ISSN 1064-2293, Eurasian Soil Science, 2026, Vol. 59:64. © Pleiades Publishing, Ltd., 2026.

GENESIS AND GEOGRAPHY OF SOILS

Data-Scarce Dryland SOC Mapping Using Landsat Spectral Indices and a Hybrid Gradient Boosting Kriging Framework

H. Gunal[a], *, A. M. Ismael[a], M. Kılıc[b], **, M. Budak[c], K. Polat[a], and N. Acir[d]

[a] *Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Harran University, Sanlıurfa, 63290 Turkiye*

[b] *Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Malatya Turgut Ozal University, Malatya, 44900 Turkiye*

[c] *Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Siirt University, Siirt, 56100 Turkiye*

[d] *Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Kırşehir Ahi Evran University, Kırşehir, 40100 Turkiye*

**e-mail: hikmetgunal@harran.edu.tr*

***e-mail: mirac.kilic@ozal.edu.tr*

Received October 20, 2025; revised January 3, 2026; accepted January 12, 2026

Abstract—Accurate estimation of soil organic carbon (SOC) is critical in arid, data-limited regions where land degradation and climatic stress disrupt carbon cycling. We developed a hybrid learning–geostatistical framework that couples Bayesian-tuned gradient boosting decision trees (GBDT) with ordinary kriging (OK) of model residuals to capture both covariate-driven variance and residual spatial autocorrelation. Field-measured SOC observations were integrated with Landsat-8/9 spectral derivatives (vegetation and soil reflectance indices). GBDT modeled nonlinear SOC–environment relationships; OK then interpolated the spatially structured residuals, and the two components were summed to form final predictions. The hybrid raised held-out performance to R^2 = 0.72 and reduced RMSE by 32.8% relative to GBDT alone. Vegetation indices explained 57% of the variance attributed to predictors (rising to 66.6% when the brightness index is included), indicating a central role for plant-soil feedbacks in carbon accumulation. The approach is readily scalable for semi-arid landscapes and supports evidence-based restoration targeting, climate adaptation planning, and MRV (monitoring, reporting, verification) of carbon outcomes.

Keywords: spatial machine learning, ensemble regression, arid agroecosystems, multispectral satellite imagery, geostatistical prediction

DOI: 10.1134/S1064229325604470

INTRODUCTION

Soil organic carbon (SOC) represents an important fraction of the terrestrial ecosystem, functioning as its most extensive carbon sink and exerting significant influence over the global carbon budget [53]. Even slight alterations in SOC can affect atmospheric CO_2 concentrations over decades, positioning SOC accumulation as an effective and economical approach to climate change mitigation [40]. In addition to its role in carbon dynamics, SOC significantly enhances soil health by influencing structure, moisture retention, nutrient turnover, and biological interactions, making it a critical indicator of land degradation [9, 47].

A multitude of factors, including parent material, climate variability, landscape position, vegetation type, and human interventions like farming practices, affect SOC's levels and spatial arrangement [4, 37]. Owing to this heterogeneity, producing detailed SOC maps is a core aim of digital soil mapping to optimize land-use and carbon-sequestration strategies; such maps help locate high carbon-storage zones and areas vulnerable to degradation or poor management [14, 20, 27].

While lab-based SOC quantification is highly accurate, sparse field sampling and costly, time-consuming analyses limit coverage and resolution, especially in unstable or remote regions [6, 29, 57]. Advances in remote sensing (RS) and machine learning (ML) address these constraints: RS provides cost-effective, wall-to-wall covariates, spectral indices, topography, and climate, and reflectance in the visible, near-infrared, and shortwave-infrared ranges correlates with organic content; indices such as NDVI, SAVI, and EVI offer indirect yet reliable indicators of surface-level carbon dynamics [1, 25, 49]. ML models non-linear relationships between these environmental covariates and observed SOC, enabling scalable, high-resolution mapping for modern land-management decisions [43, 45, 72].

Among machine learning models, ensemble methods, especially GBDT, have emerged as powerful tools for soil prediction tasks due to their ability to model complex, non-linear relationships and deal with collinearity and spatial variance [7, 44, 73]. Models such as Random Forest (RF), Classification and Regression Trees (CART), and Cubist also show robust predictive power using spectral inputs [42]. In areas like northern Iraq, where access and data are limited, XGBoost was recently shown to outperform RF, delivering high accuracy when trained on Landsat bands, NDVI, SAVI, and elevation data [26].

Despite these advances, few studies have systematically compared various GBDT approaches and tuning methods under remote-sensing-only conditions, especially in semi-arid, data-deficient landscapes. Previous work using XGBoost [26] did not explore the impact of model optimization or alternative GBDT setups, leaving significant methodological gaps.

The Amediya region in northern Iraq combines steep, dissected terrain (1046–1506 m), a Mediterranean climate with a xeric moisture regime, sparse shrub, grass cover, and extensive rainfed cropland, conditions that limit field access and yield scarce, uneven SOC observations. These constraints make a spectral-only covariate set especially pertinent: vegetation-season imagery (July) allows indices such as NDVI, SAVI, EVI, and MSAVI to proxy organic inputs and canopy productivity, whereas the late-dry season (September) bare-soil window enables soil-sensitive metrics (e.g., BI, NDSI, SRCI, GSI) to capture brightness, texture, and carbonate effects linked to SOC stabilization. At the same time, orographic controls and the land-use mosaic impart moderate short-range spatial dependence in SOC that spectral covariates may not fully explain. We therefore pair a Bayesian-tuned GBDT (to model nonlinear SOC–spectral relations) with ordinary kriging of residuals (to correct local bias and honor remaining spatial continuity), a strategy well-suited to Amediya's data-scarce, heterogeneous landscape.

This study sets out to (i) assess the extent to which Landsat-derived spectral indices, when modeled through Bayesian-optimized GBDT, can predict SOC in a semi-arid, data-limited region; and (ii) evaluate whether combining this with kriging of residuals (i.e., a GBDT-OK hybrid) enhances accuracy. We hypothesize that (a) fine-tuning hyperparameters will significantly improve GBDT model performance, and (b) regression kriging will further refine predictions by addressing spatial autocorrelation not captured by machine learning alone.

MATERIALS AND METHODS

Description of the study area. The research was conducted in the Amediya sub-district, situated approximately 70 km north of Duhok Province in northern Iraq, covering around 20.06 km^2 between latitudes 37°40′–37°60′ N and longitudes 43°27′-43°30′ E. The elevation ranges from 1046 to 1506 m above sea level (Fig. 1).

The region exhibits a thermic soil temperature regime, where annual mean soil temperatures range between 15 and 22°C, with more than a 6°C seasonal variation at 50 cm depth. Its xeric soil moisture regime is typical of Mediterranean climates, characterized by extended dry periods following the summer solstice and moist conditions during winter months. These climatic traits influence both soil development and land use [56].

Soils in the area are generally non-saline and correspond mainly to Rendzic Leptosols in the World Reference Base for Soil Resources [67]. These soils, commonly known as Rendzinas, typically form on calcareous parent materials, such as the limestone and marl prevalent in the region. They are characterized by a dark, humus-rich surface horizon directly overlying the calcareous material. The mountainous, semi-arid Mediterranean climate of the study area is characteristic for the formation of Leptosols, rather than Chernozems, which are archetypal soils of continental grassland ecosystems.

The climate corresponds to the Köppen classification type Csa, signifying a Mediterranean pattern with hot, dry summers and mild, wetter winters. Annual precipitation averages around 633.4 mm, largely concentrated in winter and spring. Rainfall is scarce between July and October, contributing to seasonal drought risk. Vegetation is relatively sparse and mainly consists of shrubs and grasses. The area's plant cover can be categorized into high and medium forests, bushlands, and grasslands. Upland zones near the Great Zap River host dry oak forests. Bush formations dominate the landscape, while grasslands, covering over 16% of the area, serve as grazing land for livestock [36].

Soil sampling and laboratory analysis. To facilitate systematic sampling, the study area was overlaid with a 500 × 500 m grid. Soil samples were collected near the corners of each grid cell (Fig. 1). Sampling was conducted from the topsoil layer (0–20 cm). When grid corners were inaccessible due to steep terrain, built-up zones, roads, or water bodies, nearby substitute locations were selected, and their GPS coordinates were recorded. Collected soils were air-dried at ambient temperature, manually cleaned to remove coarse fragments and organic debris, then ground and sieved through a 2 mm mesh before analysis. Soil organic carbon (SOC) concentrations were measured using the modified Walkley-Black dichromate oxidation method as described by Nelson and Sommers [41].

