Journal of Soils and Sediments (2025) 25:1554–1577
https://doi.org/10.1007/s11368-025-04027-5

SOILS, SEC 5 • SOIL AND LANDSCAPE ECOLOGY • RESEARCH ARTICLE

Comparison of geostatistics, machine learning algorithms, and their hybrid approaches for modeling soil organic carbon density in tropical forests

Viet Hoang Ho[1,3] · Hidenori Morita[2] · Thanh Ha Ho[3] · Felix Bachofer[4] · Thi Thuong Nguyen[3]

Received: 4 January 2025 / Accepted: 29 March 2025 / Published online: 5 April 2025

Abstract

Purpose Understanding the spatial variability of soil organic carbon density (SOCD) in tropical forests is necessary for efficient climate change mitigation initiatives. However, accurately modeling SOCD in these landscapes is challenging due to low-density sampling efforts and the limited availability of in-situ data caused by constrained accessibility. In this study, we aimed to explore the most suitable modeling technique for SOCD estimation in the context of tropical forest ecosystems.
Methods To support the research, thirty predictor covariates derived from remote sensing data, topographic attributes, climatic factors, and geographic positions were utilized, along with 104 soil samples collected from the top 30 cm of soil in Central Vietnamese tropical forests. We compared the effectiveness of geostatistics (ordinary kriging, universal kriging, and kriging with external drift), machine learning (ML) algorithms (random forest and boosted regression tree), and their hybrid approaches (random forest regression kriging and boosted regression tree regression kriging) for the prediction of SOCD. Prediction accuracy was evaluated using the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE) obtained from leave-one-out cross-validation.
Results The study results indicated that hybrid approaches performed best in predicting forest SOCD with the greatest values of R^2 and the lowest values of MAE and RMSE, and the ML algorithms were more accurate than geostatistics. Additionally, the prediction maps produced by the hybridization showed the most realistic SOCD pattern, whereas the kriged maps were prone to have smoother patterns, and ML-based maps were inclined to possess more detailed patterns. The result also revealed the superiority of the ML plus residual kriging approaches over the ML models in reducing the underestimation of large SOCD values in high-altitude mountain areas and the overestimation of low SOCD values in low-lying terrain areas.
Conclusion Our findings suggest that the hybrid approaches of geostatistics and ML models are most suitable for modeling SOCD in tropical forests.

Keywords Digital soil mapping · Hybrid approaches · Kriging · Machine learning · Soil organic carbon density · Tropical forests

1 Introduction

Forests constitute the primary terrestrial carbon pools globally (Tebeje 2020; Wang et al. 2023), with the majority of carbon sequestered in the soil (Sharma et al. 2023), significantly affecting the carbon cycle (Anderson-Teixeira et al. 2021). In forest ecosystems, the tropics cover less than 20% of the Earth's terrestrial surface (Pillay et al. 2022) but harbor a substantial fraction of non-atmospheric land-based carbon, storing up to 324 Pg (Mackey et al. 2020). In tropical forests, approximately 32% of carbon is stored in the soil under organic form (Pan et al. 2024) and considered a potential sink for elevated carbon dioxide (CO_2) emission,

Responsible editor: Armando Gómez-Guerrero

✉ Viet Hoang Ho
hoviethoang@hueuni.edu.vn

[1] Graduate School of Environmental and Life Science, Okayama University, 1 Chome- 1- 1 Tsushimanaka, Kita Ward, Okayama 700 - 8530, Japan

[2] Graduate School of Environmental, Life, Natural Science and Technology, Okayama University, 1 Chome- 1- 1 Tsushimanaka, Kita Ward, Okayama 700 - 8530, Japan

[3] University of Agriculture and Forestry, Hue University, 102 Phung Hung Str, Hue City 53000, Vietnam

[4] German Aerospace Center (DLR), Earth Observation Center, 82234 Wessling, Germany

making it necessary to maintain and enhance these significant soil organic carbon (SOC) stocks at high levels in order to address climate change effectively (Md. Shoaibur Rahman et al. 2023). Over the past several centuries, deforestation and forest degradation have been prevalent in most tropical nations due to land-use conversion, unsustainable forest practices, and human-induced deterioration, resulting in a significant decrease in forest SOC (Li et al. 2022; Qin et al. 2024). This diminution could occur in a rise in global carbon emissions, emphasizing the importance of implementing reliable monitoring systems to understand better SOC stocks and flows in tropical forests, as well as to support effective climate change mitigation initiatives (Padarian et al. 2022; Duarte et al. 2022).

