E-ISSN: 2618-0618
P-ISSN: 2618-060X

NAAS Rating (2025): 5.20
www.agronomyjournals.com
2025; SP-8(9): 305-310
Received: 27-06-2025
Accepted: 30-07-2025

G Tiwari
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Nagpur, Maharashtra, India

Dr. VN Mishra
Professor, Soil Science and Agricultural Chemistry Div., Indira Gandhi Krishi Vishwavidyalaya, Raipur, Chhattisgarh, India

Dr. RP Sharma
Senior Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Udaipur, Rajasthan, India

Dr. Sudipta Chattaraj
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Kolkata, Maharashtra, India

B Dash
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Nagpur, Maharashtra, India

A Jangir
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Udaipur, Rajasthan, India

Dr. LC Malav
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Udaipur, Rajasthan, India

Corresponding Author:
G Tiwari
Scientist, ICAR-National Bureau of Soil Survey and Land Use Planning, Nagpur, Maharashtra, India

Overcoming data scarcity in soil depth prediction: A machine learning approach for India's Black Soil Region (BSR)

G Tiwari, VN Mishra, RP Sharma, S Chattaraj, B Dash, A Jangir and LC Malav

DOI: https://www.doi.org/10.33545/2618060X.2025.v8.i9Sd.3849

Abstract
Soil depth (SoD) is a fundamental property controlling ecosystem services, agricultural productivity, and hydrological processes, yet its spatial prediction remains a challenge in data-scarce regions. This study demonstrates the effectiveness of a Quantile Regression Forest (QRF) model to predict the spatial distribution of SoD and, crucially, quantify its prediction uncertainty using a limited dataset in the black soil region (BSR) of Amravati, Maharashtra, India. Ninety-two soil profiles were integrated with a suite of environmental covariates derived from terrain, climate, and remote sensing data. Key predictors were identified through recursive feature elimination. The QRF model explained 86% of the variance ($R^2 = 0.86$) with a root mean square error of 19.99 cm (sqrt-transformed) and 12.4 cm (back-transformed). A Lin's concordance correlation coefficient (CCC) of 0.93 indicated excellent agreement between predicted and observed values. The resulting map revealed distinct patterns: deeper soils in depositional valleys and plains (>150 cm) and shallower soils on erosional plateau tops and hillslopes (<50 cm). Predictive uncertainty was lowest in well-sampled alluvial plains and highest in sparsely sampled steep landscapes. The QRF model successfully handled non-linear relationships and provided robust, interpretable predictions from sparse data. The high-resolution SoD map with quantified uncertainty is a vital tool for optimizing agricultural water use, preventing land degradation, and implementing targeted soil conservation practices in this rainfed agricultural system.

Keywords: Digital soil mapping, prediction intervals, machine learning, SCORPAN, deccan trap, spatial variability, land use planning.

1. Introduction
Soil depth (SoD) is a master variable controlling a myriad of critical landscape functions. It fundamentally influences agricultural productivity by defining rootable space and water storage capacity, regulates hydrologic processes (Pelletier & Rasmussen, 2009) [18], and determines a landscape's susceptibility to erosion and degradation (Catani *et al.*, 2010; Gu *et al.*, 2018) [2, 5]. Furthermore, SoD is a key factor in carbon sequestration, vegetation growth (Meyer *et al.*, 2007) [15], and overall land quality assessment (Yang *et al.*, 2020) [26]. Defined as the sum of the thicknesses of surface and subsurface soil horizons down to bedrock or a paralithic contact (Liu *et al.*, 2019) [10], accurate spatial prediction of SoD is therefore essential for informed land management and environmental modelling.

Despite its importance, precise and accurate spatial data on SoD is scarce in this area. This scarcity stems from several factors: the high cost and labor-intensive nature of direct field measurement, particularly for deep soils (Tesfa *et al.*, 2009) [23], the high spatial variability of SoD across landscapes (Vanwalleghem *et al.*, 2010) [24], and a historical research focus primarily on surface horizons (epipedons) for agricultural purposes (Liu *et al.*, 2013; Singh *et al.*, 2020) [11, 22].

