	0.74

Abbreviations used: RMSE: Root Mean Square Error; R^2: Coefficient of determination; MAE: Mean absolute Error.

3.5. Uncertainty

There is clear regional diversity throughout the research area in the Random Forest-derived uncertainty map of SOC estimates (Fig 8). In valley floors and generally homogenous landscapes, where field observations are more numerous and environmental circumstances are more uniform, lower uncertainty levels (≈0.79 % SOC) are typically seen. These regions show increased SOC estimate reliability and model confidence. Uncertainty values are expressed in the same units as SOC (% SOC), and each value is the standard deviation of the predictions of the individual Random Forest trees at that pixel.

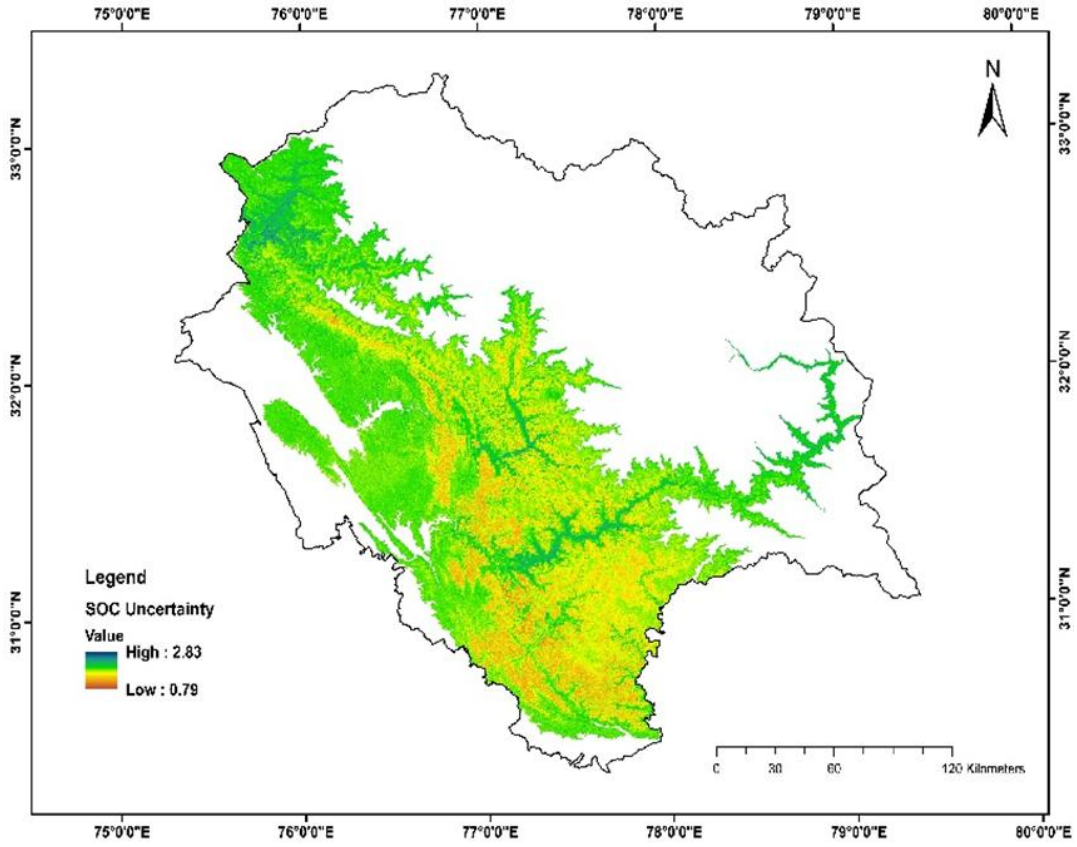


Figure. 8 Spatial distributions of Random Forest derived soil organic carbon prediction uncertainty, expressed as the standard deviation (in SOC %) of predictions across individual Random Forest trees across the study area.

On the other hand, hilly and structurally complicated areas tend to have greater uncertainty values (up to ≈2.83 % SOC). Strong soil-terrain interactions, diverse land cover, and steep elevation gradients are characteristics of these regions that raise forecast variability among individual Random Forest trees. Similar trends have been often documented in digital soil mapping investigations, where complicated topography and restricted accessibility frequently lead to uneven or scant soil sample, which increases model uncertainty.

