frontiers in Soil Science

TYPE Original Research
PUBLISHED 26 March 2026
DOI 10.3389/fsoil.2026.1745154

OPEN ACCESS

EDITED BY
Hamdi A. Zurqani,
University of Arkansas at Monticello, United States

REVIEWED BY
Tegbaru B. Gobezie,
Ontario Ministry of Agriculture, Food and Agribusiness (OMAFA), Canada
Mounir Oukhattar,
Aix-Marseille Université, France

*CORRESPONDENCE
Carlos Carbajal-Llosa
cmcarbajal@gmail.com
Rodolfo Chuchon-Remon
rodolfo.chuchon@outlook.com

RECEIVED 12 November 2025
REVISED 20 February 2026
ACCEPTED 03 March 2026
PUBLISHED 26 March 2026

CITATION
Salazar-Coronel W, Carbajal-Llosa C and Chuchon-Remon R (2026) Soil organic carbon content mapping along the coast of northern Peru: an ensemble machine learning approach. *Front. Soil Sci.* 6:1745154. doi: 10.3389/fsoil.2026.1745154

Soil organic carbon content mapping along the coast of northern Peru: an ensemble machine learning approach

Wilian Salazar-Coronel[1,2], Carlos Carbajal-Llosa[3]* and Rodolfo Chuchon-Remon[4]*

[1]Estación Experimental Agraria Vista Florida, Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), Lambayeque, Peru, [2]Programa de Doctorado en Ingeniería y Ciencias Ambientales, Universidad Nacional Agraria La Molina, Lima, Peru, [3]Centro Experimental La Molina, Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), Lima, Peru, [4]Estación Experimental Agraria El Porvenir, Dirección de Servicios Estratégicos Agrarios, Instituto Nacional de Innovación Agraria (INIA), San Martín, Peru

Introduction: Soil organic carbon (SOC) content plays a fundamental role in regulating the global carbon cycle and mitigating climate change. It is also a key marker of soil health and a vital plant component. Its distribution in space varies in dry ecosystems, where climate and land use affect it. This study aimed to estimate and map SOC in the Motupe River Basin, northern Peru, by applying machine learning algorithms and ensemble methods.

Methods: Four predictive models were evaluated: Support Vector Regression (SVR), Random Forest (RF), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost), together with two ensemble approaches—simple averaging and weighted — integrating topographic, climatic, edaphic, and vegetation indices variables. Spatial autocorrelation was minimized by spatial block cross-validation. Uncertainty was measured with bootstrapping and the Prediction Interval Ratio (PIR) derived from 90% prediction intervals.

Results and discussion: Best performance was achieved by XGBoost (R^2 = 0.83), weighted ensemble (R^2 = 0.70), and RF (R^2 = 0.63). The most influential predictors were EVI, GNDVI, temperature, TRI, and pH. SOC contents showed relatively higher concentrations (>0.7%) in areas with greater vegetation density, within a semi-arid context where SOC levels are generally low. In contrast, lower areas exhibited reduced SOC contents (< 0.6%). The uncertainty analysis indicated that SOC predictions had high to moderate confidence (PIR < 0.2) in the middle-and upper zones of the basin, and moderate confidence (0.1–0.2) in the lower areas. The results suggest that machine learning and ensemble methods improve SOC prediction, benefiting the sustainable management of soil fertility and quality in arid and semi-arid ecosystems of northern Peru.

KEYWORDS
machine learning, soil organic carbon, topographic indices, vegetation indices, digital soil mapping, ensemble modeling

1 Introduction

Extreme climatic events have intensified in recent decades as a consequence of climate change, driven mainly by the increase in atmospheric carbon dioxide (CO_2) concentrations associated with human activities such as fossil fuel combustion and deforestation (1–4). In this context, soils play a strategic role in climate change mitigation, as they store approximately three times more carbon than the atmosphere, primarily in the form of soil organic carbon (SOC) (5–7), whose content is conditioned by the interaction of soil-forming factors such as climate, relief, parent material, and vegetation cover. Beyond its climatic function, SOC content constitutes a key indicator of soil quality and functionality, given its direct effect on physical, chemical, and biological properties, nutrient availability, and plant productivity (8, 9). However, SOC content is highly dynamic and exhibits marked spatial and temporal variability, especially in arid and semi-arid ecosystems, where factors such as climate, topography, land use, and vegetation cover control its distribution (10, 11). This spatial heterogeneity limits the ability of point-based sampling to adequately represent SOC content patterns at the landscape scale, underscoring the need for robust spatial prediction approaches that allow for more accurate assessments of carbon sequestration potential and support sustainable soil management strategies (12–14).

In the Motupe River basin, the interaction of soil-forming factors along pronounced altitudinal and climatic gradients generates significant spatial heterogeneity in vegetation and edaphic properties. Key variables such as texture, pH, bulk density, and electrical conductivity characterize soil conditions, while terrain attributes and climatic factors regulate the spatial distribution patterns of SOC content throughout the basin. Given this inherent spatial variability, Digital Soil Mapping (DSM) constitutes a consolidated approach for the spatial prediction of soil properties by integrating point-based soil observations with environmental covariates and machine learning algorithms. Within this framework, topography has been widely recognized as a key factor influencing the distribution of soil organic carbon (SOC) content, due to its association with hydrological and microclimatic processes and the redistribution of materials across the landscape (15, 16). Variables such as elevation, slope, and topographic indices are commonly considered determinants of SOC spatial variability, together with land use and agricultural management practices (16, 17). Moreover, recent methodological advances have enabled the integration of remote sensing products, vegetation indices, and edaphic variables to further improve the spatial estimation of SOC (18–21).

In this context, the Machine learning (ML), mainly nonlinear deterministic models, is the preferred choice for handling big data due to its ability to forgo strict assumptions regarding data distribution or stationarity, often required with large spatial extends and legacy data (22). The ML algorithms including Random Forest (RF), Support Vector Regression (SVR), artificial neural networks (ANN), and Extreme Gradient Boosting (XGBoost)—have been extensively applied within DSM frameworks, showing strong performance in modeling complex, non-linear relationships between SOC and multiple environmental covariates across diverse regions (23–26). These approaches are particularly useful in contexts characterized by limited data availability and high spatial heterogeneity.

Nevertheless, significant information gaps persist, especially in tropical dry forest ecosystems and arid and semi-arid regions of South America. In northern Peru, these ecosystems exhibit high vulnerability to climate change and anthropogenic activities, such as agricultural and livestock expansion, which have contributed to soil degradation and the reduction of carbon stocks (27, 28). Although there are studies that have addressed the estimation of SOC content in similar ecosystems and in other regions of the world (29–31), at the Peruvian level, basin-scale analyses remain scarce. In addition, available estimates at regional or national scales often present coarse spatial resolutions and high uncertainties associated with local heterogeneity in soils, vegetation, and topography (31).

Given these limitations, basin-scale studies represent an opportunity to reduce uncertainty in SOC estimation and improve the understanding of its spatial patterns (32, 33). In this regard, the combined use of multiple machine learning algorithms, together with ensemble modeling approaches and spatial validation techniques, can contribute to generating more robust and reliable predictions. Therefore, this study aims to generate a high-resolution soil organic carbon map of the Motupe River basin, northern Peru, using locally calibrated machine learning ensemble models that integrate topographic, climatic, and vegetation covariates. Unlike global SOC products, our approach relies on site-specific field observations and spatially explicit environmental predictors to capture fine-scale SOC variability, calculate prediction uncertainty, and evaluate the contribution of ancillary soil properties as covariates within the modeling framework. The resulting locally calibrated, high-resolution SOC map should support sustainable soil management, dry forest ecosystem conservation, and the design of climate change mitigation and adaptation strategies at local and regional scales (34).

2 Materials and methods

2.1 Area of the study

The study was conducted in the Motupe watershed, located in the Lambayeque region of northern Peru (Figures 1A–C). The central point of the watershed is situated at latitude -6.314302°, and longitude: -79.59176°, within a dry forest ecosystem. The watershed covers an area of 3673.66 km², ranging from sea level up to 3500 m.a.s.l. as shown in Figure 1A.

Climatic data indicate that temperatures vary seasonally up to approximately 1000 m a.s.l., with average values of 13 °C in winter and 35 °C in summer. In these areas, rainfall is scarce, reaching up to 33 mm $month^{-1}$, in March and an annual accumulation of 69.02 mm $year^{-1}$. In the highlands, between 1500 and 3500 m a.s.l., temperatures range between 11 and 25 °C, with no marked seasonal differences. Rainfall, unlike in the lowlands, can reach 145.8 mm $month^{-1}$ in March, with an annual accumulation of 406.93 mm $year^{-1}$, and the wet period extends from September to

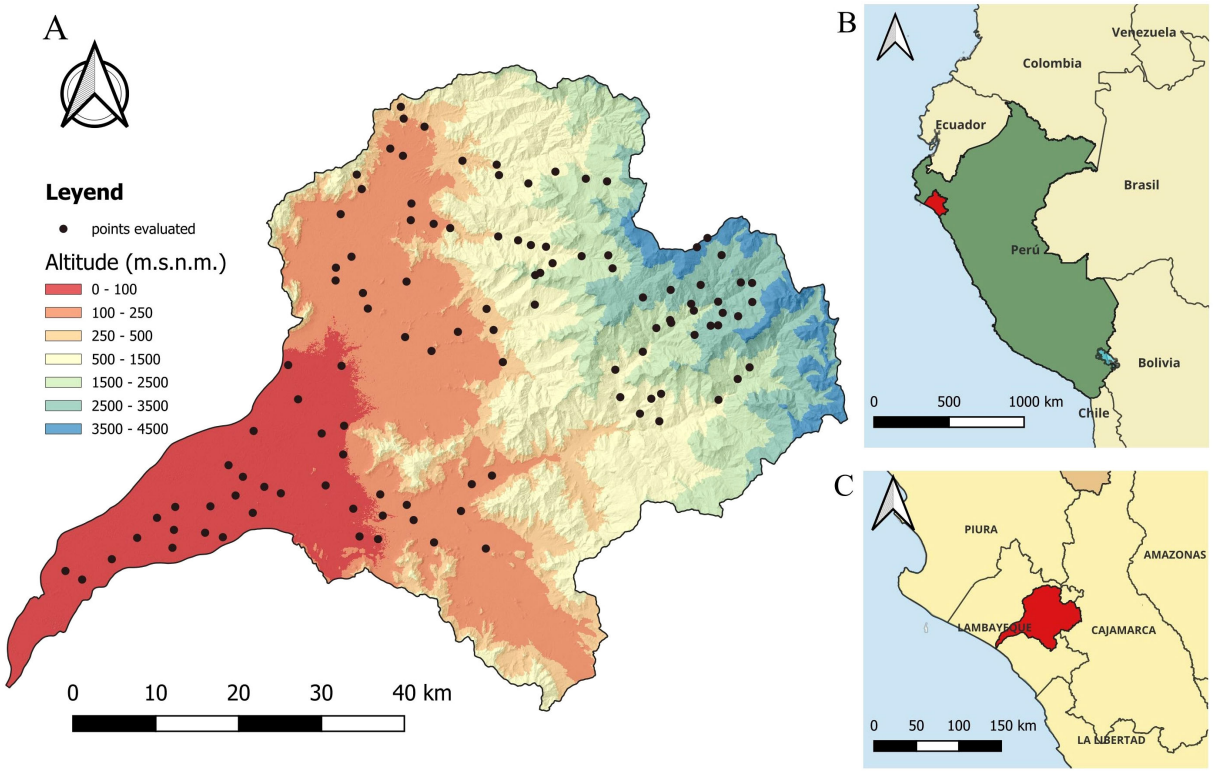

FIGURE 1
Location of the study area: **(A)** Distribution of the sampling points and elevation across the basin (meters above sea level, m.a.s.l.); **(B)** Location of the study region within Peru; and **(C)** Position of the study basin.

April. Temperature and precipitation data were obtained from SENAMHI HSR PISCO database (35). Vegetation also varies according to altitude: in lowlands (<250 m.a.s.l), vegetation is sparse due to the dry forest ecosystem, with the presence of algarrobo (*Neltuma pallida*), sapote (*Capparis scabrida*), and other characteristic species. In the mid-elevation zone, shrub vegetation predominates, with the presence of natural grasses, while in the highlands, ichu (*Stipa ichu*) and other high-Andean species are common. Agriculture is one of the main activities in the basin. In the lowlands, rice, maize, and fruit crops such as mango, avocado, and lemon predominate. In the highlands, small-scale cultivation of maize, potato, wheat, and barley is common, mainly for local consumption. Forested areas with pine and eucalyptus are also found, serving as sources of wood production.