Remote sensing variables and image processing. To support the spatial modeling of soil organic carbon (SOC), multispectral Landsat imagery was selected to capture two complementary surface conditions: one corresponding to peak vegetation cover and the other to maximum soil exposure. Specifically, Landsat 8 Operational Land Imager (OLI) data from September 2021 and Landsat 9 OLI-2 data from July 2022 were obtained from the USGS Earth Explorer platform [61]. The July 2022 imagery was selected to represent peak growing season conditions, maximizing vegeta-

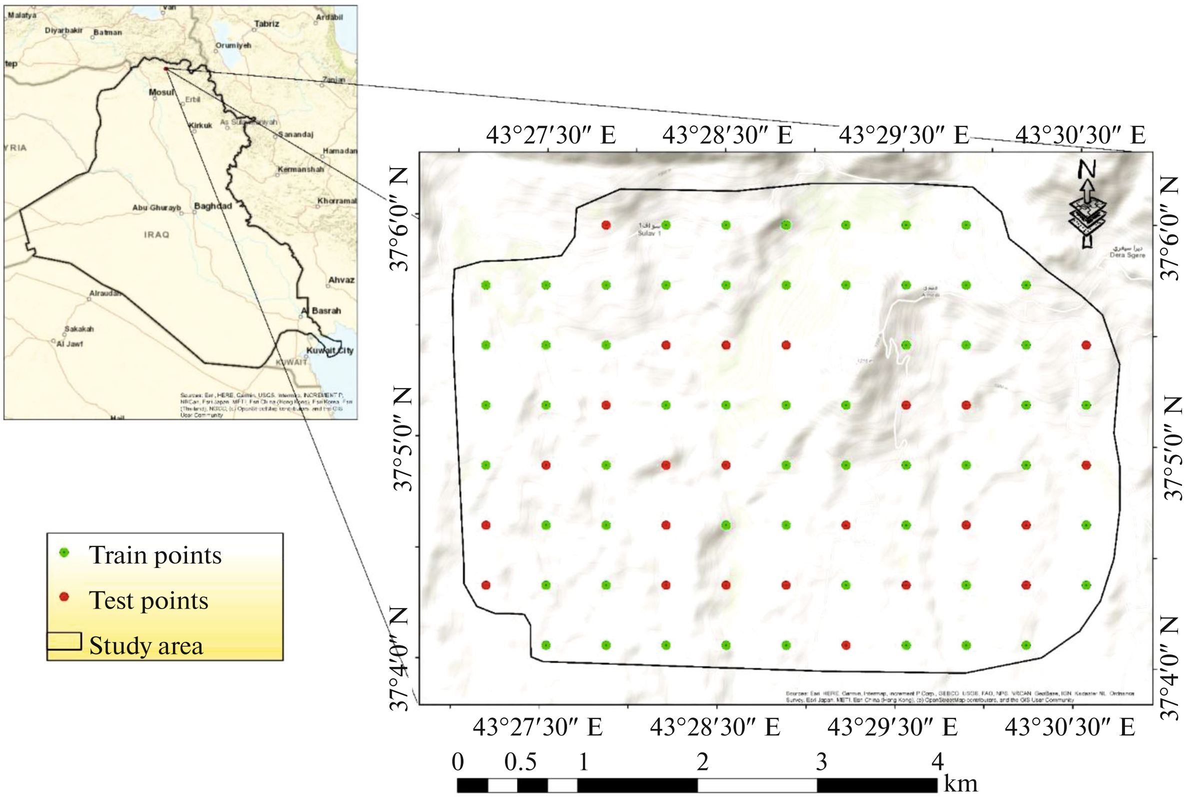

Fig. 1. Location of the study area and distribution of training (green) and test (red) points. The white–grey background represents a hillshade digital elevation model (DEM) illustrating the topographic variability of the area. Contour lines are derived from the DEM with a vertical interval of 5 m.

tion signal and facilitating the extraction of vegetation indices such as NDVI and EVI. Conversely, the September 2021 imagery was selected to capture minimal vegetation cover, corresponding to post-harvest or dry-season conditions, thereby enhancing the detectability of soil reflectance characteristics [29. 30].

The use of different Landsat platforms for each acquisition was driven by data quality and availability. Over the study area, cloud-free Landsat 8 imagery was not available for July 2022, whereas Landsat 9 provided high-quality, cloud-free imagery during this period. In contrast, suitable Landsat 8 data were available for September 2021, while Landsat 9 data for that period either contained cloud contamination or were not available. The integration of imagery from Landsat 8 and Landsat 9 is considered methodologically sound, as both sensors have near-identical spectral bands and spatial resolution, with Landsat 9 OLI-2 designed to replicate the performance of Landsat 8 OLI while improving radiometric resolution from 12 to 14 bits [62]. Studies have confirmed the geometric and radiometric consistency between the two sensors, with co-registration errors typically below 2.5 meters RMSE, well within acceptable limits for 30 m spatial resolution data [10].

All images were atmospherically corrected prior to analysis using the Dark Object Subtraction (DOS1) method implemented through the Semi-Automatic Classification Plugin in QGIS. This preprocessing step was essential for minimizing atmospheric effects and ensuring consistency across the datasets, thereby supporting accurate index derivation and SOC spatial modeling [5].

Radiometric indices. Incorporating spectral indices as auxiliary variables in soil organic carbon (SOC) prediction models improves accuracy by mitigating reflectance-related distortions [70]. This study selected a suite of vegetation, and soil-sensitive indices known for their association with SOC. Vegetation indices included the Normalized Difference Vegetation Index (NDVI) [35, 36], Enhanced Vegetation Index (EVI) [23], Soil-Adjusted Vegetation Index (SAVI) [38, 39], and its modified form, the Modified SAVI (MSAVI) [40, 41].

In addition to vegetation metrics, soil-specific indices were also utilized. The Brightness Index (BI), initially developed by Escadafal [13], was included to capture general soil reflectance. Given the significant absorption of shortwave infrared radiation by clay and carbonate minerals, relevant indices were derived from September satellite imagery, chosen for its minimal soil moisture conditions, enhancing bare soil detection [5].

Additional soil-sensitive radiometric indices were integrated to complement canopy metrics, including

Table 1. Radiometric indices chosen for application in soil organic carbon spatial prediction models

Index	Formula
Brightness Index	$\sqrt{b4^2 + b5^2}$
Simple Ratio Clay Index	$b6/b7$
Grain Size Index	$(b4 - b3)/(b4 + b3 + b2)$
Normalized Difference Vegetation Index	$(b5 - b4)/(b5 + b4)$
Soil-Adjusted Vegetation Index	$((b8 - b4))/((b8 + b4 + 0.5)) \times 1.5$
Modified Soil Adjusted Vegetation Index	$(2 \times b5 + 1 - \sqrt{((2 \times b5 + 1)^2 - 8 \times (b5 - b4)))}/2$
Enhanced Vegetation Index	$2.5 \times (b5 - b4)/(b5 + 6 \times b4 - 7.5 \times b2 + 1)$
Normalized Difference Soil Index	$(b7 - b2)/(b7 + b2)$

the Simple Ratio Clay Index (SRCI), the Topsoil Grain Size Index (GSI) [68], and the Normalized Difference Soil Index (NDSI) [11]. All indices were computed from Collection-2 surface reflectance of Landsat-8 OLI and Landsat-9 OLI-2 using the band combinations listed in Table 1. Throughout, b2–b7 denote blue, green, red, NIR, SWIR1, and SWIR2, respectively; b8 denotes the panchromatic band and is not used in these indices unless explicitly stated in Table 1.

The timing of the satellite imagery acquisition was selected to correspond closely with the soil sampling period. Distinct dates were chosen for computing vegetation-related and soil-related radiometric indices. For vegetation indices, such as NDVI, EVI, SAVI, and MSAVI, images from July 2022 were used, as this period exhibited peak vegetation vigor, reflected by the highest NDVI values. Conversely, images captured in late September 2021, representing the driest conditions and lowest NDVI values, were employed to derive soil-related indices such as SRCI, GSI, and NDSI.

Spatial Modeling of Soil Organic Carbon Distribution

Gradient boosting decision tree (GBDT). GBDT is a supervised learning algorithm that constructs predictive models through the sequential application of regression trees using the gradient boosting framework [15]. Each iteration of the Least Squares Boosting (LSBoost) algorithm aims to minimize the residuals, i.e., the differences between predicted and observed values, by optimizing a predefined loss function. With each step, a new base learner (denoted as hk(x)h_k(x)) is added to the model ensemble Fk(x), improving prediction accuracy through functional gradient descent [33]. Model development was performed using MATLAB's fit ensemble function, employing Bayesian optimization to fine-tune the hyperparameters. A 10-fold cross-validation approach was used to ensure robustness.