The combined impact of data availability and environmental variation on SOC prediction performance is reflected in the observed uncertainty distribution. Because forecast reliability is quite poor, regions with higher uncertainty should be regarded cautiously. Practically speaking, the uncertainty map identifies priority zones for further soil sample and model improvement, making it a useful decision-support layer. The robustness and transparency of digital soil mapping outputs are improved by combining such uncertainty information with SOC projections, especially for land management and environmental assessment applications. In practical terms, a value of ≈2.83

% SOC means that individual trees disagree by roughly 2.8 percentage points, which is of the same order as the whole 1.21 - 4.98% range measured in most districts; pixels with such values should be treated as low-confidence predictions and prioritised for additional sampling, whereas pixels near ≈0.79 % SOC can be used with greater confidence.

It is important to be clear about what this uncertainty layer does, and does not, represent: the mapped values are the standard deviation of predictions across the RF ensemble's individual trees, so they capture the model's internal disagreement about a pixel's SOC value given the available covariates and training samples. They do not capture SOC laboratory measurement error, spatial positional error in the sample locations, or the additional uncertainty introduced by extrapolating into covariate combinations poorly represented in the 100 calibration samples (e.g. the very high-SOC conditions found in parts of Chamba district). Users of this SOC map should therefore treat the uncertainty layer as a useful, but partial, guide to reliability, and should exercise particular caution in areas of sparse sampling or unusual covariate combinations.

4. Discussion

The present study demonstrates that the heterogeneous landscape of the Western Himalayas requires modelling approaches capable of capturing complex nonlinear relationships between SOC and environmental covariates. The superior performance of the Random Forest model aligns with recent findings from mountainous regions globally, where elevation-driven microclimate variability and topographic gradients exert strong controls on carbon accumulation (Jena et al., 2023; Hengl et al., 2015; Suleymanov et al., 2025).

This suggests that forest structural richness and SOC may be positively correlated in this area, which is consistent with research showing that higher biomass and canopy cover frequently improve the retention of organic carbon in forest soils (Ražauskaitė et al., 2020). This defies the notion that tree density and SOC are directly positively correlated, indicating that other important variables in SOC dynamics may include disturbance, litter quality, or soil microbial activity (Bradford et al., 2021). The SOC is transported from higher altitudes, where concentrations are often lower, to lower elevations, where concentrations are generally greater, by these topography-driven erosional processes (Jena et al., 2022).

Topographic indices such as MRVBF, TRI, and LS-factor were particularly influential, highlighting the role of geomorphic processes in SOC redistribution. Steep slopes facilitate erosion and SOC loss, whereas valley bottoms act as depositional zones. This agrees with earlier studies demonstrating strong terrain–carbon linkages in Himalayan and Alpine systems (Pahlavan-Rad et al., 2020; Schillaci et al., 2019).

Elevation, MRVBF, slope and TRI plausibly relate to SOC through linked effects on soil moisture, erosion, deposition and organic matter accumulation. Elevation controls temperature and precipitation: cooler and moister conditions at higher elevations slow decomposition and favour litter accumulation, which promotes organic matter build-up. Slope and TRI describe how steep and dissected the terrain is; steep, rugged surfaces shed water, promote surface runoff and erosion of the organic-rich topsoil, and support shallow, coarse and well-drained soils that retain less moisture and organic matter. In contrast, gentle and smooth surfaces retain litter and soil. MRVBF identifies flat valley bottoms and concave positions where eroded soil and organic material are deposited and

where higher soil moisture slows decomposition, favouring the accumulation and preservation of organic matter. These interpretations are consistent with the terrain-carbon linkages, they describe likely associations rather than demonstrated causal mechanisms.

The study broadly consistent with recent digital soil mapping work elsewhere in the Himalaya. In the Kashmir Himalaya, Farooq et al. (2022) similarly found Random Forest to outperform a geostatistical kriging approach for SOC mapping in rugged mountain terrain. Using the same four learning algorithms applied here (RF, SVM, Cubist and a linear model) at a comparable 30 m resolution and 0 - 15 cm sampling depth in the Upper Brahmaputra Valley of Northeastern India, Kumar et al. (2023) likewise reported RF as the top performer, with rule-based and linear models trailing behind. Rengma et al. (2023), working in the mid-Himalayan terrain of Uttarakhand, and Kalambukattu et al. (2025), working elsewhere in the North-western Himalayan region covered by the present study, both identified elevation and other terrain attributes among the leading SOC predictors using RF-based digital soil mapping, reinforcing the topography SOC linkage reported here. Collectively, these studies support RF as a robust default choice for SOC mapping in topographically complex, data-limited mountain settings, while also indicating that the relative ranking of individual covariates varies with local land use, climate and sampling design, and should not be over-generalised beyond the specific landscape studied.