Digital soil mapping (DSM) is a commonly employed approach for not only evaluating soil attributes but also monitoring their spatial distribution based on discrete samples (Fathololoumi et al. 2020; Shafizadeh-Moghadam et al. 2022). In recent studies, digital mapping of soil organic carbon density (SOCD) in forest ecosystems is a prevalent use of DSM owing to its high efficiency, accuracy, and affordability (Zhou et al. 2020c; Emadi et al. 2020; Borůvka et al. 2022). The DSM approach utilizes mathematical relationships between field soil measurements and corresponding environmental covariates (the SCORPAN framework) to quantitatively predict the spatial and temporal variation of SOCD (Dash et al. 2022). In the DSM system, a wide range of environmental covariates were introduced to leverage the covariance of the SOCD variable, commencing with soil-forming elements (e.g., climate, organisms, relief, parent material, and age) (Jenny 1994), and subsequently expanding to include geographic position variables to enhance its estimation (Zhou et al. 2022; Xia et al. 2022; Xu et al. 2022). Those environmental variables can be retrieved from various available data sources, such as remote sensing (RS) data, digital elevation model (DEM), and climatic data (Veronesi and Schillaci 2019; Sun et al. 2019; Emadi et al. 2020; Li et al. 2024b). Along with the progress in data acquisition, selecting an appropriate predictive model is required to accurately quantify SOCD (Shafizadeh-Moghadam et al. 2022; Garsia et al. 2023). Soil, in its natural condition, is the final outcome of the interplay of environmental factors through certain physical, chemical, and biological processes (Zhu et al. 2022). As a consequence of the localized variations in soil-forming processes, the determination of these influencing factors in different geographical locations is challenging, leading to intricate and uncertain spatial patterns of SOC content (Liu et al. 2015). Hence, the choice of predictive models for mapping SOCD in surface forest soils may differ depending on different study regions (Wang et al. 2020b). Numerous modeling techniques have been applied in SOCD estimation. Among these, geostatistics used quantitatively to quantify spatial autocorrelation of regionalized variables and create a spatial prediction model for interpolating unsampled locations (Chen et al. 2019a) is a long-standing and widely utilized method in DSM (Shafizadeh-Moghadam et al. 2022). Among geostatistical methods, SOCD estimation is mainly conducted using kriging models, such as ordinary kriging (OK), universal kriging (UK) – also known as kriging with internal drift, and kriging with external drift (KED) (Mishra et al. 2009; Asa et al. 2012; Nussbaum et al. 2014; Tziachris et al. 2019; Gia Pham et al. 2019). However, there was a strong progression from geostatistics to a relatively new technique—machine learning (ML) for predictive modeling, due to several reasons. Firstly, the great explosion in information technology has culminated in the proliferation of vast amounts of data (e.g., remotely sensed and high-resolution climatic data), which may not be essential for most geostatistics but are well-suited for ML algorithms (Wang et al. 2018). Secondly, geostatistics-based models may struggle to capture the intricate nonlinearities present in the data adequately and typically rely on assumptions regarding the spatial distribution of data, such as normality and stationarity, which do not accurately reflect real-world conditions (Nussbaum et al. 2014; Fouedjio and Klump 2019). Thirdly, geostatistics may exhibit lower accuracy than more sophisticated techniques and tend to generate smooth map surfaces because it is inclined to minimize the variance of predictions relative to the variation of observations to mitigate the influence of outliers (Veronesi and Schillaci 2019). Fourthly, the interpolation of geostatistics is normally restricted by the amount of observed samples and the configuration of their locations (Delmelle 2021). Given that ML is capable of addressing these drawbacks of geostatistics, ecological modelers are likely to reduce their dependence on spatial interpolation (Minasny and McBratney 2016; Veronesi and Schillaci 2019). A variety of ML techniques have been utilized in DSM for predicting SOCD, namely artificial neural networks, support vector regression, Cubist, boosted regression tree (BRT), and random forest (RF) (Lamichhane et al. 2019; Emadi et al. 2020). According to Lamichhane et al. (2019), no ML model can consistently reach optimal performance across every situation, yet the tree-based models provide exceptional predictive capabilities within numerous investigations. Recently, two tree-based ML models, including RF and BRT, have been popular in DSM methodologies for estimating SOCD (Lamichhane et al. 2019) since they possess the ability to reduce over-fitting and are insensitive to multicollinearity (Gu et al. 2019; Zhou et al. 2022; Li et al. 2024a). Although ML is typically reliable and adaptable for evaluating intricate data, it neglects to account for the possibility of spatial autocorrelation, which can lead to biased outcomes, especially in strongly heterogeneous terrain areas (Erdogan Erten et al. 2022; Liu et al. 2022; Zhu et al. 2022). Meanwhile, soil attributes, including SOC content, are considered regionalized variables, showing a strong spatial heterogeneity and autocorrelation structure (Liu et al. 2015; Marchant 2018). Consequently, the relationship between SOCD and the predictor variables, which ML regression models are unable to explain, is transmitted to the residuals of SOCD, resulting in the existence of trends and spatial dependence structure in residual components (Zhu et al. 2022). Therefore, the hybrid approaches integrate the advantages of ML algorithms and geostatistical models, which not only exploit the non-linear relationships between SOCD and predictor variables but also account for unexplained information of residuals, is considered a potential initiative to improve the prediction accuracy for estimating SOCD since they significantly mitigate errors arising from extraneous predictors and instability of predictive models (Song et al. 2017; Chen et al. 2019a; Lamichhane et al. 2019; Ho et al. 2024). Table 1 summarizes previous research on using geostatistics, ML algorithms, and their hybrid approaches in estimating SOCD in forest ecosystems. Despite the demonstrated superior performance, the hybrid approaches of geostatistics and ML algorithms have rarely been used (Zhu et al. 2022), whereas geostatistics and especially ML models appear to be utilized frequently to map SOCD in forest ecosystems that are characterized by complex terrain and dense vegetation coverage (Clough and Green 2013; Nussbaum et al. 2014; Scolforo et al. 2016; Ceddia et al. 2017; Dai et al. 2018; Wang et al. 2020b, a; Odebiri et al. 2020; Zhou et al. 2020a, b; Shafizadeh-Moghadam et al. 2022; Farooq et al. 2022; Meliho et al. 2023). Additionally, when the soil mapping community has focused extensively on exploring the efficacy of ML methods, there has been comparatively less research on the performance comparison between geostatistics and ML for SOCD mapping in forested areas (Veronesi and Schillaci 2019). Even within a recent work, Beguin et al. (2017) stated that kriging outperformed both RF and BRT in predicting soil properties in Canadian forests, causing a significant controversy surrounding whether ML is truly superior to geostatistical methods in this ecosystem category. Accordingly, it remains unclear which class of algorithms—geostatistics, ML, and hybrid approaches—performs most accurately in spatial modeling of SOCD in forest landscapes.