2.2 Data acquisition and processing

2.2.1 Field sampling and laboratory processing

A total of 109 soil samples were randomly collected across the study area between September and December 2023. Randomization was performed using the AcATaMa plugin in QGIS 3.38. The initial sampling points were established in the laboratory; however, some were relocated in the field due to access limitations. At each sampling point, four subsamples were collected with a shovel and then thoroughly mixed to form a single homogeneous composite sample. After collection, samples were transported to the LABSAF-INIA laboratory at the Vista Florida Experimental Station for processing. Soil pH was determined following the USEPA (2004) method (36), electrical conductivity (EC) was measured according to ISO 11265:2025 (37) and soil texture was analyzed following the NOM-021-RECNAT-2000, 2002 (38). Bulk density (BD) was determined from an undisturbed soil core collected concurrently with the SOC samples. The sample was oven-dried, and BD was calculated as the ratio of dry soil mass to the cylinder volume. Soil organic carbon (SOC) content was determined using the ISO 10694:1996 (39).

2.2.2 Variables used for SOC content prediction

A total of 19 predictor variables were evaluated (Table 1), grouped into three main categories. The first group comprised variables derived from the digital elevation model (DEM), including elevation, aspect, hillshade, slope, Topographic Position Index (TPI), Topographic Ruggedness Index (TRI), and Topographic Wetness Index (TWI), all computed in QGIS 3.38 with a spatial resolution of 30 m. This approach is consistent with recent studies that have integrated multiple DEM-derived variables within machine learning–based soil property modeling frameworks (46).

The second group included physicochemical soil properties determined in the laboratory. Soil property data (bulk density, pH, electrical conductivity, clay, silt, and sand content) were spatially interpolated using ordinary kriging (OK) with the fitted variogram model, yielding continuous raster layers at 30 m resolution. Leave-one-out cross-validation (LOOCV) was performed to assess interpolation accuracy. All spatial

TABLE 1 List of predictor variables for SOC estimation.

Predictor	Group	Source Equation	Reference
Elevation, Aspect, Hillshade, Slope (DEM)	Topographic	Qgis 3.38 (Raster Analysis and Terrain Analysis)	(40)
Topographic Position Index (TPI)			
Topographic Ruggedness Index (TRI)			
Topographic Wetness Index (TWI)			
pH, Electrical Conductivity (EC), DB, Sand, Clay, Silt	Soil Property	Laboratory	
Temperature	Climate	WorldClim Version 2	(41)
Precipitation			
Enhanced Vegetation Index (EVI)	Index Vegetation	$2.5_{\star}\frac{NIR - R}{(NIR + C1_{\star}R - C2_{\star}B + L)}$	(42)
Normalized Difference Vegetation Index (NDVI)		$\frac{NIR - R}{NIR + R}$	(43)
Green normalized difference vegetation index (GNDVI)		$\frac{NIR - G}{NIR + G}$	(44)
Soil Adjusted Vegetation Index (SAVI)		$(\frac{NIR - R}{NIR + R + L})_{\star}(1 + L)$	(45)

interpolations were performed in R (version 4.5x) using `gstat` (47) for variogram modeling and kriging, and `automap` (48) for automatic variogram fitting.

The third group consisted of vegetation indices—NDVI, SAVI, GNDVI, and EVI—calculated from Satellite imagery. Google Earth Engine (GEE) was used to access and analyze imagery via the Python API (ee library) and the 'geemap' package for interactive mapping. Sentinel-2 Level-2A surface reflectance data (COPERNICUS/S2_SR_HARMONIZED) were acquired for the period from August 1 to December 31, 2023. The image collection was filtered to include only scenes with less than 10% cloud cover intersecting the study area. After cloud masking, all spectral bands were scaled from digital numbers to physical reflectance by dividing by 10,000, yielding dimensionless reflectance values ranging from 0 to 1. To generate a representative cloud-free image of the study period, a temporal median composite was created from the filtered image collection. Then, vegetation indices were computed from the median composite to characterize aspects of vegetation condition, and the results were finally exported at 10 m resolution.

The final group, comprising climatic variables, specifically mean annual temperature (temp) and precipitation (precip), was incorporated at a spatial resolution of 1 km. Subsequently, all variables were resampled to a 30-meter pixel raster for spatial analysis and model implementation.

2.2.3 Multicollinearity assessment and feature selection

A Variance Inflation Factor (VIF) analysis was conducted to assess multicollinearity among 19 predictor variables, which may affect the stability of the estimates and the interpretability of the model (49, 50). The VIF quantifies the degree to which the variance of a regression coefficient is inflated due to collinearity with other predictors and was calculated for each variable using the Equation 1:

$$VIF_j = 1/(1 - R_j^2) \quad (1)$$

Where R^2_j represents the coefficient of determination obtained by regressing the j^{th} predictor against all other predictors (51, 52).

Using R software, an iterative backward elimination procedure was implemented to systematically reduce multicollinearity. At each iteration, the predictor with the highest VIF was removed, and VIF values were recalculated for the remaining variables. This process continued until all predictors exhibited VIF ≤ 10, the commonly accepted threshold for multicollinearity in environmental modeling applications (53). The algorithm included safeguards to retain at least three predictors and terminated after a maximum of 20 iterations.

Following multicollinearity removal, Recursive Feature Elimination (RFE) with Random Forest was applied to identify the optimal subset of predictors that maximized model performance (54). RFE is a backward selection wrapper method that iteratively removes the least important features based on model-derived importance scores while evaluating predictive performance through cross-validation (55, 56).

The RFE algorithm was implemented with the following specifications: (i) Random Forest as the base model with 500 trees; (ii) 10-fold cross-validation repeated three times to ensure robust performance estimation; (iii) systematic evaluation of all feature subset sizes ranging from 3 to 13 predictors; and (iv) root mean square error (RMSE) as the optimization criterion. At each iteration, the model was trained using the current feature subset, variable importance scores were computed based on mean decrease in node impurity, and the least important feature was eliminated. Model performance was assessed through cross-validation for each subset size, and the optimal number of features was determined as the subset that minimized RMSE while balancing model parsimony. Within the workflow implementation, the application of VIF+RFE for clarification regarding variable retention and elimination parameters at each stage can be visualized in Figure 2.

2.2.4 Data preprocessing and task definition

Environmental predictor values were extracted at soil sampling locations using the terra package. The extraction process linked the

predictor values to the soil sample points, creating a comprehensive dataset for model training. A spatial regression task was implemented using the mlr3spatial framework (57), with SOC content as the target variable. Coordinates were stored as spatial metadata but were not used as predictive features to ensure proper spatial modeling practices and avoid data leakage during spatial cross-validation. All predictor variables were standardized using z-score standardization (subtracting the mean and dividing by the standard deviation) to ensure equal contribution across variables with different scales and units. The scaling parameters were stored for consistent application to prediction rasters.

2.3 Machine learning algorithms

Four machine learning algorithms were implemented for SOC content prediction: Support Vector Regression (SVR), Artificial Neural Network (ANN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). To enhance model performance, systematic hyperparameter tuning using a grid search with spatial cross-validation was employed, as illustrated in Figure 2.

2.3.1 Support vector regression

The Support Vector Regression (SVR) is a supervised learning algorithm grounded in statistical learning theory to find a function that approximates the relationship between predictors and a continuous response while maximizing margin tolerance (58). To achieve this, it constructs an optimal hyperplane that maximizes the margin between groups, allowing it to generalize effectively even when the dataset is limited or exhibits complex relationships (59). One of its greatest strengths lies in the use of kernel functions, which "transform" the data into higher-dimensional spaces and facilitate the accurate and robust identification of nonlinear patterns (60). Moreover, its mathematical formulation—based on convex optimization problems—ensures global solutions and reduces the risk of overfitting, making it a robust and versatile technique for both classification and regression tasks (61).

A SVR model was implemented using the e1071 package (62) with a radial basis function (RBF) kernel. The RBF kernel was selected for its ability to capture non-linear relationships between soil properties and environmental predictors (63, 64). Initial hyperparameters were set as follows: regularization parameter (C) = 10, kernel coefficient (γ) = 0.1, and epsilon (ε) in the epsilon-insensitive loss function = 0.1. These values were subsequently optimized through systematic hyperparameter tuning.

2.3.2 Random forest

Random Forest (RF) is a supervised learning technique based on an ensemble of decision trees, introduced by (65). It combines two fundamental strategies for building an ensemble of decision trees: Bootstrap Aggregating (Bagging) and random feature selection. Once the Random Forest model has been trained, the prediction stage is based on the principle of assembling multiple base classifiers (decision trees). In a new instance, each tree in the previously trained ensemble outputs an independent prediction; these predictions are combined using an aggregation mechanism that depends on the task type: classification or regression. In regression, the arithmetic mean of the individual predictions from each tree is calculated, yielding a smoothed estimate (66).

A RF regression model was implemented using the ranger package (67), which provides a fast C++ implementation of the original algorithm (65). Initial configuration included 500 trees, with the number of variables randomly sampled at each split (*mtry*) set to [$\sqrt{p}$] where 'p' represents the number of predictor variables, following Breiman's recommendation for regression tasks. The minimum node size was set to 5 to prevent overfitting. Feature importance was calculated using the Giny impurity measure to identify the most influential predictors.

2.3.3 Artificial neural network

ANNs are mathematical models inspired by the functioning of the human brain. These networks consist of interconnected artificial "neurons" that learn from data by iteratively adjusting the weights of their connections through algorithms such as backpropagation (68). Owing to their architecture, ANNs have the ability to recognize complex and nonlinear relationships, even in datasets that are large, multidimensional, or affected by a certain degree of noise (69). This structural flexibility allows them to adopt multiple configurations—such as multilayer perceptrons (MLP), convolutional neural networks (CNN), or recurrent neural networks (RNN)—which facilitates their application across a wide range of problems and data types (70). The versatility of artificial neural networks has promoted their use in agricultural sciences, where they are employed to model crop processes, optimize farm management, and support decision-making (71).

A feed-forward neural network with one hidden layer was implemented using the 'nnet' package (72). The network architecture was constrained to use linear output activation (linout = TRUE), which is required for regression tasks in the 'nnet' implementation rather than a tunable hyperparameter. This parameter controls the output-layer activation function and must be set to TRUE for continuous-variable prediction.

2.3.4 Extreme gradient boosting

It is a machine learning algorithm based on the gradient boosting method, designed to build robust and accurate predictive models through the sequential ensemble of multiple decision trees (73). Its operation relies on the iterative optimization of a loss function using gradients, incorporating L1 and L2 regularization techniques that reduce overfitting and enhance generalization capability (74). Furthermore, its efficient architecture allows it to handle large volumes of data and correlated variables, integrating strategies such as column subsampling and tree pruning. Altogether, these properties have made this algorithm a powerful tool for classification, regression, and environmental and agricultural modeling, noted for its ability to capture nonlinear and complex relationships within data (75).