To address this, digital soil mapping (DSM) approaches have been developed to predict SoD from correlated environmental covariates. Prior research has applied various modeling paradigms, including mechanistic landscape evolution models (Bonfatti *et al.*, 2018; Liu *et al.*, 2019) [1, 10] and empirical methods ranging from geostatistics (ordinary kriging; Yan *et al.*, 2021)

[25] to machine learning (ML) algorithms (Random Forest, Cubist; Zhang *et al.*, 2021; Mulder *et al.*, 2016) [27, 16]. While ML models like Random Forest excel at capturing non-linear relationships and complex interactions inherent in soil-forming factors, they possess a significant limitation for practical application: they typically quantify prediction uncertainty inadequately or circumstantially (Ma *et al.*, 2014) [12]. This is a critical shortcoming, as ignorance of uncertainty can lead to misguided decisions. This issue is exacerbated in regions with sparse sample data, where uncertainty is inherently high and reliable estimation is most needed (Lagacherie *et al.*, 2019) [8].

This challenge is acutely felt in India's agriculturally critical black soil regions (BSRs). These regions are characterized by complex topography and non-stationary soil-environment relationships, making them ideal candidates for ML approaches. However, detailed soil survey data is often limited and spatially sparse, rendering traditional models that require dense calibration datasets less effective (Cheng *et al.*, 2019; Guo *et al.*, 2019) [4, 6] and highlighting the need for methods that can work with scarce data while providing robust uncertainty estimates.

Therefore, this study aims to bridge this gap by applying Quantile Regression Forest (QRF), a machine learning algorithm specifically designed to provide conditional quantiles and thus robust prediction intervals to predict SoD in a data-scare BSR of India. Unlike standard Random Forest or other ML models, QRF retains the full distribution of values in decision trees, allowing it to directly quantify predictive uncertainty without distributional assumptions (Meinshausen, 2006) [14], making it ideally suited for this challenge. Specifically, our objectives are to: (i) Develop a QRF model for predicting the spatial distribution of SoD using a sparse set of soil profiles and environmental covariates. (ii) Identify the key environmental factors controlling SoD variation in the region. (iii) Generate a high-resolution map of SoD with quantified prediction intervals to support risk-aware land management decisions.

2. Materials And methods

2.1. Study area

The study was conducted in the BSR of Amravati district of Maharashtra, India, covering approximately 59,758 ha (Fig. 1) and lies between 20°24' to 21°33'N and 77°06' to 78°18'E. The region is part of the Deccan Plateau, characterized by flat-topped hills (plateaus) and intervening valleys. The climate is semi-arid tropical with a mean annual rainfall of 975 mm and a mean annual temperature of 28°C. The geology is predominantly Deccan Trap basalt, with alluvial deposits in the valley of the Purna River. Rainfed agriculture is the dominant land use. The soils have an ustic moisture regime and an isohyperthermic temperature regime.