One of the key advantages of RF models over other machine learning models is that they evaluate the relative importance of the covariates in the model (Da Silva et al., 2016). Unlike Cubist methods, which keep the model intact by stepwise selection of only highly variable and soil variables, RF models do not eliminate them (Akpa et al., 2014). Pahlavan-Rad and Akbarimoghaddam, 2018 found that elevation and the CNBL were the most important factors for predicting soil characteristics in a study carried out in Himachal Pradesh.

In the upper Ganga Valley, low-elevation cultivable fields seem less assured than high-elevation cultivable plains. The difference in management practices between the more intensively farmed areas explains this (Jena et al 2022). It is possible that the variables included in these prediction models were required to sufficiently take into consideration the management factors influencing the spatial variability in the SOC content levels in these downstream-farmed areas. Because they are less often farmed, the high-elevation lands may be less affected by management and the larger influence of the terrain features. The equilibrium between carbon inputs and outputs in soils determines SOC concentrations; this equilibrium is influenced by several factors, including regional features like vegetation, topography, and environmental circumstances (Sahoo et al., 2019). Topography was the most important covariate affecting the SOC distribution in our analysis, per (Pahlavan-Rad et al., 2020).

This could be because of favourable microclimatic conditions, less disturbance, or increased litter build up (Brown and Lugo, 1982). However, even though Solan has the largest tree density, its low SOC indicates that the potential for carbon sequestration may be limited by saturation thresholds or site-specific soil conditions. Complex relationships between structural determinants, forest composition, and perhaps environmental factors are shown by the geographical distribution of SOC between districts. To further understand SOC dynamics in Himalayan oak forests, more multivariate studies combining climate, topography, soil texture, and land use history are required, as evidenced by the lack of a clear linear association between SOC.

Interestingly, SOC did not correlate directly with *Quercus leucotrichophora* density, indicating that vegetation structure alone cannot explain SOC variability. Rather, the interaction between litter quality, soil microbial

processes, moisture regimes, and past disturbance appears more important in western Himalayas (Rawat et al, 2021; Tiwari et al., 2023). This is consistent with emerging research emphasizing the dominant role of microclimatic buffering and soil biological activity in forested mountain systems (Mishra and Gorai, 2022; Bradford et al., 2021).

This study characterises SOC concentration (% by mass), not SOC stock (mass per unit area, typically expressed as Mg C ha^{-1}), because bulk density was not measured at the sampling sites. Accordingly, this study focuses on the spatial mapping of SOC concentration and does not directly estimate SOC stocks or carbon sequestration. SOC concentration and SOC stock can diverge, particularly across the wide range of slope positions and soil depths represented in this dataset, so the present maps should be used for identifying relative spatial patterns and priority areas for SOC conservation rather than for direct carbon-stock accounting; coupling these SOC concentration maps with bulk density and soil-depth measurements is a priority for future work before they are used in formal carbon accounting. SOC stock assessment requires additional measurements, in particular soil bulk density and sampling depth (and, in these stony, skeletal mountain soils, coarse fragment content), and this is identified as a requirement for future research.

Several limitations of this study should be kept in mind when interpreting the results. First, the ratio of soil samples to retained covariates (100 samples to 28 covariates, ~3.6:1) is modest for machine learning applications and is a plausible contributor to model instability, particularly for the more data-hungry Cubist and SVM models; expanding the sample size, especially in under-represented high-SOC settings such as Chamba district, would likely improve model robustness. Second, model performance was assessed using the calibration/validation partitioning, but no fully independent, spatially or temporally distinct validation dataset was available; future work should test these models against independent field data collected outside the original sampling campaign to obtain a less optimistic estimate of true predictive performance. Third, although all covariates were resampled to a common 30 m grid for mapping, several were derived from coarser native-resolution products, so the effective spatial resolution of the SOC maps in practice reflects this coarser envelope rather than a genuine 30 m level of detail. Finally, as noted above, the maps presented here describe SOC concentration rather than SOC stock. None of these limitations undermine the study's central finding that RF substantially outperformed the other three algorithms for SOC prediction in this landscape, but they should temper claims about the absolute accuracy and fine-scale spatial precision of the resulting maps.