This study aimed to compare DSM techniques in mapping SOCD in Central Vietnamese tropical forests through three geostatistical models (OK, UK, and KED), two ML algorithms (RF and BRT), and two hybrid approaches of ML and geostatistics, namely random forest regression kriging (RFRK) and Boosted regression tree regression kriging (BRTK), utilizing a range of covariates derived from RS, climatic factors, topographic attributes and geographic position data. We expect that the findings of this study can aid in selecting the most appropriate and optimal DSM methods for tropical forest ecosystems, thereby promoting more frequent monitoring and assessment of SOC pools.

2 Materials and methods

2.1 Study area

The research site, Danang city, is situated in the South Central Coastal zone of Vietnam, lying between 107.81^0–108.34^0E and 15.92^0–16.22^0N, with the total mainland covering 960 km^2 (Fig. 1). The topography of Danang city gradually slopes downward from west to east. The western areas exhibit a predominantly mountainous terrain, while the eastern part is characterized by flat plains.

The city is predominantly forested, with the rest of the land cover primarily including settlement and crop areas. Specifically, forests cover about 658.91 km^2 (nearly 68.64% of the city's whole expanse) and are almost entirely distributed in mountainous regions, with altitudes from 0 to 1659 m. Due to having a tropical monsoon climate, climatic conditions in forests are characterized by mean annual precipitation of 1959–2523 mm and mean annual temperature of 11.1–26.2^{0}C. The dominant forest type is evergreen broadleaf vegetation, such as *Syzygium levinei (Merr.) Merr.*, *Scaphium lychnophorum (Hance) Pierre, Lithocarpus annamensis (Hickel & A. Camus) Barnett*, *Canarium littorale Blume*, and *Polyalthia nemoralis Aug. DC* (Huy et al. 2016). There are three main soil types within forest landscapes of the research area: Ferralsols (56.09%), Arenic Acrisols (31.77%), and Humic Acrisols (5.62%), with pH$_{KCl}$ values of 4.0–5.5 (National Institute of Agricultural Planning and Projection of Vietnam 2005).