An XGBoost regression model was implemented using the xgboost package (73). Initial hyperparameters were configured as

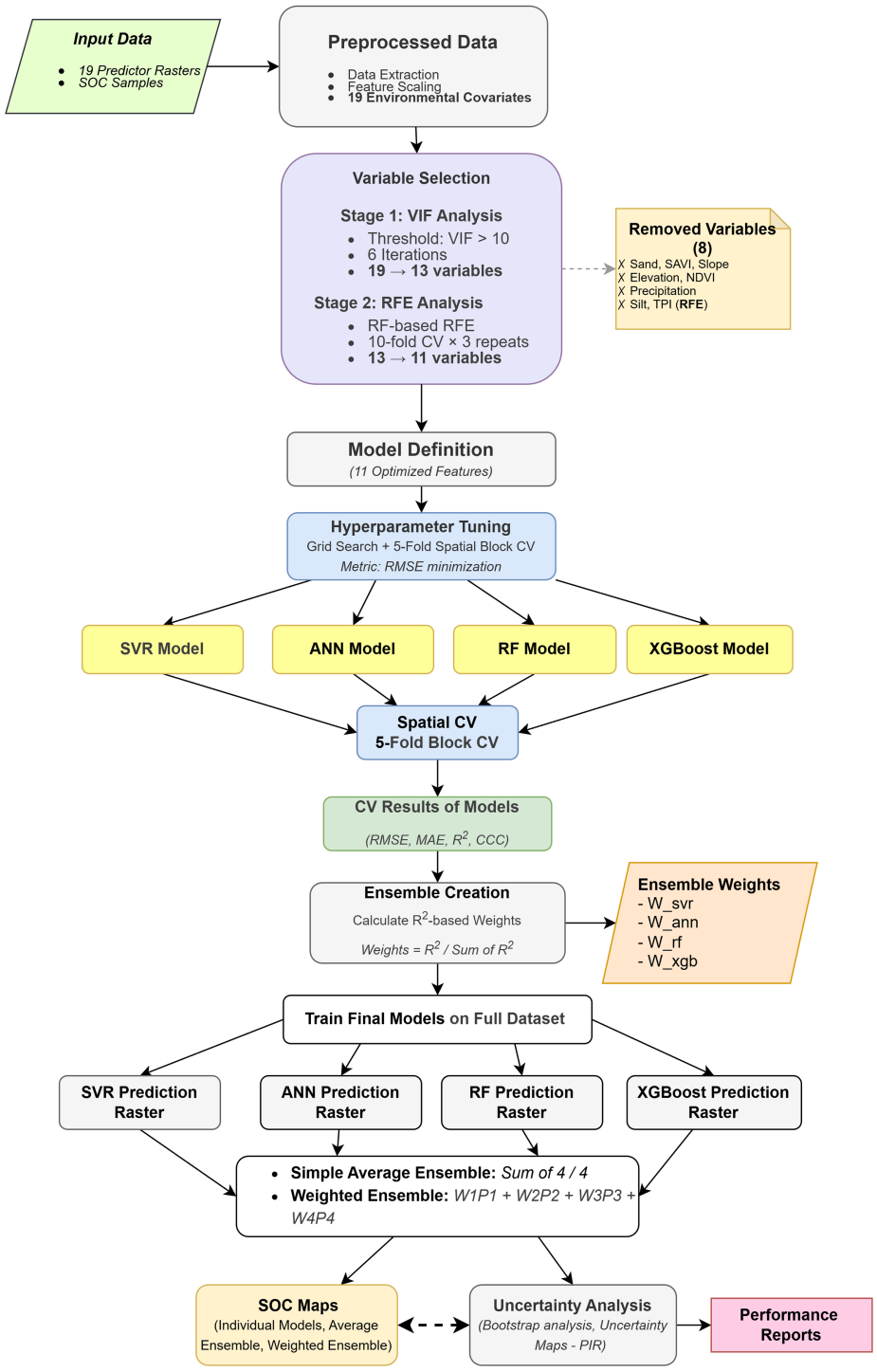

FIGURE 2
Flowchart of the SOC digital mapping framework with VIF-RFE variable selection and hyperparameter-tuned ensemble models.

follows: number of boosting rounds (nrounds) = 100, learning rate (η) = 0.1, maximum tree depth = 6, subsample ratio = 0.7, and column sampling ratio (colsample_bytree) = 0.7. These parameters were subsequently optimized through grid search.

2.3.5 Ensemble

Ensemble learning combines predictions from multiple models to produce a single, more accurate prediction. The fundamental principle is that diverse models capture different aspects of the underlying pattern, and their combination reduces individual model weaknesses while leveraging their collective strengths.

To leverage the complementary strengths of different algorithms and improve prediction robustness, two ensemble approaches were implemented: simple average ensemble and weighted average ensemble. The ensemble weights were computed from model performance in spatial cross-validation. Specifically, each model's weight was proportional to its R^2 value from spatial

CV, normalized so that weights sum to unity, as shown in Equation 2:

$$w_i = R^2_i / \Sigma(R^2_j) \quad (2)$$

where w_i is the weight for model i, R^2_i is the coefficient of determination for model i, and the summation is across all four models (j ∈ [SVR, ANN, RF, XGBoost]).

To prevent numerical instability, a minimum R^2 threshold of 0.1 was imposed. Models with $R^2 < 0.1$ were assigned $R^2 = 0.1$ for weight calculation purposes, ensuring all models contributed to the ensemble even if one performed poorly.

In the ensemble prediction, for each pixel in the study area, predictions from all four models were combined using both approaches: a simple average ensemble, which assigns equal weights (0.25 each) to all four models, and a performance-weighted ensemble. Ensemble weights were determined based on spatial cross-validation (CV) R^2 performance. Individual model R^2 values were normalized to be at least 0.1, ensuring non-negative weights, and these weights were calculated as each model's R^2 proportion of the total R^2 from all models. The final weights summed to one for proper ensemble integration. Final models were trained on the complete dataset after CV to maximize the use of available training data for prediction generation.

2.4 Statistical indicators of model efficiency

Models performance were evaluated using spatial block cross-validation to account for spatial autocorrelation in soil data (76, 77). Conventional random k-fold cross-validation can lead to overoptimistic performance estimates in spatially structured data due to information leakage between nearby training and test observations (78).

Spatial block cross-validation was implemented using the mlr3spatiotempcv package (79). The resampling strategy employed spatial block partitioning (spcv_block) with a five-fold configuration. A 3×3 grid was used to define the block structure, yielding nine spatial blocks that were randomly assigned to the five folds (Figure 3). This approach ensured that the training and test sets remained geographically separated throughout the validation process.

For each fold iteration, the model was trained on 4/5 of the spatial blocks and validated on the remaining 1/5. This process was repeated five times, with each fold serving once as the validation set. Predictions and observed values from all folds were aggregated to calculate overall performance metrics (RMSE, MAE, and R^2). Additionally, Lin's Concordance Correlation Coefficient (CCC) was calculated to assess the agreement between paired continuous measurements (80).

2.5 Uncertainty calculation

2.5.1 Bootstrap sampling strategy

A bootstrap uncertainty analysis was conducted through 10 sequential processing iterations to improve reliability when working with complex spatial objects. The bootstrap procedure comprised four key steps repeated across all iterations: first, random sampling with replacement was performed from the original training dataset; second, computationally efficient models, including linear regression, Random Forest, and decision tree algorithms, were trained; third, spatial predictions were generated across the entire study area; and finally, comprehensive quality control measures applied to ensure result validity.

2.5.2 Spatial validation and quality control

Extensive validation procedures were applied to ensure the reliability of bootstrap predictions through a multi-step approach. The validation process began with spatial consistency checks to verify that the extent and coordinate reference systems matched across all predictions. Subsequently, quality filtering criteria were implemented, requiring a minimum of 30% valid pixel values per prediction to maintain data integrity. For ensemble validation, a threshold was established, requiring at least five valid bootstrap predictions to proceed with uncertainty analysis. Throughout this process, robust error-handling protocols were incorporated, systematically excluding failed predictions while generating informative warnings to document any issues encountered.

2.5.3 Uncertainty metric calculation

Uncertainty metrics were calculated from the bootstrap prediction ensemble in order to perform a comprehensive assessment of model reliability. The analysis began with a calculation of the standard deviation to quantify the absolute uncertainty within the SOC content predictions, which offered a direct measurement of prediction variability. The coefficient of variation was then computed to express relative uncertainty as a percentage of the predicted value, enabling standardized comparisons across different prediction magnitudes. Additionally, 90% prediction intervals were determined using lower and upper bounds at the 5th and 95th percentiles, respectively (Equation 3), to capture the range of possible values. To assess the range of the prediction interval and aid interpretation of the magnitude of uncertainty, the absolute and relative widths of these 90% confidence intervals were ultimately calculated. With these values, the Prediction Interval Ratio (PIR) was implemented to provide a more robust assessment of prediction uncertainty.

$$PIR_{50} = \frac{Q_{95} - Q_{05}}{Q_{50}} \quad (3)$$

Where, Q_{50} represents the median (or 50th percentile) of the bootstrap predictions, while Q_{50} and Q_{50}; represent the 5th and 95th percentiles, respectively. Finally, the values are classified according to their confidence level associated with relative uncertainty: PIR < 0.1 (high confidence, < 10%), 0.1–0.2 (moderate, 10–20%), and 0.2–0.5 (low, 20–50%).

2.6 Assessment of covariate contribution to SOC prediction

To assess the contribution of OK-interpolated soil properties to the ML-based SOC prediction, two modeling scenarios were

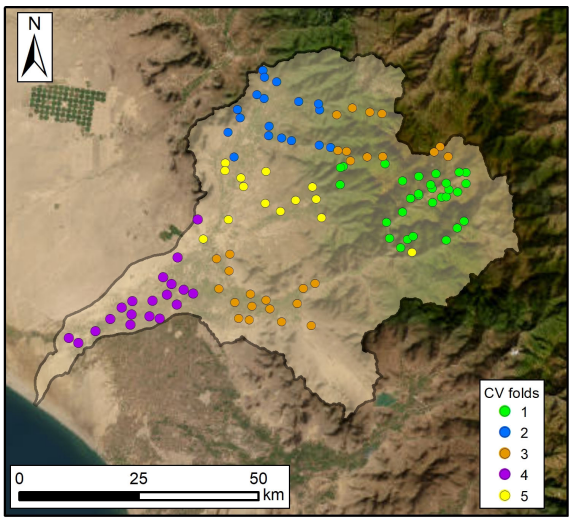

FIGURE 3
Spatial block cross-validation scheme. The 109 sampling points were grouped into five spatial cross-validation folds (CV folds 1–5) based on their spatial distribution.

compared: (A) ML models trained with all 11 covariates, and (B) ML models trained with only the 7 environmental covariates. In both of the analyzed scenarios, each ML algorithm underwent an independent training process that incorporated grid search for hyperparameter optimization, resulting in the evaluation of a comprehensive set of 134 distinct parameter combinations across the various models: 27 (SVR), 27 (ANN), 16 (RF), and 64 (XGBoost). Under the established method, an ensemble approach was used to synthesize the four individual models, after which their performance was meticulously evaluated using the metrics that had been previously mentioned in order to detect the overall efficiency of the integrated models. Similarly, the prediction uncertainty was quantified via bootstrap resampling.

3 Results

3.1 Descriptive statistics of the predictor variables and SOC content

A total of 19 predictive variables were evaluated, including those derived from the DEM, field and laboratory measurements, vegetation indices, and climatic variables. The descriptive analysis of the 109 sampling points is presented in Table 2. Bulk density (BD) ranged from 0.96 to 1.92 g cm^{-3}, with an average of 1.50 g cm^{-3}. These BD values indicate some degree of compaction in certain sampling sites. Soil pH varied from 5.5 to 8.1, with a mean of 7.24, indicating conditions ranging from slightly acidic to alkaline, but tending toward neutrality. Electrical conductivity (EC) ranged from 0.88 to 10.23 dS m^{-1}, with a mean of 1.61 dS m^{-1}, suggesting localized areas affected by salinity problems. Regarding texture, soils were dominated by the sand fraction (55.2% on average), followed by silt (27.8%) and clay (17.0%), revealing marked spatial variability and a predominance of loam to sandy loam textures. For vegetation indices, NDVI values ranged from 0.10 to 0.85, with a mean of 0.48 across the evaluated points, similar to the trends observed for EVI and GNDVI. Among the topographic variables, slope ranged from 0 to 0.62, with an average of 0.1672; the gentler areas were located in the lower and middle parts of the watershed, while steeper slopes were found in the upper regions. Concerning soil organic carbon (SOC) content, values ranged from 0.42 to 0.87%, with an average of 0.65%, indicating generally low concentrations across the study area.

3.2 Spatial interpolation performance

For soil properties, Table 3 provides a detailed overview of the goodness-of-fit statistics for the ordinary kriging models, specifically the R^2, RMSE, and MAE. By utilizing leave-one-out cross-validation, the findings indicated considerable variability in the precision of interpolation when applied to different soil properties. Soil pH exhibited the highest predictive accuracy (R^2 = 0.84, RMSE = 0.32), with a Gaussian variogram model showing strong spatial structure (nugget ratio = 7.7%). Electrical conductivity (EC) and clay content showed moderate performance (R^2 = 0.34 and 0.24, respectively), while bulk density (BD), silt, and sand content exhibited weak spatial autocorrelation with R^2 values below 0.16. The high nugget-to-sill ratio observed for BD (95.9%) indicates that spatial variation occurs predominantly at scales finer than the sampling interval or reflects measurement uncertainty.

3.3 Covariate correlation structure and feature selection

The correlation matrix is shown in Figure 4. The observed variable (SOC) exhibited correlations lower than 0.1 with aspect, TWI, pH, clay, slope, TRI, TPI, BD, and EC. The highest correlations were found with NDVI (0.24), GNDVI (0.23), SAVI (0.21), and EVI (0.20), all of which belong to the group of vegetation indices. Elevation and precipitation showed correlations of 0.12 and 0.13, respectively. Temperature displayed a negative correlation (-0.12), as did the sand percentage (-0.16); these values are expected, since higher sand content generally corresponds to lower SOC levels. Overall, the correlations between SOC and the predictive variables were weak, not exceeding 0.24 (positive) or -0.16 (negative). SOC exhibited significant correlations ($P < 0.05$) with EVI, GNDVI, NDVI, and SAVI, while no significant relationships were observed with the remaining variables.