The poor performance of the Cubist model highlights the limitations of rule-based approaches in steep, complex terrain where environmental conditions shift abruptly. RF, with its ability to integrate nonlinear interactions and reduce overfitting through bootstrapping, is more suited for Himalayan SOC prediction. These results reinforce the utility of ensemble models for DSM applications in ecologically sensitive and topographically rugged landscapes. Overall, the study provides a spatial baseline of SOC concentration (not SOC stock) that can inform land management and soil conservation priorities across the North-western Himalayas at the landscape scale permitted by the coarsest covariates.

5. Conclusion

This study successfully generated a SOC concentration map on a 30 m output grid (its effective spatial detail is limited by the coarser native resolution of some covariates) for Himachal Pradesh using four machine learning techniques. Among them, the Random Forest model demonstrated the highest predictive accuracy (R^2 = 0.85, RMSE = 0.83%), followed at a considerable distance by GLM, SVM and Cubist (Table 2); the physically implausible extremes produced by the GLM and Cubist models, further support RF as the more defensible choice for the final SOC map. Topographic variables, particularly elevation, MRVBF, slope, and TRI, were consistently among the strongest predictors of SOC distribution across models, indicating a strong association between terrain-driven processes and SOC patterns in these moist temperate oak forests; as discussed above, high covariate importance reflects predictive association rather than proof of a causal mechanism. The RF regression slope of 1.5 (>1) reflects an over-steepening of predictions that is a known artefact of tree-based bagging under skewed target distributions (here, a few very high-SOC samples from Chamba district), which reinforces why RMSE and R^2 remain the primary evaluative criteria. The resulting 30 m SOC concentration maps provide a spatially explicit baseline that forest managers, policymakers, and conservation planners can use to identify SOC-rich areas warranting protection and SOC-depleted areas that may benefit from restoration across the seven-district study area in Himachal Pradesh. Realising the maps' full potential for carbon accounting will additionally require coupling them with bulk density and soil-depth data to convert SOC concentration to SOC stock, and validating model performance against soil samples collected independently of the present calibration dataset. The maps describe SOC concentration and do not directly estimate SOC stocks or carbon sequestration; stock assessment requires measurements of soil bulk density and sampling depth, which is identified as a requirement for future research. With these caveats addressed, the RF-based workflow presented here offers a practical, scalable template for developing standardized SOC inventories elsewhere in the Himalayan region to support land-use planning and, once combined with bulk density and depth data, carbon-stock assessments relevant to climate change mitigation.

6. Acknowledgements

All authors express their gratitude and thankfulness to the Director, ICFRE-FRI and DG, ICFRE, Dehradun for providing facilities and support.

7. Authors Contribution

Writing of draft manuscript, preparation of tables and figures was done by AS. Extensive editing and review was done by TB. Overall data acquisition done by DP and RR, formal analysis was done by GM and review and editing was done by AS.

8. Declarations

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

Ethical approval

Not applicable. The study did not involve human participants or animals.

Consent to participate

571 Not applicable. The study did not involve human participants.

572 **Consent to publish**

573 Not applicable. The study involved only soil sampling in forest plots; no human participants or personal data were
574 included, and no individual person's data, images or details are published.

575 **Conflict of Interest:** Authors declare no conflict of interest.

576 **Clinical trial:** Not applicable.

577 **Data Availability:** The data supporting the findings of this study are available from the corresponding author
578 upon reasonable request.

579 **9. Funding Information**

580 The authors did not receive support from any organization for the submitted work.

581 **10. References**

582 Arrouays, D., M. G. Grundy, A. E. Hartemink, J. W. Hempel, G. B. Heuvelink, S. Y. Hong, P. Lagacherie, G.
583 Lelyk, A. B. McBratney, and N. J. McKenzie. 2014. GlobalSoilMap: Toward a fine-resolution global grid
584 of soil properties. *Advances in Agronomy* 125:93–134.

585 Akpa, S.I., Odeh, I.O., Bishop, T.F. and Hartemink, A.E., 2014. Digital mapping of soil particle-size fractions for
586 Nigeria. *Soil Science Society of America Journal*, *78*(6), pp.1953-1966.

587 Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi: 10.1023/A:1010933404324.