Hybrid approach. GBDT combined with kriging. To improve prediction accuracy, a hybrid methodology integrating machine learning with geostatistical interpolation was implemented (Fig. 2). The modeling approach followed a three-stage process. Initially, GBDT was applied to predict SOC levels based on remote sensing-derived radiometric indices. Next, OK was used to model the residuals from GBDT predictions, capturing spatial patterns that the machine learning model alone may have missed. Finally, the kriged residuals were added back to the GBDT estimates, generating a refined SOC prediction surface that leverages both the predictive strength of GBDT and the spatial continuity modeled by OK [52].

The GBDT, optimized using Bayesian algorithms, was applied to the training dataset to predict SOC values. Residuals were calculated by subtracting observed laboratory-based SOC measurements from the GBDT-predicted values at each sampling location, as shown in Eq. (1):

$$r_{(p)} = p(\mathrm{SOC_{GBDT}}) - p(\mathrm{SOC_{OBS}}), \quad (1)$$

where $r_{(\mathrm{p})}$ is denotes the residual at point p, $p(\mathrm{SOC_{GBDT}})$ is the SOC predicted by GBDT model and $p(\mathrm{SOC_{OBS}})$ is the observed SOC determined with laboratory analysis. These residuals were then spatially interpolated using OK to capture spatial patterns not accounted for by the machine learning model. The kriged residuals at point p are given in Eq. (2):

$$r'_{(\mathrm{p})} = \sum \lambda_i r_{(\mathrm{p})} \quad (2)$$

In this equation, $r'_{(\mathrm{p})}$ is kriging-interpolated residual, $\sum \lambda_i$ are optimal weights with $\sum \lambda = 1$ and $r_{(\mathrm{p})}$ are the residuals estimated from neighboring sample points. The final step is to calculate the final SOC produced as the sum of the GBDT estimates and the OK estimates of the residuals (Eq. (3)).

$$\mathrm{S\hat{O}C}_{(\mathrm{p})} = p(\mathrm{SOC_{GBDT}}) + r'_{(\mathrm{p})}. \quad (3)$$

This approach integrates the predictive capability of the GBDT model with the spatial sensitivity of geo-

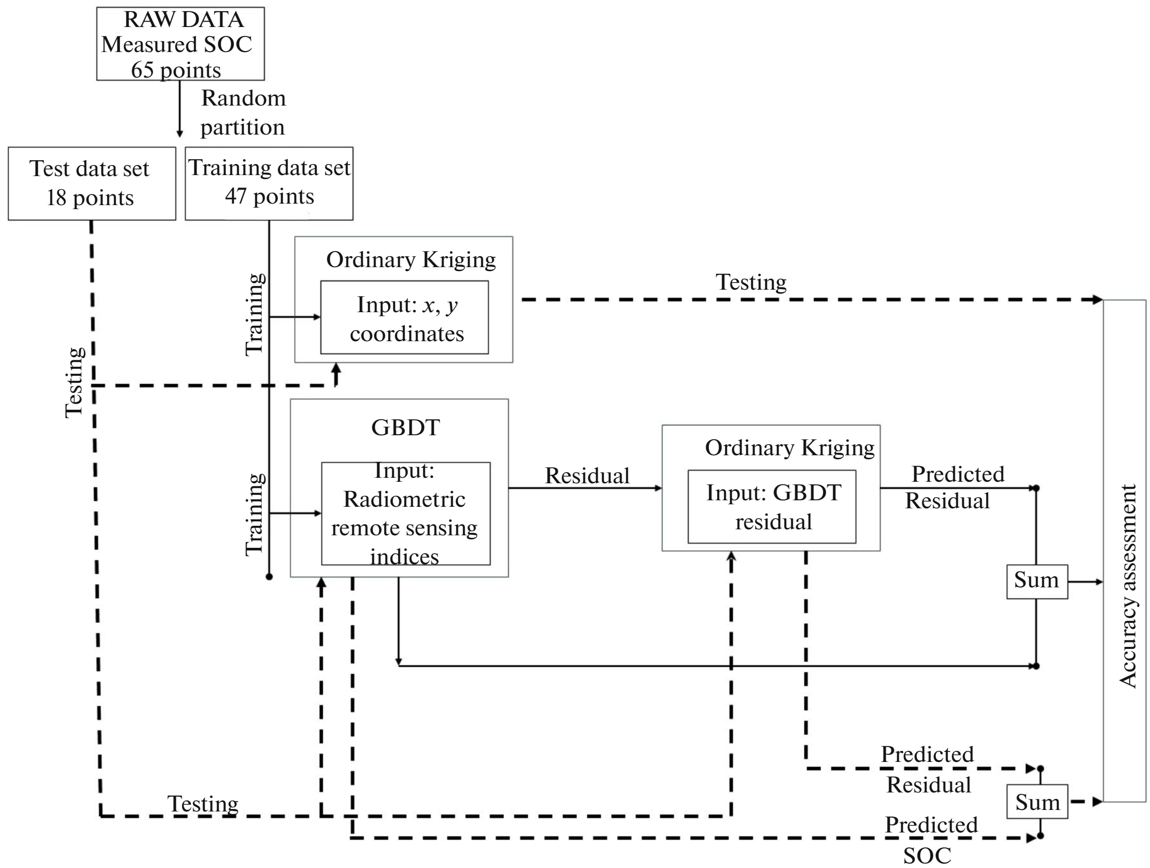

Fig. 2. Three-stage hybrid modeling workflow for spatial Soil Organic Carbon (SOC) estimation.

statistical interpolation, enhancing the accuracy of SOC spatial predictions.

Model evaluation metrics. The dataset was randomly divided into a training set (70%) and a testing set (30%) using the “Random selection within subsets” tool in QGIS. The predictive accuracy of the models was assessed using three statistical measures: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), following the approach described by Shcherbakov et al. [54].

The R^2 value, which indicates the proportion of variance in the observed data explained by the model, was calculated as (Eq. (4)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i)^2}{\sum_{i=1}^{n} (y_i - y_{\mathrm{avg}})^2}, \tag{4}$$

RMSE and MAE were used to measure prediction errors, defined as follows (Eqs. (5) and (6)):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i)^2}{n}}, \tag{5}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} (|y_i - y_i|)}{n}, \tag{6}$$

where y_i is observed SOC value, $\bar{y}_i$ is predicted SOC value, $\bar{y}$ is the mean of observed SOC values, and n is number of observations.

RESULTS

Descriptive statistics. Table 2 presents the summary statistics for SOC content in both the training and testing subsets used in the spatial prediction models. In the training set, SOC values ranged from 0.25 to 1.68%, with an average of 0.74%. The testing set showed a similar distribution, with SOC values spanning from 0.32 to 1.68% and the same mean value of 0.74%. The overall standard deviation in the full dataset was 0.40, compared to 0.42 in the training set and 0.32 in the test set. The coefficient of variation (CV) was calculated as 53% for the full dataset, 57% for the training data, and 43% for the test data, indicating moderate variability in SOC content across samples.

Spatial distribution models and accuracy assessment. *Gradient boosting decision tree (GBDT).* To enhance the predictive power of the model, hyperparameters

Table 2. Descriptive statistics for test, training and all data sets

Parameter	Min	Max	Mean	Median	Std. Dev	CV, %
All dataset	0.25	2.68	0.74	0.60	0.40	53
Training dataset	0.25	1.68	0.74	0.60	0.42	57
Test dataset	0.32	1.68	0.75	0.66	0.32	43

Table 3. GBDT Model Optimization Summary: Hyperparameters and Performance Comparison

Metric/Parameter	GBDT-1*	GBDT-2	Improvement, %	Training (GBDT-2)	Testing (GBDT-2)
Cross-validation loss (CVL)	0.39	0.26	−33		
Coefficient of Determination (R^2)	0.21	0.65	+213	0.83	0.65
Root Mean Square Error (RMSE)	0.39	0.23	−42	0.19	0.23
Mean Absolute Error (MAE)	0.17	0.13	−24	0.14	0.13

*GBDT-1 (baseline): Untuned model using library default hyperparameters; GBDT-2 (tuned): Model obtained via Bayesian hyperparameter optimization; CVL (cross-validation loss): Mean squared error averaged across K folds on the training partition (lower is better); % Improvement: Calculated as (GBDT-2−GBDT-1)/|GBDT-1| ×100; decreases in error/loss indicate improvement (negative sign).