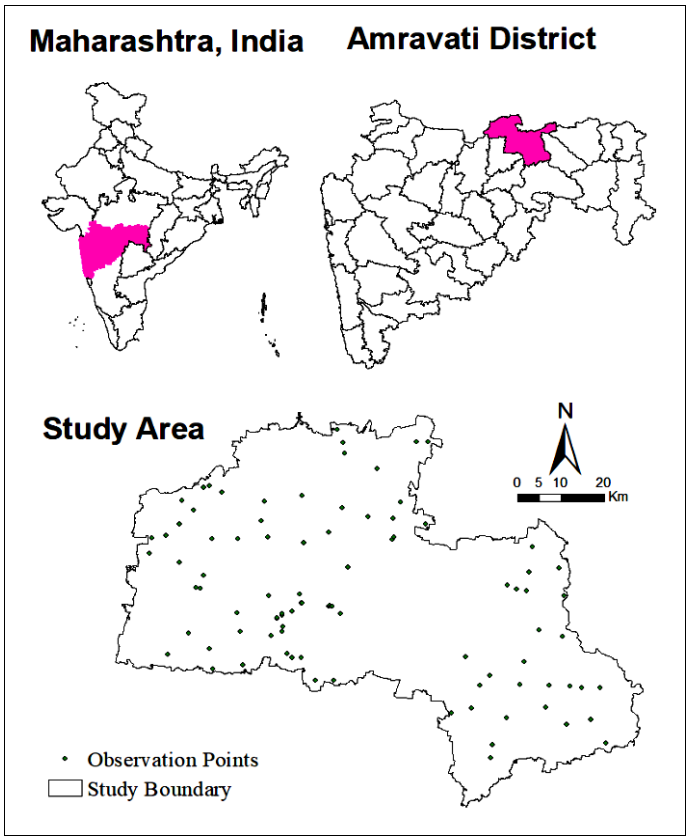

Fig 1: Location of study area

2.2. Legacy data

A total of 92 soil profiles were used, comprising 72 newly dug profiles (1: 10,000) and 20 from legacy data (1: 250,000). Profiles were excavated up to 150 cm or until a lithic/paralithic contact was encountered. Depth was recorded as the point where the volume of coarse fragments >2 mm exceeded 75% or as expert-estimated based on local topography and parent material where contact was not reached. The spatial distribution of samples was uneven, with an average density of one sample per 133 km², characterizing a sparse dataset. The raw SoD data were positively skewed. A square-root (sqrt) transformation was applied to achieve a near-normal distribution (Fig 2), which is conducive for ML modeling.

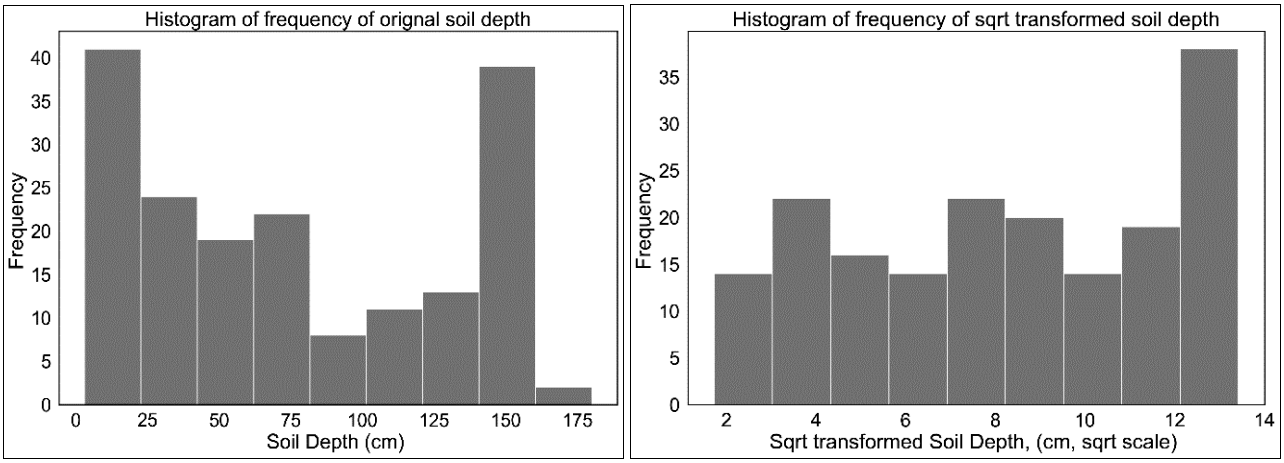

Fig 2: Histograms of frequency of original (A) and square root (sqrt)-transformed (B) SoDs obtained from field survey.

2.3 Environmental covariates

Based on the SCORPAN framework, 47 covariates representing Soil, Climate, Organisms, Relief, Parent material, Age, and Space were compiled (Table 1). Terrain attributes (30 m resolution) were derived from the SRTM DEM using SAGA GIS. Climate variables (Mean Annual Precipitation and Temperature) were sourced from *WorldClim* (1km resolution). Time-series Landsat 5 TM imagery was used to compute spectral indices (NDVI, EVI, SAVI, FV) for three seasons (*kharif*, *rabi*, *zaid*). Annual average Land Surface Temperature (LST) was derived from MODIS data (1km). All covariates were resampled to a 30 m grid using bilinear interpolation.