2.2 Soil collection and analysis

Field data for this study were obtained over the course of three months from July to September 2023. Systematic unaligned sampling, which combines features of simple random and systematic sampling designs, is suitable for forest assessment (FAO 2015) and was used for determining sampling locations. The guiding for determining sampling locations consists of two fundamental steps: (1) based on the forest administrative map in 2013 provided by the Danang Department of Forest Protection, a 2.5 km × 2.5 km grid was generated, and (2) the geographic coordinate of sampling points was selected by using a random number generator in QGIS. The produced sampling points were later transferred to portable GPS receivers (Garmin GPSMAP 64) and utilized for navigation to the field site. 104 locations were surveyed during fieldwork. At each sampling location, two soil samples were taken from the 30-cm topsoil layer: one with soil cores and the other with a soil probe. The collected samples were then bagged, labeled, and dispatched to the laboratory for SOCD measurement.

Table 1 A literature summary of using geostatistics, ML algorithms, and their hybrid approaches in estimating SOCD in forest ecosystems

Study	Research site	Models	Findings
Clough and Green (2013)	New Jersey, USA	Geostatistics	OK, co-kriging, and regression kriging achieved acceptable accuracies with RMSE values ranging from 53.43 t.ha^{-1} to 67.90 t.ha^{-1}
Nussbaum et al. (2014)	Switzerland	Geostatistics	The predictive power of the KED model was moderate ($R^2 = 0.34$ and RMSE = 0.49 for SOC stock in 0–30 cm and $R^2 = 0.40$ and RMSE = 0.56 in 0–100 cm)
Scolforo et al. (2016)	Minas Gerais State, Brazil	Geostatistics	OK, co-kriging, and regression kriging presented satisfactory results, with R^2 ranging from 0.54 to 0.67 and MAE ranging from 0.58 to 0.63
Dai et al. (2018)	Zhejiang Province, China	Geostatistics	The OK method could be utilized to interpolate the spatial distribution of SOCD effectively
Wang et al. (2020a)	Liaoning province, China	ML algorithms	BRT provided considerable prediction accuracy, with R^2 ranging from 0.53 to 0.65 and RMSE ranging from 0.07 to 0.19
(Wang et al. 2020b)	Liaoning province, China	ML algorithms	BTR was compared to other traditional regression models, and BRT was more accurate, with $R^2 = 0.56$ and RMSE = 8.50 t.ha^{-1}
Odebiri et al (2020)	KwaZulu-Natal province, South Africa	ML algorithms	Four ML models including RF, Stochastic Gradient Boosting, Support Vector Machine, and Artificial Neural Network achieved substantial accuracies, with R^2 ranging from 0.77 to 0.84 and RMSE ranging from 0.77 t.ha^{-1} to 0.92 t.ha^{-1}. Among these, RF was the most accurate, with the highest R^2 and lowest RMSE
Zhou et al. (2020a)	Heihe River Basin, northwestern China	ML algorithms	The prediction accuracies of RF and BRT varied corresponding to the data combination. Specifically, RF had R^2 ranging from 0.19 to 0.75 and RMSE ranging from 0.55 to 1.01, meanwhile, BRT had R^2 ranging from 0.22 to 0.75 and RMSE ranging from 0.55 to 1.00
Zhou et al. (2020b)	the southern part of Central Europe	ML algorithms	Four machine-learners including RF, BRT, support vector machine, and Bagged CART were used to predict the spatial distribution of SOCD. Among them, BRT achieved the highest R^2 (0.44) and the lowest RMSE (0.57)

Table 1 (continued)

Study	Research site	Models	Findings
Shafizadeh-Moghadam et al. (2022)	Golestan province, Iran	ML algorithms	RF and support vector regression provided moderate accuracies, with support vector regression being more accurate than its counterpart. Specifically, R^2 of 0.43 and RMSE of 1.32% were in RF, and R^2 of 0.58 and RMSE of 0.94% were in support vector regression
Meliho et al. (2023)	The High Atlas Mountains, Morocco	ML algorithms	Cubist, RF, support vector machine, and gradient boosting machine were used, with Cubist (R^2 = 0.86, RMSE = 11.62 t.ha^{-1}) and RF (R^2 = 0.79, RMSE = 13.26 t.ha^{-1}) exhibiting the highest predictive power
Beguin et al. (2017)	Canada	Geostatistics and ML algorithms	Kriging techniques (OK and regression kriging) consistently had higher R^2 and lower RMSE values than ML models (boosted regression trees, random forests, Cubist, and weighted k-nearest neighbors)
Farooq et al. (2022)	Wangath watershed, Western Himalayas, India	Geostatistics and ML algorithms	OK, regression kriging, and RF were used for assessing the spatial distribution. Additionally, RF (R^2 = 0.90) performed better than ordinary kriging (R^2 = 0.53) and regression kriging (R^2 = 0.29)
Zhu et al. (2022)	Jiangsu Province, China	ML algorithms and hybrid approaches	The ML models (RF, Cubist, stochastic gradient boosting, and Bayesian regularized neural networks) had moderate accuracies, with R^2 values of 0.36, 0.45, 0.34, and 0.28, respectively. Compared to these ML models, the hybrid approaches of integrating ML models with the kriging technique showed an increase in R^2 by 0.36, 0.23, 0.39, and 0.34, respectively