Table 4. Optimal semivariogram model parameters for ordinary kriging interpolation of soil organic carbon (SOC) content in the training dataset and GBDT residuals

Variogram	Model	Range, km	Nugget	Sill	Nugget/Sill, %	RMSE (Test Dataset)
SOC	Stable	2.5	0.31	0.42	74.7	0.38
Residual SOC	Stable	4.0	0.04	0.04	97	0.16

were tuned to reduce cross-validation loss (CVL). The initial GBDT-1 model yielded a CVL of 0.39, while the optimized version, GBDT-2, showed a lower CVL of 0.26, indicating a significant improvement in generalization capability. The test dataset's R^2 increased from 0.21 in GBDT-1 to 0.65 in GBDT-2 (Table 3), indicating a greater proportion of explained variance.

Error metrics further confirmed the performance gain: RMSE declined from 0.39 to 0.23 and MAE from 0.17 to 0.13. Overall, CVL decreased by 33.3%, R^2 increased by 0.44, RMSE decreased by 42%, and MAE by 24%. The optimized model (GBDT-2) achieved $R^2 = 0.83$ on the training set and $R^2 = 0.65$ on the independent test set. Key hyperparameters for GBDT-2 included: minimum parent size = 10, minimum leaf size = 5, learning rate = 0.113, number of learning cycles (trees) = 47, and maximum tree depth = 4.

Kriging-based spatial model. The optimal parameters for OK applied to SOC data are summarized in Table 4. Since the Kolmogorov-Smirnov test indicated a departure from normality, a logarithmic transformation was applied to meet the assumptions required for geostatistical modeling. Several semivariogram models, spherical, exponential, Gaussian, and stable, were tested. The Stable model provided the best fit based on cross-validation results, as it yielded the lowest RMSE and highest R^2 values. The nugget-to-sill ratio for the raw SOC data was 74.7%, suggesting moderate spatial dependence, while the residuals had a ratio of 97%, indicating weak spatial correlation.

The residual variogram displayed a low nugget effect (0.04), implying minimal spatial structure, an expected outcome given the reduced spatial autocorrelation post-GBDT modeling. The flexible nature of the Stable model made it well-suited to represent the spatial continuity characteristics of the SOC dataset, thereby enhancing the accuracy of spatial interpolation.

Integrated GBDT-ordinary kriging model. For the hybrid GBDT-OK model, out-of-fold training performance was $R^2 = 0.90$, RMSE = 0.16, and MAE = 0.01, while the independent test set yielded R^2 = 0.72, RMSE = 0.15, and MAE = 0.12; the shifts from training to testing indicate strong generalization with modest degradation primarily in absolute error (Table 5).

The GBDT prediction map (Fig. 3) showed SOC values ranging between 0.29 and 1.68%, with a median concentration of 0.62%. Given the non-normal distribution of the predictions, the first and third quartiles were 0.60 and 0.79%, respectively, while the mean value was 0.69%. This spread reflects the model's ability to capture substantial spatial variability, owing to its capacity to integrate multiple environmental predictors and model complex nonlinear interactions. The SOC distribution map reveals pronounced spatial heterogeneity, with distinct zones of high and low carbon

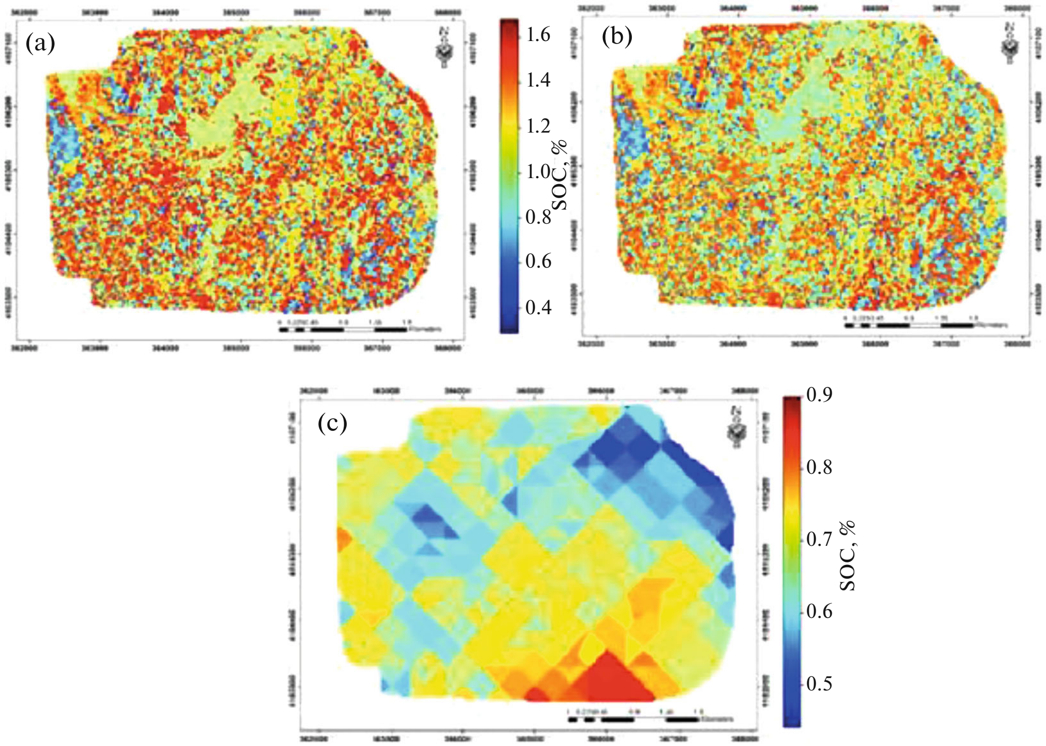

Fig. 3. Comparative spatial distributions of soil organic carbon (SOC) content (%) predicted by: (a) Gradient Boosted Decision Trees (GBDT), (b) GBDT-Ordinary Kriging hybrid model, and (c) Ordinary Kriging (OK) alone.

concentrations, likely influenced by topography, land cover, and land management practices. While this detailed spatial variability can help identify SOC-rich zones and prioritize areas for intervention, field-based verification remains essential to confirm that localized predictions align with ground-truth conditions.

The integrated GBDT–OK approach (Fig. 3b), which combines data-driven prediction with spatial interpolation of Ordinary Kriging residuals, produced SOC estimates ranging from 0.19 to 1.76%. Given the non-normal distribution of the spatial predictions, the SOC distribution is summarized using robust statistics, with a median value of 0.59% and first and third quartiles of 0.43 and 0.88%, respectively, while the mean value was 0.70%. The broader distribution compared to GBDT alone indicates that incorporating residual spatial patterns via kriging improves predictions, particularly in areas where GBDT may exhibit systematic biases.

In contrast, the standalone Ordinary Kriging model (Fig. 3c) produced a narrower range of SOC estimates, varying between 0.43 and 0.91%, with a mean value of 0.64%. This reduced variability reflects the strong smoothing effect of OK when applied to a limited number of field observations, resulting in an underrepresentation of extreme SOC values. This narrower range reflects kriging's tendency to smooth predictions and gravitate towards mean values, especially when sample density is limited or uneven. The spatial continuity evident in the OK map reflects the model's reliance on the semivariogram, which may reduce sharp variations while providing reliable interpolation under sparse sampling conditions.

Variable importance in soil organic carbon estimation. The contribution of each predictor to SOC estimation was evaluated using the feature importance scores from the GBDT model. To enable direct comparison, all importance values were normalized to a scale where the highest-ranking variable was set at 100%. This facilitated a clear assessment of which covariates had the greatest influence on SOC variability across the study area.