Table 1: Environmental covariates used for digital soil mapping of SoD

S. N	Group	Covariate	Abbr.	Res.
1	Climate	Mean annual precipitation (mm)	MAP	1 km
2	Terrain	Elevation (m)	Elv	30 m
		Slope (%)	Slope	30 m
		Relative Slope Position	RSP	30 m
		Channel Network Base Level	CNBL	30 m
		Channel Network Distance	CND	30 m
		Multi-Resolution Ridge Top Flatness Index	MRRTF	30 m
		Multi-Resolution Valley Bottom Flatness Index	MRVBF	30 m
		Valley Depth	VD	30 m
		Topographic Wetness Index	TWI	30 m
		LS-Factor	LSf	30 m
3	Vegetation (*Kharif* (k), *Rabi* (r), *Zaid* (z))	Land surface thermal conditions	LST	30 m
		Normalized Difference Vegetation Index.	NDVI	10 m
		Near infrared	NIR	10 m
		Enhanced Vegetation Index	EVI	10 m
		Fractional vegetation	FV	10 m

2.4. Variable Selection and Quantile Regression Forest Modelling

To avoid overfitting and reduce multicollinearity, recursive feature elimination (RFE) was performed using the rfe function in the caret R package. Variables were ranked by their importance (%IncMSE) from a preliminary RF model, and the optimal subset that maximized model performance was selected.

The QRF model was implemented using the ranger package in R. The model was tuned via out-of-bag (OOB) error estimation; the optimal parameters were *mtry* = 5, *num.trees* = 1000, and *min.node.size* = 5. Unlike standard RF, which estimates the conditional mean, QRF retains the entire distribution of values in the leaf nodes of each tree, allowing for the computation of any quantile of the conditional distribution. This study generated the 0.05, 0.5 (median), and 0.95 quantiles to represent the lower bound, median prediction, and upper bound of the 90% prediction interval, respectively.

2.5 Evaluation of model performance

Model performance was evaluated using a repeated (20 times) 10-fold cross-validation. Performance metrics included coefficient of determination (R^2), root mean square error (RMSE), mean error (ME), and Lin's concordance correlation coefficient (CCC). Good models have a root mean square error that is close to 0, R^2 and CCC that is equal to or close to 1.

$$\text{Coefficient of determination } (R^2) = 1-\frac{\sum_{i=1}^{n}(pi-oi)^2}{\sum_{i=1}^{n}(\overline{pi}-\overline{oi})^2} \quad \text{(i)}$$

$$\text{Mean error (ME)} = \frac{1}{n}\sum_{i=1}^{n}(pi-oi) \quad \text{(ii)}$$

$$\text{Root mean squared error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(pi-oi)^2} \quad \text{(iii)}$$

where, pi and oi are predicted and observed values, $\overline{pi}$ and $\overline{oi}$ are means of these values.

$$\text{Lin's concordance correlation coefficient (CCC)} = \frac{2\rho\sigma_o\sigma_p}{\sigma_o^2+\sigma_p^2+(\mu_o-\mu_p)^2} \quad \text{(iv)}$$

In this formula, ρ is the Pearson correlation coefficient between the observed and predicted values, μ_0 and μ_p are the means of the observed and predicted values, and σ_0^2 and σ_p^2 are the corresponding variances.

3. Results and Discussion

3.1 Variable importance

The most influential predictors of soil depth (SoD) were MRVBF, LST-z, NIR-r, NDVI-r, and FV-r (Fig. 3). Vegetation-related indices (NIR-r, NDVI-r, FV-r) were important because deeper soils store more heat and water, supporting denser vegetation, while vegetation itself reduces erosion by trapping soil particles. NIR-r was particularly significant due to its sensitivity to both plant vigor and soil moisture.