For feature selection, the VIF analysis identified six highly collinear predictors (sand, SAVI, slope, elevation, NDVI, and precipitation) among the initial 19 covariates. Subsequent RFE optimization identified 11 variables as the optimal subset, achieving RMSE = 0.105 and R^2 = 0.110 through 10-fold cross-validation with three repeats. This represented a 42% reduction in predictor dimensionality while improving model performance by 6.6% (R^2) compared to the full covariate set (Figure 5).

TABLE 2 Descriptive statistics of soil organic carbon (SOC) content and predictor variables used for model development.

Predictor	Unit	Minimum	Maximum	Mean	Range	SD	Variance	Skewness	Kurtosis
aspect	–	0.0524	6.2431	3.3324	6.1908	1.7179	2.951	-0.3521	-0.9631
BD	g cm^{-3}	0.9609	1.9181	1.5064	0.9572	0.1624	0.0264	-0.5591	1.7337
Clay	%	4.0013	55.9303	17.001	51.929	11.1848	125.1005	0.8239	0.1896
EC		0.88	10.228	1.6122	9.3479	1.1488	1.3197	4.8188	29.4943
elevation	m	15	3684.68	898.06	3669.68	1075.66	1157049.87	1.062	-0.2609
EVI	–	0.0734	0.7177	0.3064	0.6442	0.1371	0.0188	0.4493	-0.3703
GNDVI	–	0.2066	0.7577	0.4932	0.5511	0.1316	0.0173	-0.024	-0.7387
hillshade	–	0.3241	0.9398	0.6951	0.6157	0.1124	0.0126	-0.858	1.9306
NDVI	–	0.1046	0.8507	0.4779	0.7461	0.1958	0.0383	0.0545	-0.9775
pH	–	5.5017	8.0995	7.2422	2.5978	0.7985	0.6376	-0.7152	-1.1141
precip	mm $year^{-1}$	29.1578	1082.11	311.09	1052.95	274.79	75508.388	0.9063	-0.348
Sand	%	19.0038	90.9804	55.1828	71.9765	15.6439	244.7315	-0.1009	-0.4908
SAVI	–	0.0722	0.6241	0.2939	0.5519	0.1229	0.0151	0.2728	-0.6621
Silt	%	2.0343	71.9305	27.8163	69.8962	14.3947	207.2064	0.7102	-0.0058
slope	–	0	0.6215	0.1675	0.6215	0.1774	0.0315	0.6059	-1.112
temp	°C	9.4178	23.0456	20.0608	13.6277	3.7953	14.404	-1.2519	0.2649
TPI	–	-3.5589	4.3628	-0.023	7.9217	1.1118	1.2361	0.6312	4.1531
TRI	–	0	15.936	4.1433	15.936	4.4012	19.371	0.6935	-0.8901
TWI	–	4.4492	17.0625	7.9445	12.6133	2.6439	6.99	1.3528	1.9182
SOC	%	0.42	0.87	0.649174	0.45	0.105533	0.011137	0.065826	-0.63183

BD, bulk density; EC, electrical conductivity; EVI, Enhanced Vegetation Index; GNDVI, Green Normalized Difference Vegetation Index; NDVI, Normalized Difference Vegetation Index; SAVI, Soil-Adjusted Vegetation Index; TPI, Topographic Position Index; TRI, Topographic Ruggedness Index; TWI, Topographic Wetness Index.

3.4 Spatial prediction of SOC content

The spatial prediction of SOC content across the different models is shown in Figure 6, providing a reference for identifying the areas with higher concentrations. The highest SOC values, exceeding 0.8%, were observed in the upper part of the watershed in five of the six evaluated models. In contrast, the XGBoost model displayed lower concentrations (<0.5%) in these same upper areas. The RF model exhibited the least spatial variation in SOC, with values ranging between 0.6 and 0.8% (Figure 6A). The ensemble models showed the greatest spatial variability, with SOC concentrations ranging from below 0.5% to above 0.8% (Figure 6B).

3.5 Performance of individual and ensemble models

The coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and Lin's concordance correlation coefficient (CCC) were used to evaluate the performance of the models. The SVR algorithm (Figure 7A) achieved an R^2 of 0.004, with MAE, RMSE, and CCC values of 0.09, 0.01, and 0.005, respectively, for the validation set. The Artificial Neural Network (ANN) model exhibited the lowest predictive performance, with an R^2 value of 0.14 (Figure 7B), indicating a limited ability to explain the variability of SOC content. In contrast, the XGBoost algorithm

TABLE 3 Fitted variogram models and leave-one-out cross-validation results for the spatial interpolation of soil physicochemical properties.

Property	Variogram model	Nugget	Sill	Range	Nugget ratio	Cross validation		
						RMSE	MAE	R^2
BD	Sph	0.03	0.03	91,612.64	95.90	0.16	0.12	0.05
pH	Gau	0.06	0.79	24,803.14	7.70	0.32	0.23	0.84
EC	Gau	0.00	0.98	1,947.08	0.00	0.96	0.48	0.34
Clay	Gau	28.34	103.87	3,032.17	27.30	9.70	7.88	0.24
Silt	Exp	0.00	215.37	3,592.27	0.00	13.4349	10.49	0.15
Sand	Sph	91.46	256.91	11,343.75	35.60	15.1761	12.42	0.08

Soil properties include BD, bulk density (g cm^{-3}); EC, electrical conductivity (dS m^{-1}); and clay, silt, and sand contents (%). Cross-validation performance metrics include RMSE, root mean square error; MAE, mean absolute error; and R^2, coefficient of determination.

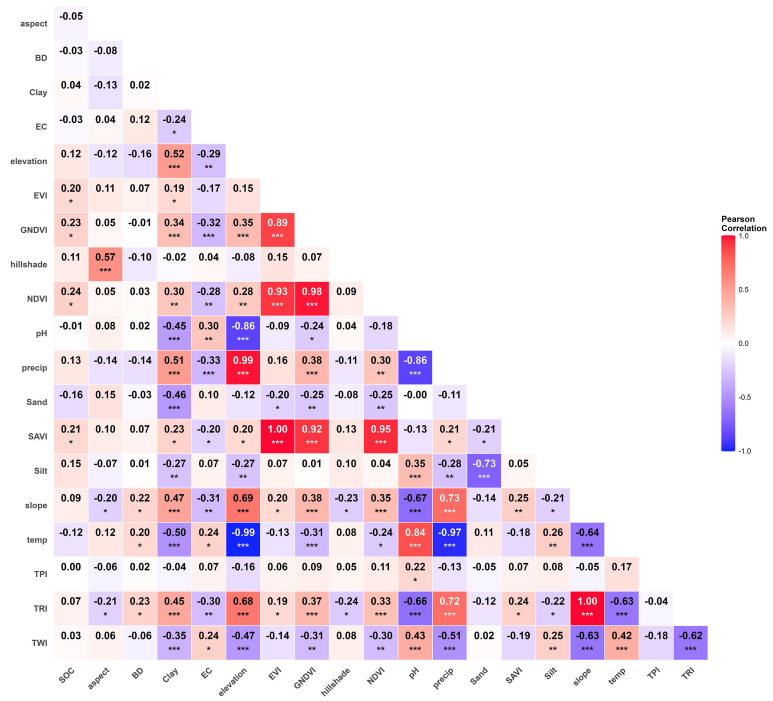

FIGURE 4
Correlation of predictive variables with SOC. The presence of *, indicates the significance level of the correlation (*p<0.05; **p<0.01; ***P<0.001).

showed strong predictive performance, achieving an R^2 of 0.83, a mean absolute error (MAE) of 0.03, a root mean square error (RMSE) of 0.002, and a Lin's concordance correlation coefficient (CCC) of 0.88 (Figure 7D). Random Forest also demonstrated adequate predictive capability (Figure 7C), with R^2, MAE, RMSE, and CCC values of 0.63, 0.05, 0.004, and 0.69, respectively. Among the ensemble approaches, the best performance was obtained with the weighted ensemble model, which achieved an R^2 of 0.70

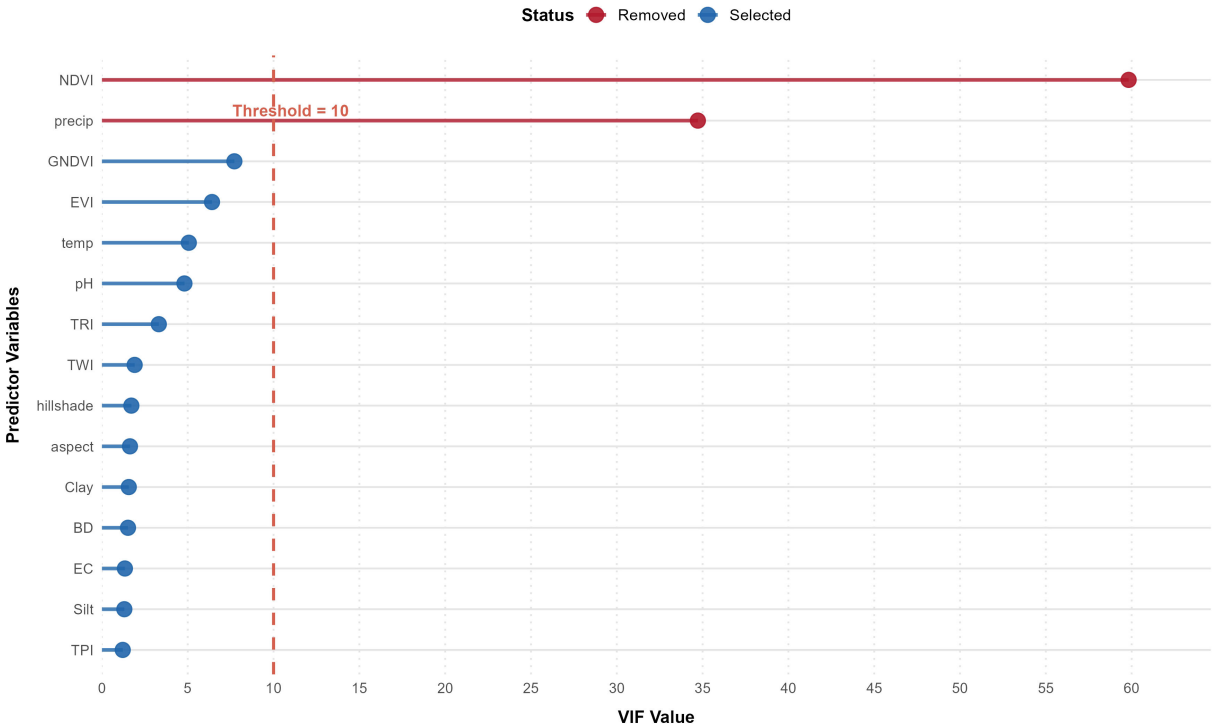

FIGURE 5
VIF values for predictor variables below 10. The red vertical dashed line represents the threshold for variable removal. (Other variables are not shown because higher VIF values reduce the relative bar size of those with lower VIF values).

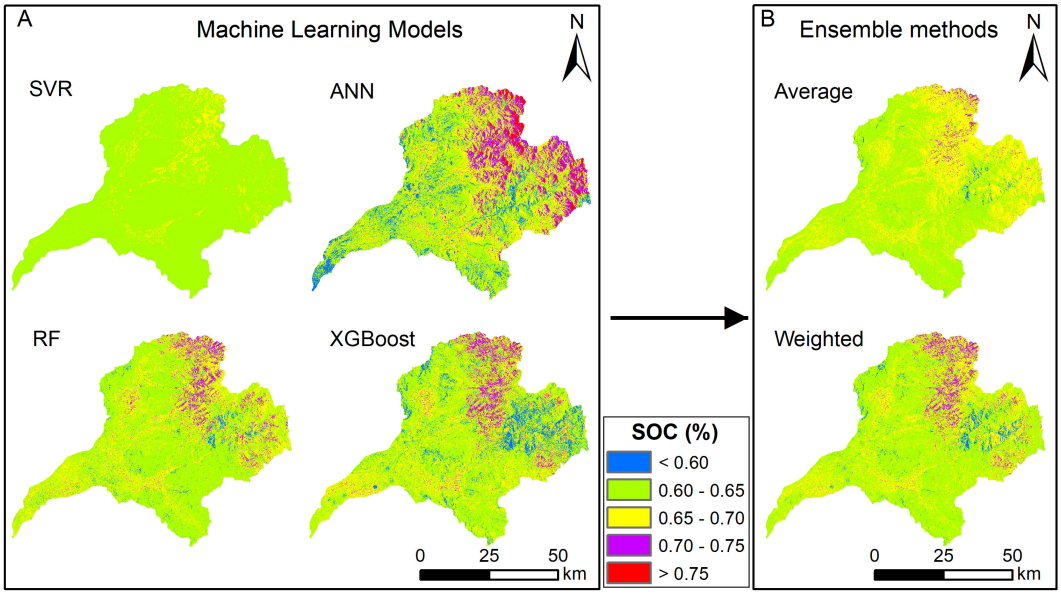

FIGURE 6
Spatial distribution of soil organic carbon (SOC, %) predicted using different machine learning models and ensemble methods. **(A)** Individual model results: Support Vector Machines (SVR), Artificial Neural Networks (ANN), Random Forest (RF), and XGBoost. **(B)** Ensemble methods: simple average (Average) and weighted average (Weighted).