588 Brown, S. and Lugo, A.E., 1982. The storage and production of organic matter in tropical forests and their role in
589 the global carbon cycle. *Biotropica*, pp.161-187.

590 Bradford, M.A., Wood, S.A., Addicott, E.T. *et al.* Quantifying microbial control of soil organic matter dynamics
591 at macrosystem scales. *Biogeochemistry* 156, 19–40 (2021). https://doi.org/10.1007/s10533-021-00789-5

592 Da Silva Chagas, C., de Carvalho Junior, W., Bhering, S.B. and Calderano Filho, B., 2016. Spatial prediction of
593 soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena*, *139*,
594 pp.232-240.

595 Dash, P. K., N. Panigrahi, and A. Mishra. 2022. Identifying opportunities to improve digital soil mapping in India:
596 A systematic review. Geoderma Regional 28: e00478. doi: 10.1016/j.geodrs. 2021.e00478

597 Dharumarajan, S., R. Hegde, N. Janani, and S. K. Singh. 2019. The need for digital soil mapping in India.
598 Geoderma Regional 16: e00204. doi: 10.1016/j.geodrs. 2019.e00204.

599 Farooq, I., Bangroo, S.A., Bashir, O., Shah, T.I., Malik, A.A., Iqbal, A.M., Mahdi, S.S., Wani, O.A., Nazir, N.
600 and Biswas, A., 2022. Comparison of random forest and kriging models for soil organic carbon mapping
601 in the Himalayan region of Kashmir. Land, 11(12), p.2180.

Hengl, Heuvelink, Kempen, Leenaars, Walsh, Shepherd, Sila, MacMillan, de Jesus, Tamene et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS ONE 2015, 10, e0125814.

Jena, R.K., Bandyopadhyay, S., Pradhan, U.K., Moharana, P.C., Kumar, N., Sharma, G.K., Roy, P.D., Ghosh, D., Ray, P., Padua, S. and Ramachandran, S., 2022. Geospatial modelling for delineation of crop management zones using local terrain attributes and soil properties. *Remote Sensing*, *14*(9), p.2101.

Jena, R.K.; Moharana, P.C.; Dharumarajan, S.; Sharma, G.K.; Ray, P.; Deb Roy, P.; Ghosh, D.; Das, B.; Alsuhaibani, A.M.; Gaber, A.; et al. Spatial Prediction of Soil Particle-Size Fractions Using Digital Soil Mapping in the North Eastern Region of India. Land 2023, 12, 1295.

Kalambukattu, J.G., Kumar, S., Das, B., Roy, T. et al., 2025. Digital mapping of soil organic carbon in the hilly and mountainous landscape of the Indian Himalayan region employing machine-learning techniques. Discover Soil, 2, p.35.

Kovačević, M., B. Bajat, and B. Gajić. 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154 (3–4):340–47. doi: 10.1016/j.geoderma.2009.11.005.

Kumar, A., Moharana, P.C., Jena, R.K., Malyan, S.K., Sharma, G.K., Fagodiya, R.K. and Doss, S.G., 2023. Digital mapping of soil organic carbon using machine learning algorithms in the Upper Brahmaputra Valley of Northeastern India. Land, 12(10), p.1841.

Liu, Zhang, Song, Li, Zhao, Yang, Wu, Yang. High-resolution and three-dimensional mapping of soil texture of China. Geoderma 2020, 361, 114061.

Mishra, G., & Gorai, A. K. (2020). Horizontal and vertical profiling of soil organic carbon stock in Nagaland, North East India. Current Science, 119(4), 632-640.

Mishra, G., Giri, K., Jangir, A., & Francaviglia, R. (2020). Projected trends of soil organic carbon stocks in Meghalaya state of Northeast Himalayas, India. Implications for a policy perspective. Science of the Total Environment, 698, 134266.

Mishra, G., Giri, K., Nath, A. J., & Francaviglia, R. (Eds.). (2023). Soil carbon dynamics in Indian Himalayan region. Singapore: Springer.

Mishra, G., Sharma, A., Barman, T., Moharana, P. C., & Suleymanov, A. (2026). Machine learning-based soil depth estimation from the diverse landscapes of Nagaland, North-Eastern Himalayas. Journal of Mountain Science, 23(7), 3104-3124.