Among terrain covariates, MRVBF effectively captured depositional versus erosional settings, with shallow soils occurring in erosion-dominated positions and deeper soils in depositional zones. LST-z also contributed strongly, reflecting the thermal buffering capacity of deeper soils. In contrast, slope, DEM, and curvature were relatively weak predictors at the regional scale, consistent with earlier findings (Scarpone *et al.*, 2016; Penízek & Borůvka, 2006; Chen *et al.*, 2019) [21, 19, 3]. However, previous small-scale studies (Patton *et al.*, 2018) [17] suggest that curvature and slope become more important in hillslope-scale environments, indicating a scale-dependent effect.

Seasonal variation in LST further highlighted that summer daytime LST best explained SoD differences, agreeing with Liu *et al.*, (2020) [9], who reported strong predictive performance of seasonal LST in low-relief areas. Overall, the results demonstrate that SoD patterns are shaped more by geomorphic processes (erosion–deposition dynamics) than by pedogenic factors. Terrain indices and vegetation signals were more decisive than lithological or climatic variables, emphasizing the dominant role of landscape processes in controlling soil depth distribution in the study region.

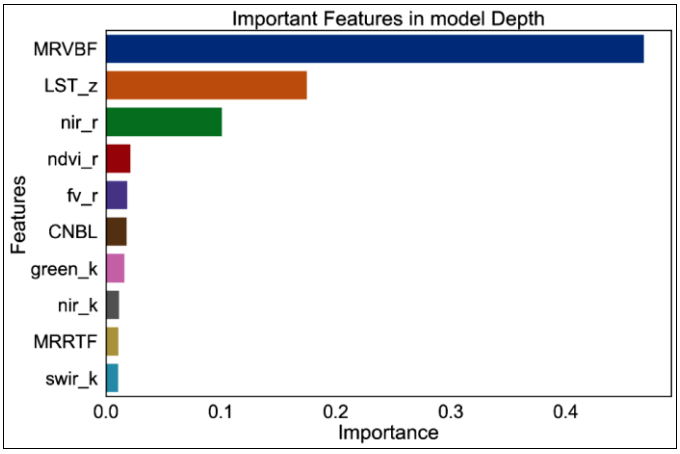

Fig 3: Relative importance (% IncMSE) of environmental covariates.

3.2 Model predictive performance

The QRF model achieved strong predictive accuracy, explaining 86% of the variance in soil depth (SoD) with a cross-validation R^2 of 0.86, RMSE of 19.99 cm (sqrt scale), and a high CCC of 0.93, indicating close agreement between predicted and observed values. Compared with earlier studies, the performance was notably higher. Zhang *et al.* (2021) [27] reported an R^2 of 0.61 for a black soil watershed in Northeast China, while Mulder *et al.* (2016) [16] achieved only 0.11 at the national scale. Piecewise models by Malone and Searle (2020) [13] in Australia yielded accuracies of 99% (rock outcrops), 85% (deep soils), and CCC = 0.77 for intermediate soils. These comparisons highlight that QRF provided robust and reliable predictions for SoD even under data-scarce conditions in India's black soil region.