The laboratory analysis of forest SOCD determination was conducted using the methodology outlined by Pearson et al. (2007). Soil samples in probes were air-dried at room temperature (approximately 25^{0}C) for seven days in a well-ventilated area to ensure consistent moisture removal, passed through a 2-mm sieve to eliminate debris and larger particles, and finely ground to achieve homogeneity for carbon concentration analysis. Carbon concentration in soil was measured by the dry combustion method with a CN-corder MT- 700 (Yanaco, Japan) (Liyanage et al. 2022). Specifically, the soil samples were combusted at around 950^{0}C in an oxygen-rich environment, converting carbon into CO_2, which was subsequently detected by a thermal conductivity detector to provide quantitative measurements. The instrument was calibrated before analysis using hippuric acid ($C_9H_9NO_3$) as a standard substance to ensure accuracy. Whereas, the soil samples in cores were oven-dried at 105^{0}C for 48 h and sieved through a 2-mm sieve for separating into coarse fragments and fine fractions to calculate the bulk density using Eq. (1).

$$SBD = \frac{ODM}{CV - (CF/RF)} \tag{1}$$

where: SBD is soil bulk density (g.cm^{-3}), ODM is oven-dry mass of fine fractions (g), CV is core volume (cm^3), CF is mass of coarse fragments (g), and RF is density of rock fragments (given as 2.65 g.cm^{-3}).

The SOCD content of soil samples was then measured by:

$$SOCD(t/ha) = SBD \times D \times C \tag{2}$$

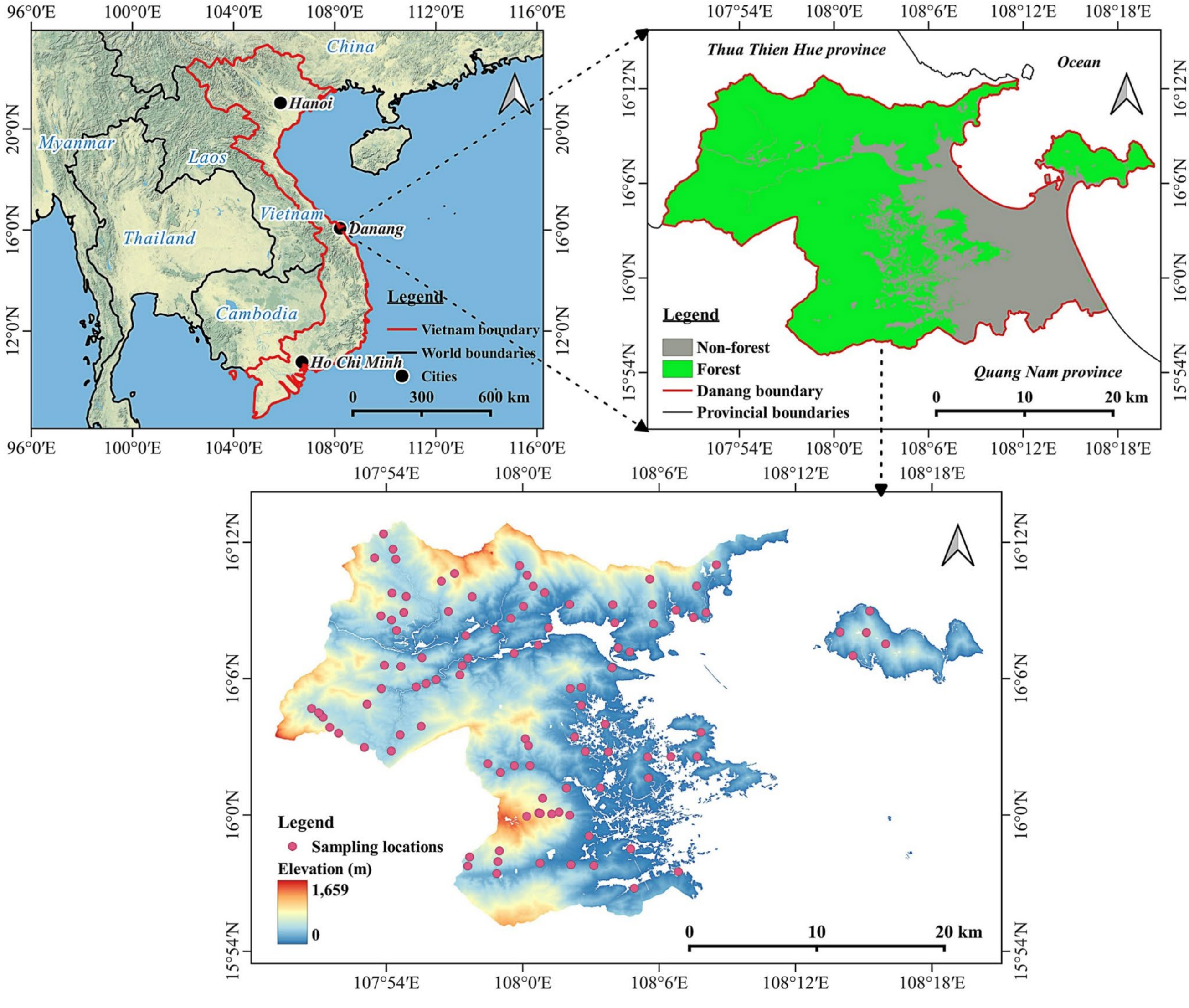


Fig. 1 Location of the research site and sampling points

where: D represents the thickness of the respective soil depth layer (cm), and C is the concentration of soil carbon (%).

2.3 Environmental covariates

In this study, thirty environmental covariates derived from RS data, terrain attributes, climatic factors, and geographic positions, as illustrated in Fig. 2, were utilized as predictors for estimating forest SOCD in ML algorithms. These covariates were projected to the WGS 84/UTM Zone 48 N coordinate reference system, resampled to a 10 m spatial resolution using bilinear interpolation, and co-registered to align with the pixel grid of the reference data (Sentinel- 2 Band 2) for further analysis.

2.3.1 Remote sensing data

The used RS data included Sentinel- 2 (S2) and Advanced Land Observing Satellite- 2/Phased Array L-band Synthetic Aperture Radar- 2 (ALOS- 2/PALSAR- 2). In forest landscapes, the effectiveness of RS in soil mapping is significantly limited by the dense vegetation that typically covers the forest floor, preventing RS sensors from directly detecting the underlying soils (Odebiri et al. 2020). In light of this context, numerous prior DSM research have utilized RS-based spectral bands and vegetation indices (VIs) as valuable predictors in predictive models for SOCD estimation in vegetation cover situations (Lamichhane et al. 2019). This method's efficacy is rooted in the ability of vegetation