Moharana, Meena, Nogiya, Jena, Sharma, Sahoo, Jha, Aditi, Vara Prasad, P.V. Impacts of Land Use on Pools and Indices of Soil Organic Carbon and Nitrogen in the Ghaggar Flood Plains of Arid India. Land 2022, 11,1180.

Moharana, P.C.; Mishra, G.; Keshavarzi, A.; Suleymanov, A. (Eds.) (2026). Harnessing Digital Mapping for Soil Resource Monitoring and Assessment, 1st ed.; CRC Press: Boca Raton, FL, USA, 2026.

Nenkam, A.M., Wadoux, A.M.C., Minasny, B., Silatsa, F.B., Yemefack, M., Ugbaje, S.U., Akpa, S., Van Zijl, G., Bouasria, A., Bouslihim, Y. and Chabala, L.M., 2024. Applications and challenges of digital soil mapping in Africa. *Geoderma*, *449*, p.117007.

Pahlavan-Rad, M.R. and Akbarimoghaddam, A., 2018. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). *Catena*, *160*, pp.275-281.

Pahlavan-Rad, M.R., Dahmardeh, K., Hadizadeh, M., Keykha, G., Mohammadnia, N., Gangali, M., Keikha, M., Davatgar, N. and Brungard, C., 2020. Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. *Catena*, *194*, p.104715.

Rawal, R., Sharma, A., Mishra, G., Barman, T., Chaturvedi, R. K., & Tewari, L. M. (2026). Variation in Soil Organic Carbon Along an Altitudinal Gradient Across Different Aspects in the Timberline Zone of the Western Himalaya, India. Plants, 15(18), 2775.

Ražauskaitė, Vanguelova, Cornulier, Smith, Randle and Smith (2020) A New Approach Using Modeling to Interpret Measured Changes in Soil Organic Carbon in Forests; The Case of a 200 Year Pine Chronosequence on a Podzolic Soil in Scotland. *Front. Environ. Sci.* 8:527549. doi: 10.3389/fenvs.2020.527549

Rao, A. and Kaushal, S., 2024. The Multifaceted Soils of Himachal Pradesh, India: Types and Characteristics. *Archives of Current Research International*, *24*(11), pp.145-160.

Rawat, D., Sati, S. P., Khanduri, V. P., Riyal, M., & Mishra, G. (2021). Carbon sequestration potential of different land use sectors of Western Himalaya. In Advances in carbon capture and utilization (pp. 273-294). Singapore: Springer Singapore.

Rengma, N.S., Yadav, M., Kalambukattu, J.G. and Kumar, S., 2023. Machine learning-based digital mapping of soil organic carbon and texture in the mid-Himalayan terrain. Environmental Monitoring and Assessment, 195(8), p.994.

Rossiter, D. G. 2018. Past, present & future of information technology in pedometrics. Geoderma 324:131–37. doi:10.1016/j.geoderma.2018.03.009query.

R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: https://www.R-project.org/ (accessed on 5 February 2023).

Sahoo, U.K., Singh, S.L., Gogoi, A., Kenye, A. and Sahoo, S.S., 2019. Active and passive soil organic carbon pools as affected by different land use types in Mizoram, Northeast India. *PloS one*, *14*(7), p.e0219969.

Schillaci, Acutis, Vesely, Saia. A simple pipeline for the assessment of legacy soil datasets: An example and test with soil organic carbon from a highly variable area. Catena 2019, 1, 110–122.

Singh, A., Khanduri, V. P., Rawat, D., Singh, B., Riyal, M. K., Thakur, T. K., ... & Chaturvedi, R. K. (2026). Himalayan altitude gradient drives divergent carbon storage: conifer biomass peaks, broadleaf soils stabilize. Carbon Research, 5(1), 48.

Tiwari, A., Khanduri, V. P., Rawat, D., Singh, B., Riyal, M. K., Mishra, G., & Kumar, M. (2023). Vegetation composition, soil properties, and carbon stock of montane forests along a disturbance in the Garhwal Himalaya, India. Frontiers in Forests and Global Change, 6, 1193792.

Yigini, Panagos, P. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Sci. Total Environ. 2016, 557–558, 838–850.

Suleymanov, R., Yurkevich, M., Bakhmet, O., Popova, T., Kungurtsev, A., Zakirov, D., ... & Suleymanov, A. (2025). Interpretable machine learning and remote sensing data reveal soil biogeochemistry patterns in agricultural systems. Land, 14(9), 1881.