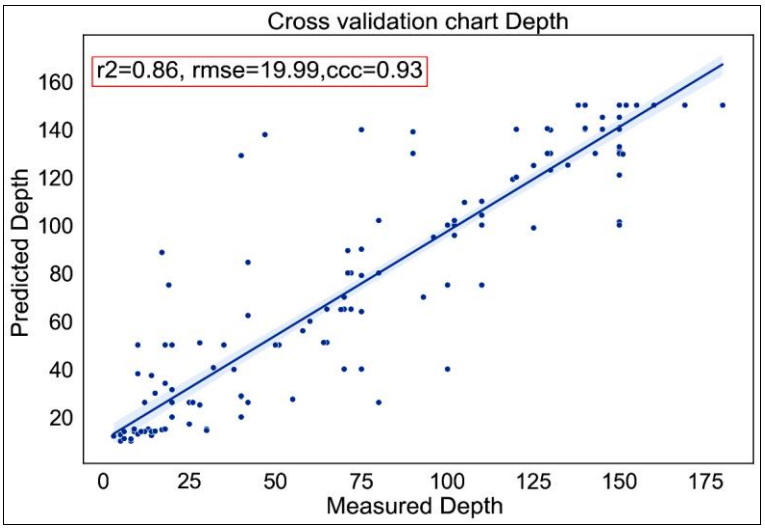

Fig 4: Cross-validation performance of the QRF model

3.3 Predicted Soil Depth Distribution

The predicted soil depth (SoD) map for the study area is shown in Fig. 5. Soil is predicted to be deepest in the plain areas and shallowest in the high plateau. Because gravity, freeze–thaw, and water erosion are more powerful on the slopes and ridges of mountains than in valley bottoms, the soils on the plateau top are typically shallower. Near the ridge crests, soils are extremely thin or absent, exposing bare rock.

In the pediment, the terrain becomes gentle and open, and the eroded material carried from the upper slopes is deposited in valleys and plains, leading to overall deeper soils than the pediment itself. The valley areas have deeper soils than the plains due to deposition of alluvial, colluvial, and aeolian materials. Lower plains tend to have deeper soils than the middle and upper plains, as soil materials are continuously transported and deposited by wind and water.

Two small windows in the northern and central parts of the research area (Fig. 5B and C) demonstrate that our predictions effectively captured local variations in SoD in addition to regional patterns. SoD changes linked to vertical zonality and microtopography are evident across both small-window maps. For instance, SoD increases from the plateau to the pediment in Fig. 5B, while Fig. 5C shows marked SoD variability among plateaus, pediments, valleys, and surrounding landforms. Such spatial patterns are consistent with findings from Zhang *et al.*, (2021) [27] and Henderson *et al.*, (2005) [7], who reported strong topographic control on soil depth in hilly terrains.

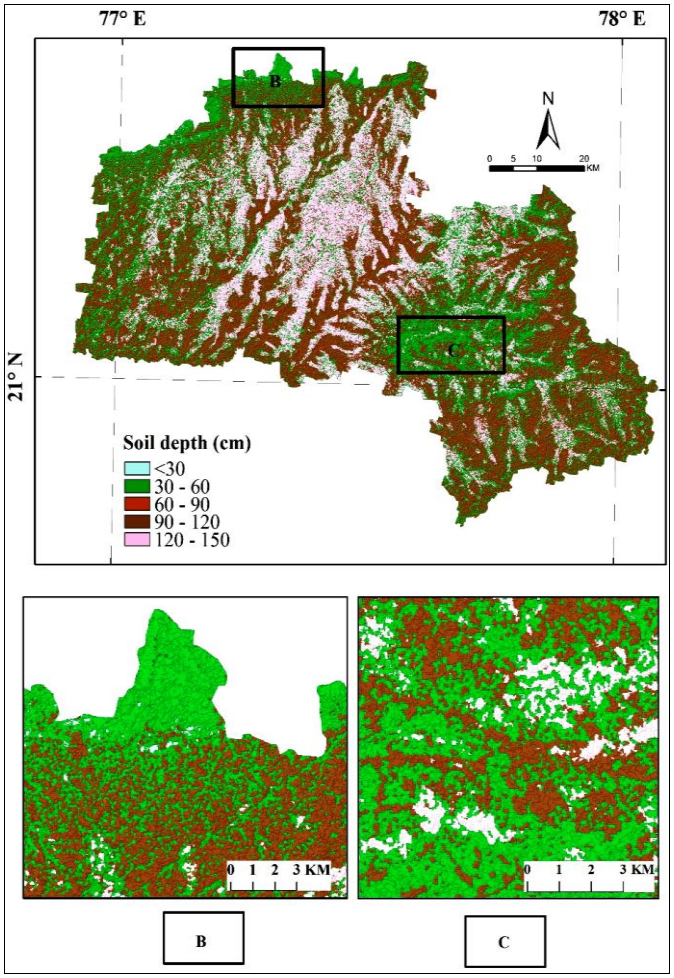

Fig 5: The predicted SoD map

3.4 Spatial Uncertainty of Soil Depth Prediction

The spatial distribution of uncertainty associated with the SoD prediction is shown in Fig. 6. Areas of the pediment, plateau, and hillslope exhibit relatively high uncertainty, represented by dark blue shades. These regions were difficult to access, resulting in sparse or absent soil survey data. In contrast, valleys and alluvial plains, where field accessibility was easier and survey density was higher show lower uncertainty, represented by green and yellow shades.