Table 5. Accuracy assessment of the GBDT-OK hybrid model

Data set	R^2	MAE, %	RMSE, %	Relative Importance-GBDT, %
Training	0.90	0.01	0.16	15.7
Test	0.72	0.12	0.15	32.8

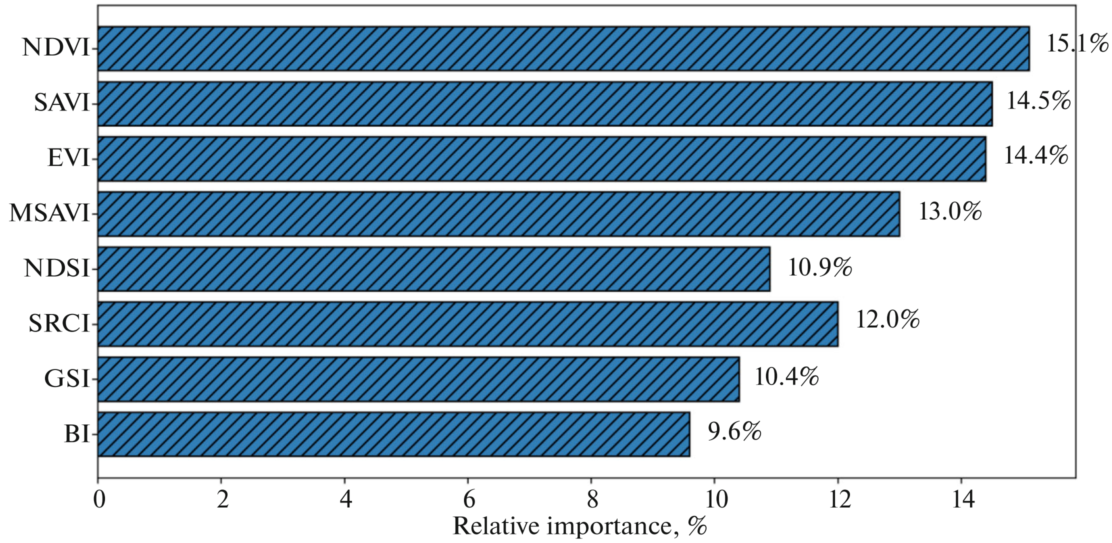

Fig. 4. Relative importance (%) of vegetation and soil spectral indices in soil organic carbon prediction.

Among the vegetation-related indices, NDVI contributed the most at 15.1%, followed by SAVI (14.5%), EVI (14.4%), MSAVI (13.0%), and BI (9.6%). Collectively, these vegetation indicators accounted for 66.6% of the total variance explained in the model (Fig. 4). Regarding soil spectral indicators, NDSI emerged as the top contributor (12.0%), with SRCI and GSI explaining 10.9 and 10.4% of the variance, respectively. Altogether, the soil index variables contributed 21.3% to the overall model performance.

DISCUSSION

Analyzing descriptive statistics is a crucial first step in digital soil mapping, offering insight into the underlying patterns and variability of SOC data. This statistical overview not only contextualizes raw observations but also enhances the interpretability of machine learning models, defined as the ability to understand how and why a model generates specific outputs [39]. The observed consistency in summary metrics supports the reliability of SOC data for modeling its spatial distribution.

Influence of Model Optimization and Parameter Tuning. Tuning model parameters improved the accuracy of SOC predictions, with R^2 values increasing from 0.65 to 0.72 following the integration of geostatistical residuals with machine learning predictions (Tables 3, 5). These improvements are consistent with earlier work by Emadi et al. [12], Zhang et al. [74], and [44] Parvizi and Fatehi, which emphasized the importance of combining statistical optimization with spatial modeling techniques to refine SOC mapping. Moreover, variations in optimal hyperparameter settings reported in studies such as Budak et al. [7] and Guo et al. [19] indicate the context-specific nature of model tuning, as differences in predictor variables and dataset structures can influence overall performance.

The optimal setting for the Minimum Leaf Size hyperparameter in gradient boosting models is highly context-dependent, influenced by the specific characteristics of the dataset, the nature of the prediction task, and interactions with other tuning parameters. In this study, the Minimum Parent Size, representing the smallest number of observations required in a node before further splitting, was set to 10. This falls within the broader range of values (2 to 10) reported in previous SOC spatial modeling efforts [7, 16, 72]. Ye et al. [71] employed a related hyperparameter, min_child_weight, within the XGBoost framework to predict soil organic matter in Nhui Province using GF-6 satellite data. Their model, tuned with a min_child_weight of 4, achieved an R^2 of 0.63, demonstrating that parameters controlling model complexity can substantially impact prediction accuracy. These hyperparameters are sensitive to factors such as the type of covariate data (e.g., UAV imagery, Sentinel products), sample size, and the underlying model architecture. The satisfactory performance of the GBDT-2 model in this study, achieved through LSBoost and Bayesian optimization, demonstrates that properly tuned hyperparameters contribute meaningfully to model robustness and predictive power.

Effectiveness of the GBDT-OK hybrid model. The spatial structure of soil organic carbon (SOC) in this study exhibited moderate spatial dependence, as reflected by a nugget-to-sill ratio of 74.7%. This level of dependence is consistent with findings from prior studies (e.g., [3], which demonstrated that SOC variability is shaped by environmental gradients such as land use patterns, vegetative cover, and terrain complexity. Conversely, the residuals derived from the

GBDT model displayed a much higher nugget/sill ratio of approximately 97%, signifying that any remaining spatial variation in the residuals is weak or nearly random.

Despite this low spatial autocorrelation in the residuals, the use of OK in the regression-kriging framework remains methodologically sound. Incorporating OK provides a means to capture subtle spatial signals or anomalies not fully explained by the machine learning model. Even when residuals exhibit near-randomness, kriging can function as a spatial smoothing tool, addressing localized model inaccuracies and reducing prediction noise [28, 60]. This integration enhances the visual consistency and interpretability of SOC spatial maps, especially in data-sparse areas or transition zones between land use types.

Moreover, from a theoretical standpoint, regression kriging tends to default to the regression output when residuals lack spatial structure, ensuring model parsimony and reducing the risk of overfitting [18]. This principle was reflected in our findings, where the minimal nugget variance (0.041) indicated limited spatial correlation in residuals, validating the GBDT model's explanatory strength. Nevertheless, incorporating OK remains a standard geostatistical procedure in digital soil mapping, used even in cases of weak spatial dependence to account for overlooked local trends and reinforce spatial continuity [21, 75].

The results demonstrate the practical advantage of integrating OK with machine learning for SOC prediction. Although the residuals exhibited weak spatial dependence, the hybrid GBDT-OK model achieved superior performance. This indicates that the geostatistical layer captured additional spatial nuances beyond those explained by covariates alone. Specifically, RMSE was reduced by 15.7% in the training set. An even greater reduction of 32.8% was observed in the test set compared with the optimized GBDT-2 model. These gains demonstrate the hybrid model's ability to generalize better and detect subtle spatial heterogeneity.

When juxtaposed with other spatial modeling approaches, such as artificial neural networks (ANNs) using elevation, temperature, precipitation, and NDVI as inputs, our model demonstrated comparable improvements in RMSE (approximately 26.9%) to those observed in kriging-enhanced systems, including universal kriging and inverse distance weighting methods. This suggests that combining GBDT with kriging is not only statistically sound but also highly effective in digital soil carbon mapping.

Contribution of vegetation and soil texture indicators to SOC prediction. The results of this study corroborate earlier findings by Parvizi and Fatehi [44] and Liu et al. [30], which demonstrated that hybrid geostatistical approaches, such as regression kriging and stratified kriging using spatial detectors, can enhance the precision of SOC spatial predictions. Their reported RMSE improvements, ranging from 13.7 to 21.6%, align well with the performance gains observed in our GBDT-OK model. These improvements are especially meaningful in the context of the inherent variability in SOC content, indicated by the high coefficient of variation and moderate spatial dependence observed in our dataset.

A key driver of this enhanced predictive capacity lies in the effective optimization of the GBDT model. Unlike other nonlinear models such as multilayer perceptron neural networks, GBDT iteratively corrects errors by concentrating on underperforming estimators (weak learners), thereby directly minimizing the loss function. Coupled with Bayesian hyperparameter tuning, this strategy strengthens the model's ability to learn complex patterns while avoiding overfitting [51, 59].

The variable importance analysis showed that vegetation indices (e.g., NDVI, SAVI, EVI, MSAVI) collectively explained 66.6% of the variance in SOC predictions. These indices are sensitive to canopy density and plant productivity. Both factors are closely linked to organic matter inputs and carbon cycling processes. Soil indices including NDSI, SRCI, and GSI contributed an additional 21.3%. This reflects the influence of surface texture and mineral composition in SOC estimation. The comparable contributions of vegetation and soil-based spectral indicators indicate the multifactorial controls on SOC distribution. This finding supports the inclusion of diverse covariates in spatial modeling frameworks.

The findings reveal the dominant influence of vegetation-related variables in predicting SOC levels across the study area. Among the vegetation indices analyzed, NDVI, SAVI, and EVI emerged as the most influential predictors of SOC, consistent with the observations of Zhong et al. [76] and Wang et al. [65], who reported a similar predictive strength of NDVI and BI in SOC mapping across diverse landscapes. The strong contribution of SAVI in this study is particularly relevant, as this index is particularly effective in environments characterized by limited vegetation cover. SAVI compensates for soil brightness and reduces reflectance variability caused by sparse vegetation and topographic shading, thereby improving accuracy in semi-arid and arid ecosystems [2, 24].