(Figure 7F), compared with the simple average ensemble that reached an R^2 of 0.49 (Figure 7E). According to the results, the XGBoost and weighted ensemble models provided the most accurate predictions of SOC.

3.6 Spatial uncertainty of soil organic carbon content predictions

The spatial distribution of prediction uncertainty, expressed as PIR_{90}, ranged from 0.023 to 8.516 across the study area (Figure 8), with a mean value of 0.142 ± 0.044. Approximately 14.89% of the study area exhibited low uncertainty (PIR < 0.1) or high prediction level, predominantly in agricultural areas and in uncultivated areas in the lower part of the basin. The majority of the area, specifically 75.09%, exhibited moderate uncertainty (PIR = 1.0 - 0.2), which was the most common form and was distributed over all land cover types within the basin. Finally, high uncertainty (PIR > 0.2) was observed in 10.02% of the area located in the top zone reflecting variability in environmental conditions and maybe due to sparse sampling.

3.7 Variable importance

The most important variable for the SOC model was the EVI index, followed by GNDVI, temperature, TRI, and pH (Figure 9). The remaining eight variables showed lower importance under the environmental conditions of the study area. Among them, Silt and TWI were the least influential variables for the models. EVI accounted for approximately 8.75 % of the total importance, GNDVI for 6.30%, temperature for 6.25 %, TRI for 5.5%, and pH for 5.1%. Altogether, these five variables contributed about 31.9% of the total importance. The remaining 68.1% corresponded to other variables derived from the DEM and soil texture.

3.8 Impact of covariate selection on model performance

The Table 4 showed how the scenario A (all 11 covariates), the weighted ensemble achieved an RMSE of 0.057, MAE of 0.046, and R^2 of 0.70, representing a 45.6% RMSE reduction compared to the best individual model (SVM, RMSE = 0.105). Among individual models, XGBoost exhibited the highest explanatory power (R^2 = 0.83), while SVM attained the lowest RMSE but poor variance explanation (R^2 = 0.004). The ensemble prediction uncertainty was low, with a coefficient of variation of 5.07% and a mean 90% prediction interval width of 0.093 SOC units. In Scenario B (without soil properties), the weighted ensemble achieved an RMSE of 0.059, MAE of 0.051, and R^2 of 0.68, with individual model performance remaining comparable (SVM RMSE = 0.102, RF RMSE = 0.106, XGBoost RMSE = 0.104). The comparison between scenarios revealed that excluding the OK-interpolated soil properties resulted in negligible differences in ensemble prediction accuracy (RMSE increase of only 0.002), indicating that the environmental covariates alone captured the spatial variability of SOC effectively.

Mean PIR values of 0.142 (A) and 0.143 (B) were found in the uncertainty evaluation of the weighted ensemble model. This means the width of the 90% prediction interval is about 14% of the predicted SOC. Additionally, most of the study area, ranging from 75% (A) to 77% (B), was categorized as moderate, and the high-prediction-level classification showed only 1.09% variation across scenarios (14.89-13.80%).

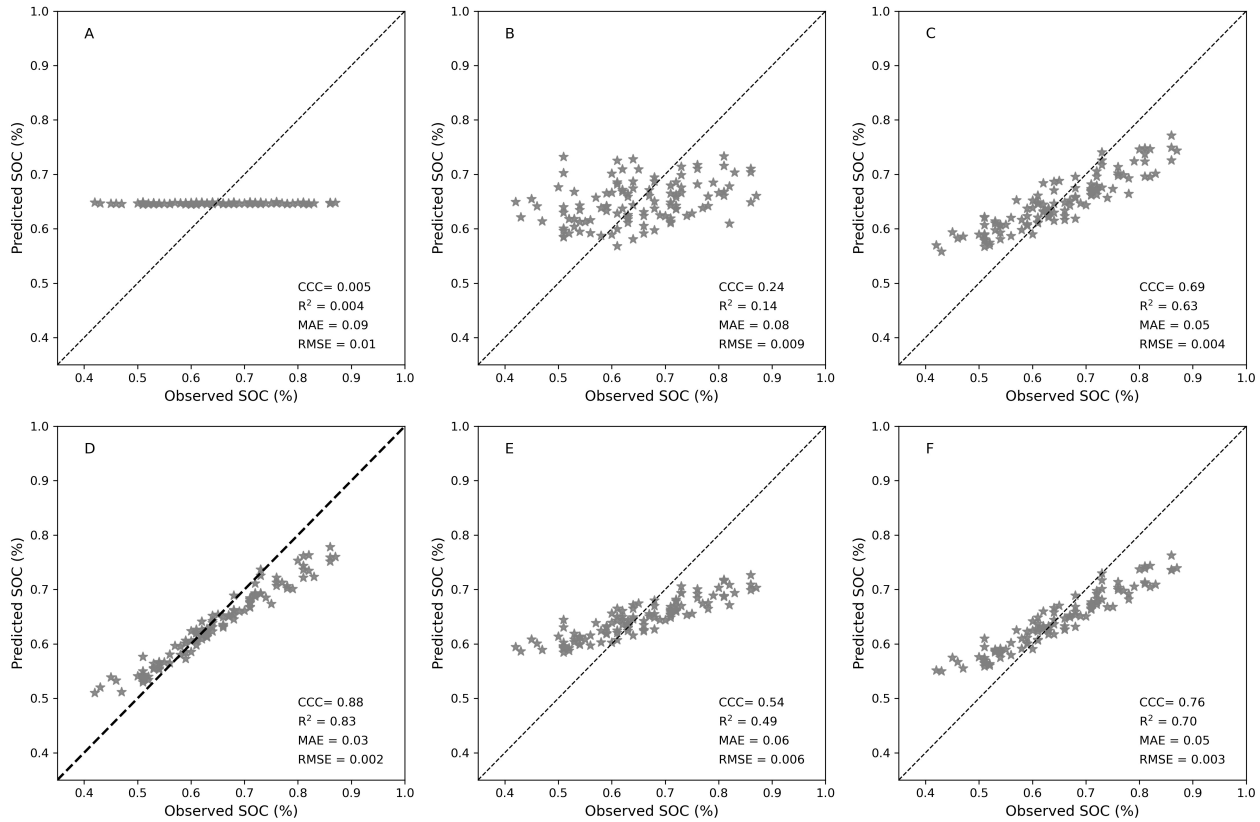

FIGURE 7
Comparison between observed and predicted SOC values across the different models: **(A)** SVR, **(B)** ANN, **(C)** Random Forest, **(D)** XGBoost, **(E)** Ensemble Average, and **(F)** Ensemble Weighted.

4 Discussions

4.1 Variability of SOC content concentrations and their prediction

SOC values obtained in the laboratory ranged from 0.42 to 0.87%, with a mean of 0.65%. The best-performing model (XGBoost) estimated a range of 0.50 to 0.79%, with a mean of 0.64%. Although differences were observed in the ranges, the mean values are very similar, with a variation of only 0.01% compared to the laboratory-derived mean. These values are higher than those reported by (81) for an arid zone with an average annual precipitation of 229.2 ± 59.5 mm under olive cultivation. According to (82), soils require adequate organic matter to maintain fertility and functionality; therefore, the SOC levels recorded in the Motupe River basin are considered low, reflecting soils with limited carbon reserves that may reduce their water and nutrient retention capacity. The low SOC concentrations may also be associated with the high sand content (83), which in the study area reached up to 90%, with an average of 55.18%, while silt and clay contents were 27.8% and 17%, respectively. The spatial distribution of SOC showed that higher concentrations (> 0.8%) were mainly found in the upper part of the basin, likely due to higher annual precipitation (406.93 mm year^{-1}). The correlation of precipitation with NDVI in the study basin was 0.30, with a significance of 0.01 (84). reported that total organic carbon and nitrogen increase with altitude, and that SOC is positively correlated with both annual mean temperature and precipitation gradients. It may also be due to the presence of areas with pine forestation (*Pinus sylvestris*), which can increase the SOC values. Changes of soil carbon along precipitation gradients in three typical desert vegetation types were observed (85). In some upper areas, however, low SOC concentrations were also detected, likely due to steep slopes, pasture burning, and agricultural activities such as grazing and crop cultivation. In contrast, the lower part of the basin showed a more orange-toned zone, indicating SOC concentrations between 0.6 and 0.7%. This area corresponds to the Pomac Historical Sanctuary, a dry forest dominated by *Neltuma pallida* (algarrobo), which is protected by the Peruvian government.

4.2 Effect of variables on SOC content prediction

The correlation between SOC concentration and the predictive variables was generally low; soil texture components (sand, silt, and clay) showed correlation coefficients not exceeding 0.16. This contrasts with the findings of (86), who reported higher correlations of 0.67 and 0.79 for clay and silt, respectively. Although both studies found a negative relationship between sand content and SOC, the magnitude differed: while (86) reported a correlation of –0.79, this study found only –0.16. These results are consistent with the variable importance analysis, where silt and clay

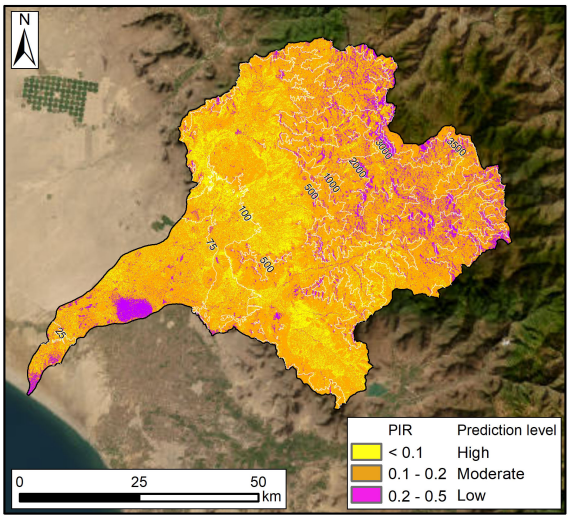

FIGURE 8
Prediction Interval Ratio (PIR) map of the Motupe River basin. Values are classified according to their confidence level associated with relative uncertainty: PIR < 0.1 (high confidence, < 10%), 0.1–0.2 (moderate confidence, 10–20%) and 0.2–0.5 (low confidence, 20–50%).

content were identified as having the lowest importance. This discrepancy may be attributed to the spatial variability and textural heterogeneity across the study area. In (87), precipitation, temperature, and DEM variables were used to predict SOC using regression kriging, achieving an R^2 of 0.98. The present study similarly found temperature and DEM-derived variables, particularly TPI to be important predictors of SOC However, precipitation was excluded due to high collinearity with other predictors. Along with precipitation, NDVI, elevation, sand percentage, SAVI, and slope were also removed. The careful selection of predictor variables is essential to enhance model accuracy and eliminate redundant information (88).

The consistent ranking of vegetation indices (EVI, GNDVI) across the RF indicated strong vegetation-SOC relationships in the study area. Vegetation indices derived from remote sensing data are effective proxies for representing vegetation status, which correlates with soil properties including SOC, due to the role of vegetation in carbon inputs to soil through litter and root biomass (89). The importance of temperature and pH reflected climatic and chemical controls on SOC processes. Temperature influences SOC primarily through its effect on the rates of organic matter decomposition and microbial activity. Higher temperatures generally accelerate decomposition, potentially reducing SOC stocks, whereas cooler temperatures slow decomposition and can promote SOC accumulation (90–92). Topographic variables (TPI, TRI, TWI) demonstrated the significance of landscape position in controlling SOC distribution through effects on moisture, erosion, and deposition processes. Landscape topography controls soil formation and properties by regulating gravity-driven soil movement induced by runoff and tillage activities, thereby influencing soil redistribution and SOC patterns (93). Empirical evidence supports this mechanism, with hillslope curvature and aspect alone explaining up to 94% of fine-scale SOC variation within catchments when high-resolution digital elevation models and complete soil profiles are employed (94). In a previous study, TWI has been identified as the predominant variable in regulating SOC density in agricultural fields, as evidenced by its capacity to explain over 62% of the variation in SOC density across the field locations (95).