This pattern highlights the importance of observation density for predictive reliability, as also noted by Malone and Searle (2020) [13] and Poggio *et al.*, (2021) [20]. By explicitly quantifying spatial uncertainty, the SoD map can be provided to end-users with clear confidence levels, ensuring its usefulness for decision-making in land evaluation, agricultural planning, and environmental management.

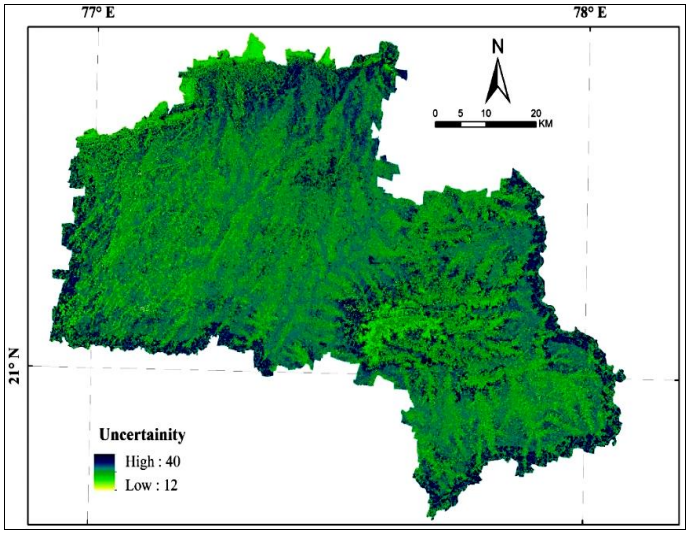

Fig 6: Map of predictive uncertainty for SoD

4. Conclusion

This study provides a robust framework for high-resolution soil depth mapping in data-scarce regions. The QRF model successfully handled sparse data and non-linear relationships to deliver accurate predictions with quantifiable uncertainty. The findings underscore the dominance of geomorphic processes in shaping SoD patterns in this landscape. The final map is a decision-support tool for enhancing agricultural water productivity and soil conservation. Future work will focus on integrating proximal sensing data to validate depth estimates and scaling this approach to the entire black soil belt of Central India.

5. Acknowledgment

I am deeply grateful to the faculty of the Soil Science Division, IGKV, Raipur, for their invaluable academic guidance. My sincere thanks to the faculty of the SRS Division and Director of ICAR-NBSS & LUP, for their expert mentorship and for providing an exceptional research environment.

References

1. Bonfatti BR, Hartemink AE, Giasson E, Iório JFS, Demattê JAM. Digital mapping of soil carbon in a degraded watershed of the Brazilian savanna. Geoderma. 2018;324:48-59.
2. Catani F, Segoni S, Falorni G. An empirical geomorphology-based approach to the spatial prediction of soil thickness at catchment scale. Water Resour Res. 2010;46(5):W05508.
3. Chen S, Arrouays D, Mulder VL, Poggio L, Minasny B. Digital mapping of soil carbon stocks with machine learning and environmental covariates in the northern Circumpolar Permafrost Region. Earth Sci Rev. 2021;223:103858.
4. Cheng M, Wang Y, Zhang J, Liu Y, Li Y. Evaluation of different machine learning approaches for prediction of soil organic matter and soil moisture content in a forested watershed, China. Soil Sci Soc Am J. 2019;83(4):1107-1118.
5. Gu Z, Duan X, Shi Y, Li Y, Pan X. Spatiotemporal prediction of soil moisture content using multiple-linear regression and random forest in a small catchment of the Loess Plateau, China. Catena. 2018;171:583-594.

6. Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma. 2019;237:49-59.
7. Henderson BL, Bui EN, Moran CJ, Simon DAP. Australia-wide predictions of soil properties using decision trees. Geoderma. 2005;124(3-4):383-398.
8. Lagacherie P, Arrouays D, Bourennane H, Gomez C, Martin MP, Saby NPA. How far can the uncertainty on a digital soil map be known? A numerical experiment using pseudo values of clay content obtained from Vis-NIR hyperspectral imagery. Geoderma. 2019;337:1320-1328.
9. Liu F, Wu H, Zhao Y, Li D, Yang J, Song X. Mapping high resolution National Soil Information Grids of China. Sci Bull. 2020;67:1-12.
10. Liu H, Zhang XL, Wu W, Tang Z. Prediction of soil depth in a large watershed using environmental covariates and machine learning. Soil Sci Soc Am J. 2019;83(5):1402-1413.
11. Liu X, Zhang G, Heathman GC, Wang Y, Huang C. Fractal features of soil particle size distribution as affected by plant communities in the forested region of Mountain Yujia, China. Geoderma. 2013;206:74-82.
12. Ma Y, Minasny B, Malone BP, McBratney AB. Pedology and digital soil mapping (DSM). Eur J Soil Sci. 2014;70(1):2-12.
13. Malone BP, Searle R. Improvements to the Australian national soil thickness map using an integrated data mining approach. Geoderma. 2020;377:114579.
14. Meinshausen N. Quantile regression forests. J Mach Learn Res. 2006;7:983-999.
15. Meyer SE, García-Moya E, Lagunes-Espinoza LDC. Soil depth and fertility effects on biomass and nutrient allocation in jaragua grass. J Range Manag. 2007;60(4):388-395.
16. Mulder VL, Lacoste M, Richer-de-Forges AC, Arrouays D. GlobalSoilMap France: A high-resolution spatial database of soil properties. Geoderma. 2016;279:1-12.
17. Patton NR, Lohse KA, Godsey SE, Crosby BT. Predicting soil thickness on soil mantled hillslopes. Nat Commun. 2018;9(1):3329.
18. Pelletier JD, Rasmussen C. Geomorphically based predictive mapping of soil thickness in upland watersheds. Water Resour Res. 2009;45(9):W09417.
19. Penízek V, Borůvka L. Soil depth prediction supported by primary terrain attributes: a comparison of methods. Plant Soil Environ. 2006;52(9):424-430.
20. Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, *et al*. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil. 2021;7(1):217-240.
21. Scarpone C, Schmidt MG, Bulmer CE, Knudby A. Modelling soil thickness in a forested watershed of southern British Columbia, Canada. Geoderma. 2016;282:59-69.
22. Singh K, Mishra AK, Singh B, Singh RP, Patra DD. Tillage effects on crop yield and physicochemical properties of soil organic matter in a wheat–soybean double-cropping system. Soil Tillage Res. 2020;199:104596.
23. Tesfa TK, Tarboton DG, Chandler DG, McNamara JP. Modeling soil depth from topographic and land cover attributes. Water Resour Res. 2009;45(10):W10438.
24. Vanwalleghem T, Stockmann U, Minasny B, McBratney AB. A quantitative model for integrating landscape evolution and soil formation. J Geophys Res Earth Surf. 2010;115(F4):F04013.
25. Yan F, Shangguan W, Zhang J, Hu B. Estimation of soil organic carbon stock in China based on high-density soil sampling. Sci Total Environ. 2021;754:142150.
26. Yang QJ, Zhang DP, Liu MB, Li R, Yang Y. Effects of soil depth on the spatial patterns of soil moisture and vegetation on a hillslope. J Hydrol. 2020;589:125135.
27. Zhang S, Huang Y, Shen C, Ye H, Du Y. Prediction of soil organic carbon and its spatial distribution in a small watershed based on machine learning. Soil Sci Soc Am J. 2021;85(4):1085-1098.