Despite the prominence of vegetation indices, abiotic constraints remain pivotal in arid rangelands. Deane McKenna et al. [35] reported that precipitation regimes and intrinsic soil properties often exert greater control over carbon dynamics than vegetation or management practices. This suggests that while vegetation metrics are essential for capturing aboveground influences on SOC, their impact may be modulated by deeper soil and climatic controls, especially in water-limited environments.

Limitations and insights in remote sensing-based soil organic carbon estimation. Estimating SOC using remote sensing data presents significant challenges due to the inherent spatial heterogeneity of soil properties. This variability complicates the establishment of precise relationships between soil characteristics and data derived from medium- to low-resolution satellite imagery. However, empirical spectral indices such as the NDSI offer a practical solution by improving the ability to identify exposed soil surfaces and accentuate relevant soil features [11]. In the present study, NDSI emerged as the most influential soil index, reaffirming its effectiveness in landscapes marked by high spatial variability, particularly within agricultural zones.

Soil texture parameters, specifically the SRCI and GSI, also significantly contributed to SOC estimation, collectively accounting for 21.3% of the model's explanatory power. These findings are in agreement with prior research [66, 69], which emphasized that soil texture exerts a pivotal influence on SOC dynamics through its effects on moisture retention, nutrient cycling, and general soil fertility. The observed correlation between SOC and clay content in this study is consistent with earlier literature, especially in hot, arid regions where finer soil particles play a key role in protecting organic matter from microbial decomposition [66, 76]. Classic studies by [55] and [34] further support this relationship, noting that the role of soil texture in regulating SOC becomes more pronounced in warmer climatic zones.

The moderate spatial resolution of Landsat imagery (30 m) can limit the accuracy of SOC modeling in heterogeneous terrains like Amediya. Each pixel, covering approximately 0.09 hectares, often encompasses a mixture of soil types and land covers, leading to an averaging effect that obscures finer variations in SOC at the sub-pixel level [77]. In regions such as Amediya, where agricultural lands intersect with mountainous features, SOC distribution tends to fluctuate at scales smaller than a single Landsat pixel. As a result, detailed SOC anomalies, such as enriched patches or depleted zones, can be either diminished or overlooked entirely [22]. Previous studies indicate that coarse spatial resolution can mask local variation in soil properties, whereas finer spatial resolution tends to provide more detailed and often more accurate SOC/SOM maps, particularly in heterogeneous terrain and over limited extents (e.g., Sentinel-2 at 10–20 m outperforming or revealing more detail than Landsat at 30 m) [8, 31, 78]. Multi-resolution analyses further show that model performance is sensitive to the scale of predictors, with finer resolutions generally improving spatial detail and, in many cases, predictive accuracy, although the optimal mapping resolution can vary with study extent and environmental heterogeneity [17, 46]. Accordingly, the ideal grid size should align with zones of relative SOC homogeneity; exclusive reliance on 30 m data in complex landscapes risks underrepresenting true SOC variability and may compromise the precision of predictive models and maps.

CONCLUSIONS

This study demonstrates that phenology-targeted, spectral covariates paired with a hybrid learning-geostatistical framework can deliver reliable SOC predictions in data-scarce, semi-arid terrain. The design isolates two complementary windows, peak greenness for canopy signals and late dry season for bare-soil signals and shows that vegetation indices are the principal drivers of SOC variability, accounting for the majority of modeled importance alongside soil-reflectance metrics. A Bayesian-tuned GBDT coupled with residual kriging improves held-out accuracy (Test $R^2 = 0.72$) and reduces error relative to GBDT alone, indicating value in correcting short-range spatial structure not fully captured by spectral covariates. Conceptually, the results clarify when regression-kriging remains beneficial despite weak residual autocorrelation, adding nuance to model-selection guidance for digital soil mapping in heterogeneous landscapes. Practically, the approach enables higher-resolution SOC surfaces suitable for restoration targeting, climate-adaptation planning, and MRV workflows in regions where dense sampling is infeasible. Limitations include potential sub-pixel mixing at 30 m resolution and seasonal sensitivity; future work should incorporate multi-temporal stacks and terrain-climate covariates and assess transferability across neighboring basins.

AUTHOR CONTRIBUTION

H. Günal: Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization, Supervision, Writing—review and editing, Writing original draft. A.M. Ismael: Data curation, Investigation, Methodology, Validation, Writing, review and editing, Writing original draft. M. Kılıç: Investigation, Software, Validation, Visualization, Writing, review and editing, Formal analysis. M. Budak: Software, Validation, Visualization, Writing original draft. K. Polat: Data curation, Validation, Visualization.

FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

DATA AVAILABILITY

The data that supports the findings of this study are available from the corresponding author upon reasonable request.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This work does not contain any studies involving human and animal subjects.

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. V. Allory, A. Cambou, P. Moulin, Ch. Schwartz, P. Cannavo, L. Vidal-Beaudet, and B. G. Barthès, "Quantification of soil organic carbon stock in urban soils using visible and near infrared reflectance spectroscopy (VNIRS) in situ or in laboratory conditions," *Sci. Total Environ.* **686**, 764–773 (2019). https://doi.org/10.1016/j.scitotenv.2019.05.192
2. N. Gholami Baghi and J. Oldeland, "Do soil-adjusted or standard vegetation indices better predict above ground biomass of semi-arid, saline rangelands in North-East Iran?," *Int. J. Remote Sens.* **40**, 8223–8235 (2019). https://doi.org/10.1080/01431161.2019.1606958
3. S. A. Bangroo, G. R. Najar, E. Achin, and P. N. Truong, "Application of predictor variables in spatial quantification of soil organic carbon and total nitrogen using regression kriging in the North Kashmir forest Himalayas," *Catena* **193**, 104632 (2020). https://doi.org/10.1016/j.catena.2020.104632
4. M. Birol and H. Günal, "Field scale variability in soil properties and silage corn yield," *Soil Studies* **11**, 27–34 (2022). https://doi.org/10.21657/soilst.1143238
5. S. Bousbih, M. Zribi, Ch. Pelletier, A. Gorrab, Z. Lili-Chabaane, N. Baghdadi, N. Ben Aissa, and B. Mougenot, "Soil texture estimation using radar and optical data from Sentinel-1 and Sentinel-2," *Remote Sens.* **11**, 1520 (2019). https://doi.org/10.3390/rs11131520
6. M. Budak and H. Gunal, "Carbon storage potentials of soils under different land uses in Upper Tigris Basin," *Anatolian Journal of Forest Research* **4**, 61–74 (2018).
7. M. Budak, E. Günal, M. Kılıç, İ. Çelik, M. Sırrı, and N. Acir, "Improvement of spatial estimation for soil organic carbon stocks in Yuksekova plain using Sentinel 2 imagery and gradient descent–boosted regression tree," *Environ. Sci. Pollut. Res.* **30**, 53253–53274 (2023). https://doi.org/10.1007/s11356-023-26064-8
8. F. Castaldi, "Sentinel-2 and Landsat-8 multi-temporal series to estimate topsoil properties on croplands," *Remote Sens.* **13**, 3345 (2021). https://doi.org/10.3390/rs13173345
9. İ. Çelik, H. Günal, N. Acir, Z. B. Barut, and M. Budak, "Soil quality assessment to compare tillage systems in Cukurova Plain, Turkey," *Soil Tillage Res.* **208**, 104892 (2021). https://doi.org/10.1016/j.still.2020.104892
10. M. J. Choate, R. Rengarajan, J. C. Storey, and M. Lubke, "Landsat 9 geometric characteristics using underfly data," *Remote Sens.* **14**, 3781 (2022). https://doi.org/10.3390/rs14153781
11. Yi. Deng, Ch. Wu, M. Li, and R. Chen, "RNDSI: A ratio normalized difference soil index for remote sensing of urban/suburban environments," *Int. J. Appl. Earth Obs. Geoinformation* **39**, 40–48 (2015). https://doi.org/10.1016/j.jag.2015.02.010
12. M. Emadi, R. Taghizadeh-Mehrjardi, A. Cherati, M. Danesh, A. Mosavi, and T. Scholten, "Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran," *Remote Sens.* **12**, 2234 (2020). https://doi.org/10.3390/rs12142234
13. R. Escadafal, "Remote sensing of arid soil surface color with Landsat thematic mapper," *Adv. Space Res.* **9**, 159–163 (1989). https://doi.org/10.1016/0273-1177(89)90481-x
14. Ch. J. Feeney, D. A. Robinson, A. R. C. Thomas, P. Borrelli, D. M. Cooper, and L. May, "Agricultural practices drive elevated rates of topsoil decline across Kenya, but terracing and reduced tillage can reverse this," *Sci. Total Environ.* **870**, 161925 (2023). https://doi.org/10.1016/j.scitotenv.2023.161925
15. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.* **29**, 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
16. P. Fu, Ch. Clanton, K. M. Demuth, V. Goodman, L. Griffith, M. Khim-Young, J. Maddalena, K. LaMarca, L. A. Wright, D. W. Schurman, and J. R. Kellner, "Accurate quantification of 0–30 cm soil organic carbon in croplands over the continental united states using machine learning," *Remote Sens.* **16**, 2217 (2024). https://doi.org/10.3390/rs16122217
17. Yo. Garosi, Sh. Ayoubi, M. Nussbaum, and M. Sheklabadi, "Effects of different sources and spatial resolutions of environmental covariates on predicting soil organic carbon using machine learning in a semi-arid region of Iran," *Geoderma Reg.* **29**, e00513 (2022). https://doi.org/10.1016/j.geodrs.2022.e00513
18. P. Goovaerts, *Geostatistics for Natural Resources Evaluation* (Oxford University Press, New York, 1997).
19. H. Guo, J. Wang, D. Zhang, J. Cui, Yo. Yuan, H. Bao, M. Yang, J. Guo, F. Chen, W. Zhou, G. Wu, Ya. Guo, H. Wei, B. Qiao, and Sh. Zhao, "Mapping surface soil organic carbon density of cultivated land using machine learning in Zhengzhou," *Environ. Geochem. Health* **47**, 1 (2024). https://doi.org/10.1007/s10653-024-02313-8
20. R. He, L. Luo, A. Shamsuddin, and Q. Tang, "Corporate carbon accounting: a literature review of carbon accounting research from the Kyoto Protocol to the Paris Agreement," *Accounting and Finance* **62**, 261–298 (2022). https://doi.org/10.1111/acfi.12789
21. T. Hengl, G. B. M. Heuvelink, and A. Stein, "A generic framework for spatial prediction of soil variables based on regression-kriging," *Geoderma* **120**, 75–93 (2004). https://doi.org/10.1016/j.geoderma.2003.08.018
22. M. Hoffmann, N. Jurisch, J. Garcia Alba, E. Albiac Borraz, M. Schmidt, V. Huth, H. Rogasik, H. Rieckh, G. Verch, M. Sommer, and J. Augustin, "Detecting small-scale spatial heterogeneity and temporal dynamics of soil organic carbon (SOC) stocks: A comparison between automatic chamber-derived C budgets and repeated soil inventories," *Biogeosciences* **14**, 1003–1019