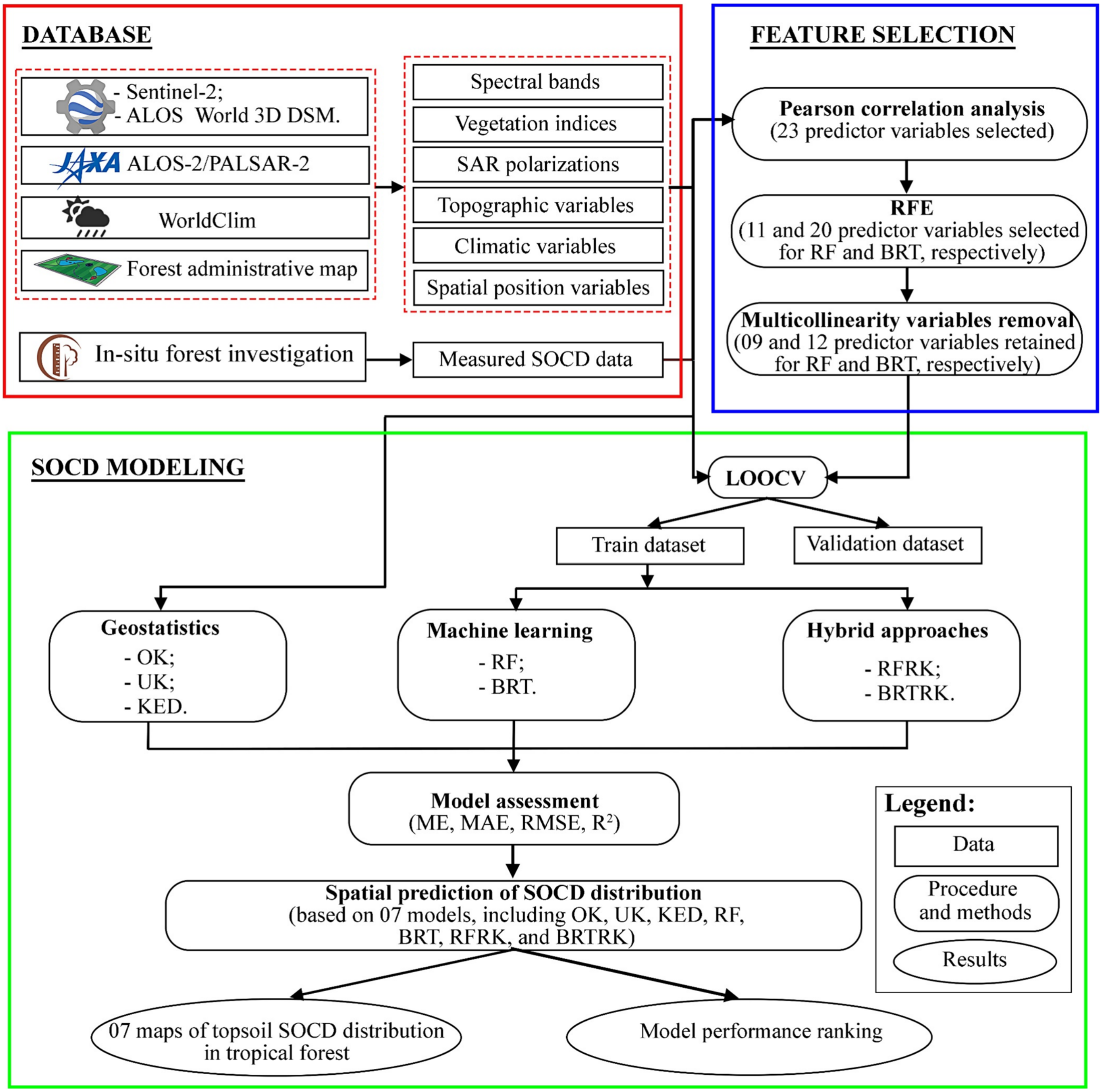

DATABASE
- Sentinel-2;
- ALOS World 3D DSM.
ALOS-2/PALSAR-2
WorldClim
Forest administrative map
Spectral bands
Vegetation indices
SAR polarizations
Topographic variables
Climatic variables
Spatial position variables
In-situ forest investigation
Measured SOCD data
FEATURE SELECTION
Pearson correlation analysis
(23 predictor variables selected)
RFE
(11 and 20 predictor variables selected for RF and BRT, respectively)
Multicollinearity variables removal
(09 and 12 predictor variables retained for RF and BRT, respectively)
SOCD MODELING
LOOCV
Train dataset
Validation dataset
Geostatistics
- OK;
- UK;
- KED.
Machine learning
- RF;
- BRT.
Hybrid approaches
- RFRK;
- BRTRK.
Model assessment
(ME, MAE, RMSE, R²)
Spatial prediction of SOCD distribution
(based on 07 models, including OK, UK, KED, RF, BRT, RFRK, and BRTRK)
07 maps of topsoil SOCD distribution in tropical forest
Model performance ranking
Legend:
Data
Procedure and methods
Results

Fig. 2 Flowchart of the comparison of geostatistics, ML algorithms, and hybrid models for estimating SOCD

to influence the spatial variability of soil characteristics through the modulation of soil biophysical processes (Zhou et al. 2020a).

The European Space Agency (ESA) S2 satellites are outfitted with a multispectral (MSI) instrument capable of detecting 13 spectral bands, offering spatial resolutions of 10, 20, and 60 m, and a revisit time of around four days (Vizzari 2022). S2 imageries was acquired from Google Earth Engine (https://code.earthengine.google.co.in). Having cloud-free coverage of S2 imagery for the research site was challenging because of the persistent cloud cover in tropical regions. Thus, a cloud-free Sentinel- 2 composite was generated for a period of approximately five months, spanning from May to September 2023, with a total of 270 images from both S2 A and S2B level- 2 A. The creation of this composite adhered to three fundamental steps: (1) Sentinel- 2 Surface Reflectance and Cloud Probability image collections in Google Earth Engine (GEE) were refined using spatial and temporal criteria, (2) a cloud masking function is employed to exclude pixels with a cloud probability over

30%, (3) a median composite image was generated from the filtered S2 datasets and the cloud-masked dataset. For this study, band 2 (B2), band 3 (B3), band 4 (B4), band 5 (B5), band 6 (B6), band 7 (B7), band 8 (B8), band 8 A (B8 A), band 11 (B11), and band 12 (B12) were used. Besides, five VIs including NDVI, GNDVI, RVI, SAVI, and EVI, were also computed and extracted from GEE (see Table 2). These VIs were proven as applicable variables for forest SOCD estimation (Wang et al. 2020a, b; Odebiri et al. 2020).