4.3 Analysis of the prediction models

The current trend in SOC content assessment involves the use of machine learning algorithms combined with remote sensing, mainly to address issues of temporal and spatial variability (96). In this study, four models (SVR, RF, XGBoost, and ANN) and two ensemble approaches based on these models were used: one employing the simple average of the four algorithms and another determined by weighted model assignment. When ranking the models according to their R2 values, the following order was obtained: XGBoost (R2 = 0.83; MAE = 0.03; RMSE = 0.002; CCC = 0.88) > weighted ensemble (R2 = 0.70; MAE = 0.05; RMSE = 0.003; CCC = 0.76) > RF (R2 = 0.63; MAE = 0.05; RMSE = 0.004; CCC = 0.69) > simple average ensemble (R2 = 0.49; MAE = 0.06; RMSE = 0.006; CCC = 0.54) > ANN (R2 = 0.14; MAE = 0.08; RMSE = 0.009; CCC = 0.24) > SVR (R2 = 0.00; MAE = 0.09; RMSE = 0.011; CCC = 0.01). The weighted ensemble, RF, and simple average ensemble models outperformed those reported by (26), with R2 values of 0.60 and 0.57 for RF and XGBoost, respectively. Similarly (97), reported good performance of the RF algorithm, with an RMSE of 3.32, a value notably higher than that obtained in the present study (0.004). Likewise, when comparing the performance of RF with the results reported by (98), the model developed in this study showed better performance in terms of error, with an RMSE

TABLE 4 Performance comparison of ML models under two covariate scenarios: Scenario A (all 11 covariates including OK-interpolated soil properties) and Scenario B (7 environmental covariates only). Metrics derived from 5-fold spatial block cross-validation.

Model	RMSE		MAE		R^2	
	A	B	A	B	A	B
SVR	0.105	0.102	0.086	0.086	0.004	0.806
ANN	0.113	0.112	0.092	0.091	0.136	0.102
RF	0.108	0.106	0.090	0.088	0.626	0.616
XGBoost	0.111	0.104	0.090	0.085	0.833	0.478
Ensemble (Average)	0.075	0.069	0.062	0.058	0.485	0.566
Ensemble (Weighted)	0.057	0.059	0.046	0.051	0.702	0.683

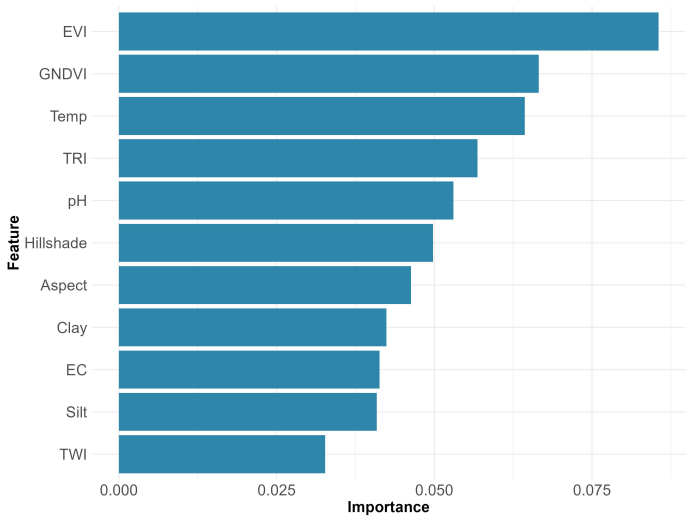

FIGURE 9
Relative importance of the predictor variables for soil organic carbon (SOC) content in the Random Forest model. EVI, Enhanced Vegetation Index; GNDVI, Green Normalized Difference Vegetation Index; Temp, mean temperature; pH, soil hydrogen potential; TPI, Topographic Position Index; Aspect, terrain orientation; TRI, Terrain Ruggedness Index; Hillshade, terrain shading; Clay, clay content; EC, electrical conductivity; Silt, silt content; TWI, Topographic Wetness Index; BD, bulk density.

of 0.52 and an MAE of 0.34; however, the R^2 value reported by those authors (0.76) is higher than that obtained in this study. These findings indicate that the models are suitable for predicting SOC in dry forest watersheds. The low performance of the ANN and SVR models can be attributed to both the limited sample size and the constraints imposed by the 'nnet' package implementation. Specifically, the ANN model was restricted to a single hidden layer architecture with fixed activation functions and lacked batch normalization capabilities. These limitations constrained the model's ability to adequately capture the complex nonlinear relationships inherent in soil-landscape interactions. In the case of XGBoost, the R^2 value obtained in this study was better than that reported by (99), who achieved R^2 = 0.76, RMSE = 4.47,and MAE = 3.91, when estimating carbon stock in the lower Brazos River basin, Texas, USA. Furthermore, the Concordance Correlation Coefficient (CCC) obtained for the XGBoost model (0.88) exceeds the value reported by (100), who achieved a CCC of 0.73, and whereas for RF that study reported higher values (0.77) than those obtained in this study (0.69), based on satellite imagery and topographic parameters Sun et al. (101) reported CCC values of 0.63 for RF, which are lower than those obtained in this study (0.69), where the model was developed using digital elevation data, remote sensing, and climatic variables. Likewise, study (102) reported CCC values of 0.54 for their models, which are lower than those obtained in the present study. This improvement highlights the robustness of the XGBoost and RF models developed for representing SOC content in dry forest ecosystems, particularly when combined with weighted ensemble modeling strategies, which further enhance predictive consistency and agreement between observed and predicted values.

Another alternative is the use of ensemble modeling. In this study, two ensemble approaches were implemented: the first used the simple average of the four base models (SVR, RF, ANN, and XGBoost), and the second applied a weighted scheme assigning different weights to each individual model. Ensemble models can enhance the robustness and accuracy of predictions by combining multiple algorithms (103, 104). The best ensemble identified in this study was the weighted one, which achieved an R^2 of 0.83. This result is consistent with that reported by (105), who obtained an R^2 of 0.85 using an ensemble composed of the same base models. Another study that evaluated different combinations of models for ensemble prediction of SOC, the best results were obtained when XGBoost and SVR were included—two of the algorithms also used in the present study (106). However, there are also ensemble models that did not outperform individual algorithms, as reported by (107), a similar situation occurred in this study, where the ensemble did not exceed the performance of the XGBoost model, which achieved an R^2 of 0.83. Another combination of algorithms that has shown good performance for SOC estimation in ensemble frameworks are RF and SVR, with reported R^2 values of 0.74 (108).

4.4 Evaluation of uncertainties

The PIR metric provides a robust measure of prediction uncertainty with a mean value of 13.4%, indicating generally reliable SOC predictions across the Motupe study area. This metric quantifies the uncertainty of SOC predictions relative to their predicted magnitude, making it easy to compare uncertainty across different locations and models (109, 110). The classification of uncertainty according to PIR into low, medium, and high levels helps to visualize the reliability of the obtained results. A low prediction confidence indicates high PIR values. These low values may be associated with complex topography, which makes field data collection difficult (111). Additional uncertainty may also arise from the spatial predictive variables used, since not all points were measured directly—for instance, variables such as clay and silt percentages were estimated using kriging interpolation (112).

4.5 Alternatives for improving SOC content concentration

Modeling can help to understand the spatial distribution of SOC across the entire watershed and identify the factors that contribute to its loss—factors that are not only associated with anthropogenic activities but also with the intrinsic characteristics of the dry forest ecosystem. Determining SOC concentration is essential for assessing soil fertility and the environmental sustainability of the ecosystem (113). Likewise, identifying the factors that influence SOC prediction and its spatial distribution is crucial for designing effective improvement strategies (98). The scarce vegetation cover, resulting from low and irregular rainfall and prolonged dry periods, may predispose the watershed to low SOC concentrations. In areas where higher SOC levels would be expected, agricultural activities are commonly developed, which can be affected by erosive processes. This study highlights the overall low SOC content throughout the watershed and underscores the urgent need to implement soil management practices aimed at increasing and conserving SOC. An increase in SOC would enable farmers to reduce their dependence on synthetic fertilizers and better cope with the impacts of climate change. It is also important to acknowledge that certain reforested areas exist in the upper watershed, along with some plots that incorporate infiltration trenches to reduce runoff; however, these efforts remain insufficient and should be further strengthened and expanded.

During compared scenarios with or without soil properties as covariates, the results founded suggest that the OK-interpolated soil properties did not provide substantial complementary predictive information beyond what was already captured by the environmental variables, possibly due to propagated interpolation uncertainty or multicollinearity between the kriged soil variables and the environmental covariates. The similar case occurred when the uncertainty was assessed, where both scenarios yielded nearly identical uncertainty levels (mean PIR ~0.14), indicating that the exclusion of OK-interpolated soil properties did not increase prediction uncertainty.

Despite promising results, this study has limitations. First, the sampling design was restricted by logistical challenges and complex, dense terrain, leading to lower sampling density in less accessible areas. Second, the cost of field campaigns and lab analyses limited the total sample size. While sufficient for calibration, a higher density would better capture short-range variability and reduce uncertainty in heterogeneous areas. Finally, the digital soil mapping relied on available environmental covariates. The lack of certain high-resolution auxiliary data (e.g., specific airborne geophysics or higher-resolution climatic grids) meant models relied on satellite imagery and DEM derivatives, potentially missing micro-scale pedological processes.

Future research should focus on overcoming these barriers by integrating proximal soil sensing (e.g., portable XRF or Vis-NIR spectroscopy) to increase data density at a lower cost, and by incorporating emerging high-resolution remote sensing data to refine the predictor stack.

5 Conclusions

Studying SOC content levels in a dry forest watershed is important since this ecosystem is quite susceptible to climate change and human impact. Low SOC levels mean we must reduce carbon loss and increase soil accumulation. With easily obtainable data such as soil properties, terrain data, vegetation indices, and climatic factors, four of the six machine learning algorithms were able to accurately predict SOC concentrations. Before model implementation, high collinearity variables were eliminated using VIF. This reduced the variables from 19 to 13, simplifying models. The algorithms showing the best statistical performance were XGBoost, ensemble average, and RF, achieving coefficients of determination (R^2) above 0.60. In contrast, ANN and SVR exhibited lower R^2 values of 0.14 and 0.004, respectively. The most important variables for prediction were EVI, GNDVI, temperature, TRI, and pH, which together accounted for approximately 31.9% of importance.

The spatial uncertainty analysis, performed through the Prediction Interval Ratio (PIR), indicated a moderate to high level of confidence in the model predictions. The spatial distribution of SOC content demonstrated the highest concentrations within the upper watershed. Nevertheless, these values are suboptimal, indicating a requirement for management strategies, including reforestation and sustainable agricultural practices. Future research should focus on identifying appropriate management plans for this dry forest ecosystem and determining which land uses most negatively affect SOC content concentrations. This will help decision-makers select the best alternatives for soil conservation.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

WS-C: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft. CC-L: Formal analysis, Methodology, Software, Visualization, Writing – original draft. RC-R: Methodology, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was funded by the CUI 2487112 INIA project “Mejoramiento de los servicios de investigación y transferencia tecnológica en el manejo y recuperación de suelos

agrícolas degradados y aguas para riego en la pequeña y mediana agricultura en los departamentos de Lima, Áncash, San Martín, Cajamarca, Lambayeque, Junín, Ayacucho, Arequipa, Puno y Ucayali.

Acknowledgments

The authors would like to acknowledge the support of Eng. Ivan Vilchez, Eng. Issac Castro, and Bach. Johan Rivas for their contribution during the soil sampling process. The authors also express their gratitude to the LABSAF staff at the Vista Florida Agricultural Experimental Station (INIA) and LABSAF Lima for their support in the analysis of the soil samples.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Brasseur GP, Granier C. Mitigation, adaptation or climate engineering? *Theor Inquiries Law*. (2013) 14:1–20. doi: 10.1515/til-2013-003

2. Levy BS, Patz JA. Climate change. In: Levy BS, Wegman DH, Baron SL, Sokas RK, editors. *Occupational and environmental health*. New York: Oxford University Press (2017). doi: 10.1093/oso/9780190662677.003.0032

3. Cologna V, Meiler S, Kropf CM, Lüthi S, Mede NG, Bresch DN, et al. Extreme weather event attribution predicts climate policy support across the world. *Nat Clim Change*. (2025) 15:725–35. doi: 10.1038/s41558-025-02372-4

4. Panepinto D, Riggio VA, Zanetti M. Analysis of the emergent climate change mitigation technologies. *Int J Environ Res Public Health*. (2021) 18:6767. doi: 10.3390/ijerph18136767

5. Hicks Pries CE, Castanha C, Porras R, Torn M. The whole-soil carbon flux in response to warming. *Science*. (2017) 355:1420–3. doi: 10.1126/science.aal1319

6. Farooqi ZUR, Hussain MM, Qadeer A, Ayub MA. Chapter 2 - Role of carbon cycle in soil productivity and carbon fluxes under changing climate. In: Aftab T, Hakeem KR, editors. *Frontiers in plant-soil interaction*. London: Academic Press (2021). p. 29–48. doi: 10.1016/B978-0-323-90943-3.00017-1

7. Food and Agriculture Organization of the United Nations (FAO). *Captura de carbono en los suelos para un mejor manejo de la tierra. Informes sobre recursos mundiales de suelos* (2002). Available online at: https://www.fao.org/4/y2779s/y2779s05.htm (Accessed September 12, 2025).