(2017).
https://doi.org/10.5194/bg-14-1003-2017

23. A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," *Remote Sens. Environ.* **83**, 195–213 (2002).
https://doi.org/10.1016/s0034-4257(02)00096-2
24. A. R. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sens. Environ.* **25**, 295–309 (1998).
https://doi.org/10.1016/0034-4257(88)90106-x
25. X. Ji, C. Li, X. Ma, and E. Veldkamp, "Digital mapping of soil organic carbon: Advances in remote sensing, modeling, and environmental drivers," PhD Thesis (University of Groningen, 2025).
https://doi.org/10.33612/diss.1394968210
26. H. S. Khalaf, Ya. T. Mustafa, and M. A. Fayyadh, "Digital mapping of soil organic matter in northern Iraq: Machine learning approach," *Appl. Sci.* **13**, 10666 (2023).
https://doi.org/10.3390/app131910666
27. M. Kılıç, R. Gündoğan, and H. Günal, "Development and applications of digital soil mapping," in *New Research in Food, Environment, Agroforestry, and Agriculture for Sustainability*, Ed. by K. Bellitürk, M. F. Baran, and A. Çelik (IKSAD Publishing House, 2021), pp. 53–87.
28. S. Kumar, R. Lal, and D. Liu, "A geographically weighted regression kriging approach for mapping soil organic carbon stock," *Geoderma* **189**, 627–634 (2012).
https://doi.org/10.1016/j.geoderma.2012.05.022
29. A. A. J. Lima, J. C. Lopes, R. P. Lopes, T. de Figueiredo, E. Vidal-Vázquez, and Z. Hernández, "Soil organic carbon assessment using remote-sensing data and machine learning: A systematic literature review," *Remote Sens.* **17**, 882 (2025).
https://doi.org/10.3390/rs17050882
30. Ya. Liu, Yi. Chen, Z. Wu, B. Wang, and Sh. Wang, "Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity," *Catena* **196**, 104953 (2021).
https://doi.org/10.1016/j.catena.2020.104953
31. Ch. Luo, W. Zhang, X. Zhang, and H. Liu, "Mapping soil organic matter content using Sentinel-2 synthetic images at different time intervals in Northeast China," *Int. J. Digital Earth* **16**, 1094–1107 (2023).
https://doi.org/10.1080/17538947.2023.2192005
32. Y.-J. Ma, T. Xie, and X.-Y. Li, "Spatial variation of soil organic carbon in the Qinghai Lake watershed, northeast Qinghai-Tibet Plateau," *Catena* **213**, 106187 (2022).
https://doi.org/10.1016/j.catena.2022.106187
33. D. Marvin, L. Nespoli, D. Strepparava, and V. Medici, "A data-driven approach to forecasting ground-level ozone concentration," *Int. J. Forecast.* **38**, 970–987 (2022).
https://doi.org/10.1016/j.ijforecast.2021.07.008
34. P. A. McDaniel and L. C. Munn, "Effect of temperature on organic carbon-texture relationships in Mollisols and Aridisols," *Soil Sci. Soc. Am. J.* **49**, 1486–1489 (1985).
https://doi.org/10.2136/sssaj1985.03615995004900060031x
35. M. D. McKenna, S. E. Grams, M. Barasha, A. J. Antoninka, and N. C. Johnson, "Organic and inorganic soil carbon in a semi-arid rangeland is primarily related to abiotic factors and not livestock grazing," *Geoderma* **419**, 115844 (2022).
https://doi.org/10.1016/j.geoderma.2022.115844
36. A. Mehri, H. Karimi, Ya. T. Mustafa, A. M. F. Al-Quraishi, and S. Galalizadeh, "Predicting soil erosion using RUSLE model in Duhok governorate, Kurdistan region of Iraq," in *Natural Resources Deterioration in MENA Region: Land Degradation, Soil Erosion, and Desertification*, Ed. by A. M. F. Al-Quraishi and Y. T. Mustafa (Springer, 2024), pp. 171–187.
https://doi.org/10.1007/978-3-031-58315-5_9
37. B. Minasny, A. B. McBratney, B. P. Malone, and I. Wheeler, "Digital mapping of soil carbon," in *Advances in Agronomy* (Elsevier, 2013), Vol. 118, pp. 1–47.
https://doi.org/10.1016/b978-0-12-405942-9.00001-3
38. V. L. Mulder, S. de Bruin, M. E. Schaepman, and T. R. Mayr, "The use of remote sensing in soil and terrain mapping—A review," *Geoderma* **162**, 1–19 (2011).
https://doi.org/10.1016/j.geoderma.2010.12.018
39. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: Definitions, methods, and applications," arXiv Preprint (2019). arXiv:1901.04592
40. J. Navarro-Pedreño, M. B. Almendro-Candel, and A. A. Zorpas, "The increase of soil organic matter reduces global warming, myth or reality?," *Sci* **3**, 18 (2021).
https://doi.org/10.3390/sci3010018
41. D. W. Nelson and L. E. Sommers, "Total carbon, organic carbon, and organic matter," in *Methods of Soil Analysis, Part 3: Chemical Methods* (Wiley, 1996), Vol. 5, pp. 961–1010.
https://doi.org/10.2136/sssabookser5.3.c34
42. O. Odebiri, J. Odindi, and O. Mutanga, "Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review," *Int. J. Appl. Earth Obs. Geoinformation* **102**, 102389 (2021).
https://doi.org/10.1016/j.jag.2021.102389
43. J. Padarian, B. Minasny, and A. B. McBratney, "Machine learning and soil sciences: A review aided by machine learning tools," *Soil* **6**, 35–52 (2020).
https://doi.org/10.5194/soil-6-35-2020
44. Ya. Parvizi and Sh. Fatehi, "Geospatial digital mapping of soil organic carbon using machine learning and geostatistical methods in different land uses," *Sci. Rep.* **15**, 4449 (2025).
https://doi.org/10.1038/s41598-025-88062-9
45. C. V. Patriche, B. Roşca, R. G. Pîrnău, and I. Vasiliniuc, "Spatial modelling of topsoil properties in Romania using geostatistical methods and machine learning," *PLoS One* **18**, e0289286 (2023).
https://doi.org/10.1371/journal.pone.0289286
46. Ch. Piedallu, E. Pedersoli, E. Chaste, F. Morneau, I. Seynave, and J.-C. Gégout, "Optimal resolution of soil properties maps varies according to their geographical extent and location," *Geoderma* **412**, 115723 (2022).
https://doi.org/10.1016/j.geoderma.2022.115723
47. R. Prăvălie, I. A. Nita, C. Patriche, M. Niculiță, M. V. Birsan, B. Roșca, and G. Bandoc, "Global

changes in soil organic carbon and implications for land degradation neutrality and climate stability," *Environ. Res.* **201**, 111580 (2021). https://doi.org/10.1038/s41467-024-48252-x

48. J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian, "A modified soil adjusted vegetation index," *Remote Sens. Environ.* **48**, 119–126 (1994). https://doi.org/10.1016/0034-4257(94)90134-1
49. G. P. O. Reddy, "Satellite remote sensing sensors: Principles and applications," in *Geospatial Technologies in Land Resources Mapping, Monitoring and Management* (Springer, 2018), pp. 21–43. https://doi.org/10.1007/978-3-319-78711-4_2
50. J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the great plains with ERTS," in *Proceedings of the 3rd Earth Resources Technology Satellite-1 Symposium* (1974), pp. 309–317.
51. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* **61**, 85–117 (2015). https://doi.org/10.1016/j.neunet.2014.09.003
52. A. P. Sergeev, A. G. Buevich, E. M. Baglaeva, and A. V. Shichkin, "Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals," *Catena* **174**, 425–435 (2019). https://doi.org/10.1016/j.catena.2018.11.037
53. A. Shakoor, E. Pendall, and C. A. Macdonald, "Microbial mechanisms of interactive climate-driven changes in soil N_2O and CH_4 fluxes: A global meta-analysis," *J. Environ. Manage.* **376**, 124380 (2025). https://doi.org/10.1016/j.jenvman.2025.124380
54. M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, V. A. Kamaev, et al., "A survey of forecast error measures," *World Applied Sciences Journal* **24**, 171–176 (2013). https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80032
55. Z. R. Sims and G. A. Nielsen, "Organic carbon in montana soils as related to clay content and climate," *Soil Sci. Soc. Am. J.* **50**, 1269–1271 (1986). https://doi.org/10.2136/sssaj1986.036159950050000-50037x
56. D. W. Smith, *Soil survey staff: keys to soil taxonomy* (Natural Resources Conservation Service, 2014).
57. P. T. Sorenson, S. A. Quideau, and B. Rivard, "High resolution measurement of soil organic carbon and total nitrogen with laboratory imaging spectroscopy," *Geoderma* **315**, 170–177 (2018). https://doi.org/10.1016/j.geoderma.2017.11.032
58. A. Stevens, B. van Wesemael, H. Bartholomeus, D. Rosillon, B. Tychon, and E. Ben-Dor, "Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils," *Geoderma* **144**, 395–404 (2008). https://doi.org/10.1016/j.geoderma.2007.12.009
59. H. Tahraoui, A. Amrane, A.-E. Belhadj, and J. Zhang, "Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting," *Environ. Technol. Innovation* **27**, 102419 (2022). https://doi.org/10.1016/j.eti.2022.102419
60. B. Takoutsing and G. B. M. Heuvelink, "Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors," *Geoderma* **428**, 116192 (2022). https://doi.org/10.1016/j.geoderma.2022.116192
61. USGS, United State Geological Survey, Earth Explorer, (2022). https://earthexplorer.usgs.gov/.
62. USGS, United States Geological Survey, *Landsat 9 Science Users Handbook* (2021). https://www.usgs.gov/media/files/landsat-9-data-users-handbook.
63. L. Wang, Zh. Li, D. Wang, Sh. Liao, X. Nie, and Ya. Liu, "Factors controlling soil organic carbon with depth at the basin scale," *Catena* **217**, 106478 (2022). https://doi.org/10.1016/j.catena.2022.106478
64. B. Wang, C. Waters, S. Orgill, A. Cowie, A. Clark, D. Li Liu, M. Simpson, I. McGowen, and T. Sides, "Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia," *Ecol. Indic.* **88**, 425–438 (2018). https://doi.org/10.1016/j.ecolind.2018.01.049
65. K. Wang, Ya. Qi, W. Guo, J. Zhang, and Q. Chang, "Retrieval and mapping of soil organic carbon using Sentinel-2A spectral images from bare cropland in autumn," *Remote Sens.* **13**, 1072 (2021). https://doi.org/10.3390/rs13061072
66. M.-Y. Wang, X.-Z. Shi, D.-S. Yu, S.-X. Xu, M.-Z. Tan, W.-X. Sun, and Y.-C. Zhao, "Regional differences in the effect of climate and soil texture on soil organic carbon," *Pedosphere* **23**, 799–807 (2013). https://doi.org/10.1016/s1002-0160(13)60071-5
67. IUSS Working Group WRB, *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, *World Soil Resources Reports* No. 106 (2022), pp. 1–191.
68. J. Xiao, Y. Shen, R. Tateishi, and W. Bayaer, "Development of topsoil grain size index for monitoring desertification in arid land using remote sensing," *Int. J. Remote Sens.* **27**, 2411–2422 (2006). https://doi.org/10.1080/01431160600554363
69. H. Xiao, Zh. Li, X. Chang, B. Huang, X. Nie, Ch. Liu, L. Liu, D. Wang, and J. Jiang, "The mineralization and sequestration of organic carbon in relation to agricultural soil erosion," *Geoderma* **329**, 73–81 (2018). https://doi.org/10.1016/j.geoderma.2018.05.018
70. B. Xie, J. Ding, X. Ge, X. Li, L. Han, and Zh. Wang, "Estimation of soil organic carbon content in the Ebinur Lake wetland, Xinjiang, China, based on multisource remote sensing data and ensemble learning algorithms," *Sensors* **22**, 2685 (2022). https://doi.org/10.3390/s22072685
71. Zh. Ye, Z. Sheng, X. Liu, Yo. Ma, R. Wang, Sh. Ding, M. Liu, Z. Li, and Q. Wang, "Using machine learning algorithms based on GF-6 and google earth engine to predict and map the spatial distribution of soil organic matter content," *Sustainability* **13**, 14055 (2021). https://doi.org/10.3390/su132414055
72. X. Zhang, J. Xue, S. Chen, N. Wang, T. Xie, Yi. Xiao, X. Chen, Zh. Shi, Yu. Huang, and Zh. Zhuo, "Fine resolution mapping of soil organic carbon in croplands with feature selection and machine learning in northeast plain China," *Remote Sens.* **15**, 5033 (2023). https://doi.org/10.3390/rs15205033

73. W. Zhang, Zh. Du, D. Zhang, Sh. Yu, and Yu. Hao, "Boosted regression tree model-based assessment of the impacts of meteorological drivers of hand, foot and mouth disease in Guangdong, China," *Sci. Total Environ.* **553**, 366–371 (2016). https://doi.org/10.1016/j.scitotenv.2016.02.023
74. W.-C. Zhang, H.-S. Wan, M.-H. Zhou, W. Wu, and H.-B. Liu, "Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques," *Ecol. Indic.* **143**, 109420 (2022). https://doi.org/10.1016/j.ecolind.2022.109420
75. Ya. Zhang, W. Ji, D. D. Saurette, T. H. Easher, H. Li, Zh. Shi, V. I. Adamchuk, and A. Biswas, "Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging," *Geoderma* **366**, 114253 (2020). https://doi.org/10.1016/j.geoderma.2020.114253
76. Z. Zhong, Zh. Chen, Ya. Xu, Ch. Ren, G. Yang, X. Han, G. Ren, and Yo. Feng, "Relationship between soil organic carbon stocks and clay content under different climatic conditions in Central China," *Forests* **9**, 598 (2018). https://doi.org/10.3390/f9100598
77. T. Zhou, Ya. Geng, Ch. Ji, X. Xu, H. Wang, J. Pan, J. Bumberger, D. Haase, and A. Lausch, "Prediction of soil organic carbon and the C : N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images," *Sci. Total Environ.* **755**, 142661 (2021). https://doi.org/10.1016/j.scitotenv.2020.142661
78. D. Žížala, R. Minařík, and T. Zádorová, "Soil organic carbon mapping using multispectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions," *Remote Sens.* **11**, 2947 (2019). https://doi.org/10.3390/rs11242947