8. Zhao B, Li Z, Li P, Xu G, Gao H, Cheng Y, et al. Spatial distribution of soil organic carbon and its influencing factors under the condition of ecological construction in a hilly-gully watershed of the Loess Plateau, China. *Geoderma*. (2017) 296:10–7. doi: 10.1016/j.geoderma.2017.02.010

9. Meena RS, Singh AK, Jatav SS, Rai S, Pradhan G, Kumar S, et al. Chapter 10 - Significance of soil organic carbon for regenerative agriculture and ecosystem services. In: Singh K, Ribeiro MC, Calicioglu Ö, editors. *Biodiversity and bioeconomy*. Amsterdam, Netherlands: Elsevier (2024). p. 217–40. doi: 10.1016/B978-0-323-95482-2.00010-9

10. Zhang T, Li Y, Wang M. Remote sensing-based prediction of organic carbon in agricultural and natural soils influenced by salt and sand mining using machine learning. *J Environ Management*. (2024) 352:120107. doi: 10.1016/j.jenvman.2024.120107

11. Zhu Y, Dong M, Wang X, Chen D, Zhang Y, Liu X, et al. Spatiotemporal distribution characteristics of soil organic carbon and its influencing factors in the loess plateau. *Agronomy*. (2025) 15:2260. doi: 10.3390/agronomy15102260

12. Zhang Y, Zhang G, Pan J, Fan Z, Chen F, Liu Y. Soil organic carbon distribution in relation to terrain & land use—a case study in a small watershed of Danjiangkou reservoir area, China. *Global Ecol Conserv*. (2019) 20:e00731. doi: 10.1016/j.gecco.2019.e00731

13. Vasques GM, Grunwald S, Comerford NB, Sickman JO. Regional modelling of soil carbon at multiple depths within a subtropical watershed. *Geoderma*. (2010) 156:326–36. doi: 10.1016/j.geoderma.2010.03.002

14. Demenois J, Torquebiau E, Arnoult MH, Eglin T, Masse D, Assouma MH, et al. Barriers and strategies to boost soil carbon sequestration in agriculture. *Front Sustain Food Syst*. (2020) 4:37. doi: 10.3389/fsufs.2020.00037

15. Boubehziz S, Khanchoul K, Benslama M, Benslama A, Marchetti A, Francaviglia R, et al. Predictive mapping of soil organic carbon in Northeast Algeria. *CATENA*. (2020) 190:104539. doi: 10.1016/j.catena.2020.104539

16. Shi P, Zhang Y, Li P, Li Z, Yu K, Ren Z, et al. Distribution of soil organic carbon impacted by land-use changes in a hilly watershed of the Loess Plateau, China. *Sci Total Environment*. (2019) 652:505–12. doi: 10.1016/j.scitotenv.2018.10.172

17. Wang L, Li Z, Wang D, Liao S, Nie X, Liu Y. Factors controlling soil organic carbon with depth at the basin scale. *CATENA*. (2022) 217:106478. doi: 10.1016/j.catena.2022.106478

18. Hau NX, Tuan NT, Trung LQ, Chi TT. Estimation of soil organic carbon content using visible and near-infrared spectroscopy in the Red River Delta, Vietnam. *Chemometrics Intelligent Lab Systems*. (2024) 255:105253. doi: 10.1016/j.chemolab.2024.105253

19. Zhang Z, Gao X, Zhang L, Zhang X, Zhao X. Deep learning approach with coupled weighted loss function for estimation and prediction of soil organic carbon in China. *Eur J Soil Science*. (2025) 76:e70189. doi: 10.1111/ejss.70189

20. Ding Z, Liu K, Grunwald S, Smith P, Ciais P, Wang B, et al. Advancing soil organic carbon prediction: A comprehensive review of technologies, AI, process-based and hybrid modelling approaches. *Advanced Science*. (2025) 12:e04152. doi: 10.1002/advs.202504152

21. Tian X, de BS, Simoes R, MS I, Minarik R, Ho YF, et al. Spatiotemporal prediction of soil organic carbon density in Europe (2000–2022) using earth observation and machine learning. *PeerJ*. (2025) 13:e19605. doi: 10.7717/peerj.19605

22. Chen S, Arrouays D, Leatitia Mulder V, Poggio L, Minasny B, Roudier P, et al. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma*. (2022) 409:115567. doi: 10.1016/j.geoderma.2021.115567

23. Agaba S, Ferré C, Musetti M, Comolli R. Mapping soil organic carbon stock and uncertainties in an alpine valley (Northern Italy) using machine learning models. *Land*. (2024) 13:78. doi: 10.3390/land13010078

24. Chinilin A, Savin I. Combining machine learning and environmental covariates for mapping of organic carbon in soils of Russia. *Egyptian J Remote Sens Space Sci*. (2023) 26:666–75. doi: 10.1016/j.ejrs.2023.07.007

25. Emadi M, Taghizadeh-Mehrjardi R, Cherati A, Danesh M, Mosavi A, Scholten T. Predicting and mapping of soil organic carbon using machine learning algorithms in northern Iran. *Remote Sensing*. (2020) 12:2234. doi: 10.3390/rs12142234

26. Mahmoudzadeh H, Matinfar HR, Taghizadeh-Mehrjardi R, Kerry R. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Regional.* (2020) 21:e00260. doi: 10.1016/j.geodrs.2020.e00260

27. Powers JS, Feng X, Sanchez-Azofeifa A, Medvigy D. Focus on tropical dry forest ecosystems and ecosystem services in the face of global change. *Environ Res Lett.* (2018) 13:090201. doi: 10.1088/1748-9326/aadeec

28. Barboza E, Salazar W, Gálvez-Paucar D, Valqui-Valqui L, Valqui L, Zagaceta LH, et al. Cloud computing application for the analysis of land use and land cover changes in dry forests of Peru. *IJEI.* (2024) 7:505–14. doi: 10.18280/ijei.070312

29. Macías CAS, Escobar KM, Sancán GS, Chávez WA, Chóez AM, Cedeño FV, et al. Influencia del gradiente altitudinal sobre la estimación del carbono almacenado en biomasa aérea viva y en suelos del 'Bosque y vegetación protector El Artesan - EcuadorianHands'. *Joa Jipijapa: Ecosistemas.* (2020) 29:1973–3. doi: 10.7818/ECOS.1973

30. Gandhi DS, Sundarapandian S. Soil carbon stock assessment in the tropical dry deciduous forest of the Sathanur reserve forest of Eastern Ghats, India. *J Sustain Forestry.* (2017) 36:358–74. doi: 10.1080/10549811.2017.1308870

31. Dube T, Muchena R, Masocha M, Shoko C. Estimating soil organic and aboveground woody carbon stock in a protected dry Miombo ecosystem, Zimbabwe: Landsat 8 OLI data applications. *Phys Chem Earth Parts A/B/C.* (2018) 105:154–60. doi: 10.1016/j.pce.2018.03.007

32. Teng M, Zeng L, Xiao W, Huang Z, Zhou Z, Yan Z, et al. Spatial variability of soil organic carbon in Three Gorges Reservoir area, China. *Sci Total Environment.* (2017) 599–600:1308–16. doi: 10.1016/j.scitotenv.2017.05.085

33. Durante P, Guevara M, Vargas R, Oyonarte C. Predicting soil organic carbon with different approaches and spatial resolutions for the southern Iberian Peninsula, Spain. *Geoderma Regional.* (2024) 37:e00780. doi: 10.1016/j.geodrs.2024.e00780

34. Akumu CE, McLaughlin JW. Modeling peatland carbon stock in a delineated portion of the Nayshkootayaow river watershed in Far North, Ontario using an integrated GIS and remote sensing approach. *CATENA.* (2014) 121:297–306. doi: 10.1016/j.catena.2014.05.025

35. SENAMHI HSR PISCO. (2025). Available online at: https://iridl.ldeo.columbia.edu/SOURCES/.SENAMHI/.HSR/.PISCO/index.html?Set-Language=es (Accessed September 18, 2025).

36. USEPA. METHOD 9045D. In: *SOIL AND WASTE pH 2004.* Washington, DC, USA (2004). Available online at: https://www.epa.gov/sites/default/files/2015-12/documents/9045d.pdf (Accessed July 31, 2024).

37. ISO (International Organization for Standardization). *ISO 11265:2025 Environmental solid matrices — Determination of the specific electrical conductivity* (2025). Available online at: https://www.iso.org/obp/ui/es/iso:std:iso:11265:ed-2:v1:en (Accessed September 29, 2025).

38. Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). *NOM-021-RECNAT-2000 — Norma Oficial Mexicana Que Esta blece Las Especificaciones de Fertilidad, Salinidad y Clasificación de Suelos.* Estudios: Muestreo y Análisis (2002). Available online at: https://faolex.fao.org/docs/pdf/mex50674.pdf (Accessed October 10, 2025).

39. ISO 10694. *1995 Soil quality - Determination of organic and total carbon after dry combustion (elementary analysis) - Sky Bear Technical Standards* (2025). Available online at: https://www.skybearstandards.com/https://www.skybearstandards.com/documents/10694-soil-quality-determination-of-organic-and-total-carbon-after-dry-combustion-elementary-analysis/.

40. QGIS Development Team. Spatial without Compromise. In: *QGIS web site.* Beaverton (OR): Open Source Geospatial Foundation Project. (2025). Available online at: https://qgis.org/.

41. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Intl J Climatology.* (2017) 37:4302–15. doi: 10.1002/joc.5086

42. Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens Environment.* (2002) 83:195–213. doi: 10.1016/S0034-4257(02)00096-2

43. Rouse JW, Haas RH, Schell JA, Deering DW. Monitoring vegetation systems in the Great Plains with ERTS, in: (1974). Available online at: https://ntrs.nasa.gov/citations/19740022614 (Accessed October 17, 2025).

44. Gitelson AA, Kaufman YJ, Merzlyak MN. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens Environment.* (1996) 58:289–98. doi: 10.1016/S0034-4257(96)00072-7

45. Huete AR. A soil-adjusted vegetation index (SAVI). *Remote Sens Environ.* (1988) 25:295–309. doi: 10.1016/0034-4257(88)90106-X

46. Gobezie TB, Scott SD, Daggupati P, Bedard-Haughn A, Biswas A. Soil data recency: The foundation for harmonizing soil data across time. *J Environ Management.* (2024) 364:121484. doi: 10.1016/j.jenvman.2024.121484

47. Pebesma EJ. Multivariable geostatistics in S: the gstat package. *Comput Geosciences.* (2004) 30:683–91. doi: 10.1016/j.cageo.2004.03.012

48. Hiemstra PH, Pebesma EJ, Twenhöfel CJW, Heuvelink GBM. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Comput Geosciences.* (2009) 35:1711–21. doi: 10.1016/j.cageo.2008.10.011

49. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. . *Ecography.* (2013) 36:27–46. doi: 10.1111/j.1600-0587.2012.07348.x

50. Zheng Z, Liu K, Zhou Y, Debliquy M, Bittencourt C, Zhang C. Response to letter to the editor from Y. Takefuji on "Beyond principal component analysis: Enhancing feature reduction in electronic noses through robust statistical methods. *Trends Food Sci Technology.* (2025) 157:104918. doi: 10.1016/j.tifs.2025.104918

51. O'brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.* (2007) 41:673–90. doi: 10.1007/s11135-006-9018-6

52. Xiao C, Duan A, Tang Y, Tang B, Wang Q, Yang X. Machine learning prediction of summer extreme precipitation days in the middle and lower Yangtze River with SHAP explanation. *Atmospheric Res.* (2026) 330:108614. doi: 10.1016/j.atmosres.2025.108614

53. Hair JF, Black WC, Babin BJ. Multivariate data analysis: a global perspective. In: *Pearson education* Upper Saddle River (NJ): Pearson Education (2010). Available online at: https://books.google.com.pe/books?id=SLRPLgAACAAJ (Accessed July 26, 2025).

54. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* (2018) 19:65. doi: 10.1186/s12863-018-0633-8

55. Kuhn M. Building predictive models in R using the caret package. *J Stat Software.* (2008) 28:1–26. doi: 10.18637/jss.v028.i05

56. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learning.* (2002) 46:389–422. doi: 10.1023/A:1012487302797

57. Schratz P, Becker M, Lang M, Brenning A. mlr3spatiotempcv: spatiotemporal resampling methods for machine learning in R. *J Stat Software.* (2024) 111:1–36. doi: 10.18637/jss.v111.i07

58. Steinwart I, Christmann A. Support vector machines. In: *(Information science and statistics).* Springer, New York, NY (2008). doi: 10.1007/978-0-387-77242-4

59. Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* Cambridge, Massachusetts, USA: MIT Press (2002). p. 658.

60. Cristianini N, Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge: Cambridge University Press (2000). doi: 10.1017/CBO9780511801389

61. Vapnik VN. *Statistical learning theory.* Statistical learning theory: New York: Wiley (1998). p. 768.

62. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. *Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien* (2025). Available online at: https://CRAN.R-project.org/package=e1071 (Accessed October 20, 2025).

63. Kavzoglu T, Colkesen I. A kernel functions analysis for support vector machines for land cover classification. *Int J Appl Earth Observation Geoinformation.* (2009) 11:352–9. doi: 10.1016/j.jag.2009.06.002

64. Hong H, Pradhan B, Bui DT, Xu C, Youssef AM, Chen W. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). *Geomatics Natural Hazards Risk.* (2017) 8:544–69. doi: 10.1080/19475705.2016.1250112

65. Breiman L. Random forests. *Mach Learning.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

66. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* (2002) 2:18–22. Available online at: https://CRAN.R-project.org/doc/Rnews/ (Accessed October 14, 2025).

67. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J Stat Software.* (2017) 77:1–17. doi: 10.18637/jss.v077.i01

68. Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An introductory review of deep learning for prediction models with big data. *Front Artif Intell.* (2020) 3:3. doi: 10.3389/frai.2020.00004

69. Manuylovich E, Argüello Ron D, Kamalian-Kopae M, Turitsyn SK. Robust neural networks using stochastic resonance neurons. *Commun Eng.* (2024) 3:169. doi: 10.1038/s44172-024-00314-0

70. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

71. Unigarro C, Hernandez J, Florez H. Artificial neural networks for image processing in precision agriculture: A systematic literature review on mango, apple, lemon, and coffee crops. *Informatics.* (2025) 12:46. doi: 10.3390/informatics12020046

72. Venables WN, Ripley BD. *Modern applied statistics with S. 4th edn.* New York: Springer (2002). Available online at: https://www.stats.ox.ac.uk/pub/MASS4/ (Accessed October 26, 2025).

73. Chen T, Guestrin C. (2016). XGBoost: A scalable tree boosting system, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, . pp. 785–94. San Francisco California USA: ACM. doi: 10.1145/2939672.2939785

74. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot.* (2013) 7:7. doi: 10.3389/fnbot.2013.00021

75. Islam MM, Alharthi M, Alkadi RS, Islam R, Masum AKM, Islam MM, et al. Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia. *AIMSAGRI.* (2024) 9:980–1003. doi: 10.3934/agrfood.2024053

76. Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *2012 IEEE Int Geosci Remote Sens Symposium*. (2012), 5372–5. Available online at: https://ieeexplore.ieee.org/document/6352393 (Accessed October 15, 2025).

77. Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ Model Software*. (2018) 101:1–9. doi: 10.1016/j.envsoft.2017.12.001

78. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography.* (2017) 40:913–29. doi: 10.1111/ecog.02881

79. Schratz P, Muenchow J, Iturritxa E, Cortés J, Bischl B, Brenning A, et al. Monitoring forest health using hyperspectral imagery: does feature selection improve the performance of machine-learning techniques? *Remote Sens*. (2021) 13:4832. doi: 10.3390/rs13234832

80. Akoglu H. User's guide to correlation coefficients. *Turkish J Emergency Med*. (2018) 18:91–3. doi: 10.1016/j.tjem.2018.08.001

81. Gargouri K, Rigane H, Arous I, Touil F. Evolution of soil organic carbon in an olive orchard under arid climate. *Scientia Horticulturae.* (2013) 152:102–8. doi: 10.1016/j.scienta.2012.11.025

82. Brady NC, Weil RR. *The nature and properties of soils* Upper Saddle River, NJ, USA: Pearson, *15th edition*. (2025). doi: 10.2136/sssaj2016.0005br.

83. Ortner M, Seidel M, Semella S, Udelhoven T, Vohland M, Thiele-Bruhn S. Content of soil organic carbon and labile fractions depend on local combinations of mineral-phase characteristics. *SOIL.* (2022) 8:113–31. doi: 10.5194/soil-8-113-2022

84. Zhang Y, Ai J, Sun Q, Li Z, Hou L, Song L, et al. Soil organic carbon and total nitrogen stocks as affected by vegetation types and altitude across the mountainous regions in the Yunnan Province, south-western China. *CATENA.* (2021) 196:104872. doi: 10.1016/j.catena.2020.104872

85. Zhu X, Si J, Jia B, He X, Zhou D, Wang C, et al. Changes of soil carbon along precipitation gradients in three typical vegetation types in the Alxa desert region, China. *Carbon Balance Manage*. (2024) 19:19. doi: 10.1186/s13021-024-00264-2

86. Zhang W, Song T, Dong W, Su X, Duan Y. Sun J. Particle composition, nutrient content, and alkalinity determine organic carbon variations in saline-alkali soils across different land-use types. *J Environ Management.* (2025) 394:127565. doi: 10.1016/j.jenvman.2025.127565

87. Ahmed IS, Hassan FA, Sulieman MM, Keshavarzi A, Elmobarak AA, Yousif KM, et al. Using environmental covariates to predict soil organic carbon stocks in Vertisols of Sudan. *Geoderma Regional.* (2022) 31:e00578. doi: 10.1016/j.geodrs.2022.e00578

88. Huang J, Liu J, Ye Y, Jiang Y, Lai Y, Qin X, et al. Mapping soil properties in the haihun river sub-watershed, yangtze river basin, China, by integrating machine learning and variable selection. *Sensors.* (2024) 24:3784. doi: 10.3390/s24123784

89. Kunkel VR, Wells T, Hancock GR. Modelling soil organic carbon using vegetation indices across large catchments in eastern Australia. *Sci Total Environment.* (2022) 817:152690. doi: 10.1016/j.scitotenv.2021.152690

90. Yang S, Xu X, Peng F, Zhu Z, Xu C, Ju C, et al. Unexpected responses of SOC decomposition and its temperature sensitivity to plant invasion across soil layers: Implications for plantation understory management. *CATENA.* (2025) 256:109110. doi: 10.1016/j.catena.2025.109110

91. Shoumik BAA, MdZ K, Gülser C. Climate sensitivity of soil organic carbon and nutrient stocks under different land uses in europe. *Eur J Soil Science.* (2025) 76:e70156. doi: 10.1111/ejss.70156

92. Chen Y, Han M, Yuan X, Zhou H, Zhao X, Schimel JP, et al. Long-term warming reduces surface soil organic carbon by reducing mineral-associated carbon rather than "free" particulate carbon. *Soil Biol Biochem.* (2023) 177:108905. doi: 10.1016/j.soilbio.2022.108905

93. Li X, McCarty GW, Karlen DL, Cambardella CA. Topographic metric predictions of soil redistribution and organic carbon in Iowa cropland fields. *CATENA.* (2018) 160:222–32. doi: 10.1016/j.catena.2017.09.026

94. Patton NR, Lohse KA, Seyfried MS, Godsey SE, Parsons SB. Topographic controls of soil organic carbon on soil-mantled landscapes. *Sci Rep*. (2019) 9:6390. doi: 10.1038/s41598-019-42556-5

95. Li X, McCarty GW. Use of principal components for scaling up topographic models to map soil redistribution and soil organic carbon. *JoVE.* (2018) 140):58189. doi: 10.3791/58189

96. Petropoulos T, Benos L, Busato P, Kyriakarakos G, Kateris D, Aidonis D, et al. Soil organic carbon assessment for carbon farming: A review. *Agriculture.* (2025) 15:567. doi: 10.3390/agriculture15050567

97. Zhang Y, Wang Y, Bai Y, Zhang R, Liu X, Ma X. Prediction of spatial distribution of soil organic carbon in helan farmland based on different prediction models. *Land.* (2023) 12:1984. doi: 10.3390/land12111984

98. Mosaid H, Barakat A, John K, Faouzi E, Bustillo V, El Garnaoui M, et al. Improved soil carbon stock spatial prediction in a Mediterranean soil erosion site through robust machine learning techniques. *Environ Monit Assess.* (2024) 196:130. doi: 10.1007/s10661-024-12294-x

99. Tikuye BG, Ray RL. Soil organic carbon retrieval using a machine learning approach from satellite and environmental covariates in the Lower Brazos River Watershed, Texas, USA. *Appl Computing Geosciences*. (2025) 26:100252. doi: 10.1016/j.acags.2025.100252

100. Liu X, Zhang M, Ma Z. Enhanced soil organic carbon mapping in Gannan's alpine meadows: A comparative analysis of machine learning models and satellite data. *Ecol Indicators.* (2025) 177:113800. doi: 10.1016/j.ecolind.2025.113800

101. Su L, Heydari M, Jaafarzadeh MS, Mousavi SR, Rezaei M, Fathizad H, et al. Incorporating forest canopy openness and environmental covariates in predicting soil organic carbon in oak forest. *Soil Tillage Res*. (2024) 244:106220. doi: 10.1016/j.still.2024.106220

102. Chun ZW, Shuang WH, Hou ZM, Wu W, Bin LH. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecol Indicators.* (2022) 143:109420. doi: 10.1016/j.ecolind.2022.109420

103. Pham VT, Le Thi HA, Luu HPH, Damel P. DCA-based weighted bagging: A new ensemble learning approach. In: Nguyen NT, Boonsang S, Fujita H, Hnatkowska B, Hong TP, Pasupa K, et al, editors. *Intelligent information and database systems*. Springer Nature, Singapore (2023). p. 121–32. doi: 10.1007/978-981-99-5837-5_11

104. Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects. *A Survey Ensemble Learning: Concepts Algorithms Applications Prospects. IEEE Access*. (2022) 10:99129–49. doi: 10.1109/ACCESS.2022.3207287

105. Sun Y, Ma J, Zhao W, Qu Y, Gou Z, Chen H, et al. Digital mapping of soil organic carbon density in China using an ensemble model. *Environ Res.* (2023) 231:116131. doi: 10.1016/j.envres.2023.116131

106. Tang K, Zhao X, Xu Z, Sun H. A stacking ensemble model for predicting soil organic carbon content based on visible and near-infrared spectroscopy. *Infrared Phys Technology.* (2024) 140:105404. doi: 10.1016/j.infrared.2024.105404

107. Adeniyi OD, Brenning A, Bernini A, Brenna S, Maerker M. Digital mapping of soil properties using ensemble machine learning approaches in an agricultural lowland area of lombardy, Italy. *Land.* (2023) 12:494. doi: 10.3390/land12020494

108. Tajik S, Ayoubi S, Zeraatpisheh M. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Regional.* (2020) 20:e00256. doi: 10.1016/j.geodrs.2020.e00256

109. Lilburne L, Helfenstein A, Heuvelink GBM, Eger A. Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids. *Geoderma.* (2024) 450:117052. doi: 10.1016/j.geoderma.2024.117052

110. Dvorakova K, Heiden U, Pepers K, Staats G, Van Os G, Van Wesemael B. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. *Geoderma*. (2023) 429:116128. doi: 10.1016/j.geoderma.2022.116128

111. Kakhani N, Alamdar S, Kebonye NM, Amani M, Scholten T. Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. *Remote Sensing.* (2024) 16:438. doi: 10.3390/rs16030438

112. Hoffmann U, Hoffmann T, Johnson EA, Kuhn NJ. Assessment of variability and uncertainty of soil organic carbon in a mountainous boreal forest (Canadian Rocky Mountains, Alberta). *CATENA.* (2014) 113:107–21. doi: 10.1016/j.catena.2013.09.009

113. Zhang S, Dai H, Chen C, Wei J, Guan Z, Niu X. Prediction of regional cropland soil organic carbon content and distribution using deep learning: a case study of the Northeast China Plain. *Environ Monit Assess.* (2025) 197:1159. doi: 10.1007/s10661-025-14622-1