Along with MSI imagery, ALOS- 2/PALSAR- 2 imagery was also used due to its L-band's high sensitivity to forest structure (Shimada 2011). L-band ALOS- 2/PALSAR- 2 sensor leverage enhanced image contrasts achieved by multi-looking capabilities, allowing for the quantification of within-canopy properties and minimizing vulnerability to spectrum saturation relative to MSI data (Mutanga et al. 2023). Compared to other shorter-frequency SAR sensors (X- and C-band), the L-band possesses the capability to penetrate more deeply into vegetation canopies, making them resistant to the saturation issue in vegetation estimation (Ji et al. 2024). To encompass the full research region, two scenes of ALOS- 2/PALSAR- 2 level 2.1 fine beam dual (HV, HH) format image tiles with a 6.25-m spatial resolution on 4th June 2023 were obtained from the Japan Aerospace Exploration Agency (JAXA) via the https://www.eorc.jaxa.jp website. Before mosaicking, the images underwent preprocessing with the ESA SNAP toolbox. This was accomplished by transforming digital numbers (DN) into gamma-nought (γ^0) backscattering coefficients using the following equation:

$$\sigma^0 = 10.log_{10}(DN)^2 + CF \tag{3}$$

$$\gamma^0 = \frac{\sigma^0}{cos\varphi} \tag{4}$$

where σ^0 is sigma-nought backscattering coefficients (dB), φ is the incidence angle, and CF is the calibration factor. For ALOS- 2/PALSAR- 2 data, the CF is − 83.0 dB (Shimada et al. 2009).

The capacity of the difference and ratio between ALOS- 2/PALSAR- 2 backscatters for SOCD prediction was demonstrated (Wang et al. 2020a). Hence, HH and HV were used together with HH-HV and HH/HV as predictor variables in this study.

2.3.2 Topographic attributes

Geo-morphometric data are frequently utilized as predictors for SOCD in forest ecosystems (Clough and Green 2013; Wiesmeier et al. 2013; Nussbaum et al. 2014) because they can influence the hydrothermal conditions, which might potentially impact SOCD variation by speeding soil loss and runoff water drainage (Clough and Green 2013; Zhou et al. 2022). This research employed the ALOS World 3D-30 m digital surface model (DSM) from GEE to compute topographic attributes using the SAGA-GIS 9.4.1 software. Topographic variables consisted of aspect (ASPECT), channel network base level (CNBL), elevation (ELEV), slope (SLOPE), valley depth (VD), topographic wetness index (TWI), and terrain ruggedness index (TRI).

2.3.3 Climatic factors

Climate exerts influence over the amount of aboveground net primary output in tropical forests (Cleveland et al. 2011), which subsequently impacts the spatial variability of SOCD (Leff et al. 2012). Thus, we employed two climatic variables in the form of raster files, namely mean annual temperature (MAT) and mean annual precipitation (MAP), to estimate SOCD in the study area. These climate indices were sourced from the https://www.worldclim.org website, which was proposed by Fick and Hijmans (2017).

2.3.4 Geographical positions

Forest biomass and soil parameters exhibited considerable variations as the distance from the road to the inner part of the forest increased (Zhou et al. 2020c). In addition, rivers and streams can contribute to the ongoing deposition of soil materials and cause instability (Zhou et al. 2022). Therefore, two raster files were generated: the distance to roads (DTR) based on the road layer in the forest administrative map and the distance to rivers/streams (DTW) determined from DEM data. These variables were expected to correlate with SOCD in the study site.

Table 2 Details of Sentinel- 2 vegetation indices

Indices	Descriptions	References
NDVI	Normalized difference vegetation index:$(B8 - B4)/(B8 + B4)$	Tucker (1979)
GNDVI	Green normalized difference vegetation index:$B7 - B3/B7 + B3$	Gitelson and Merzlyak (1998)
RVI	Ratio vegetation index:$B8/B4$	Broge and Leblanc (2001)
SAVI	Soil adjusted vegetation Index:$(1 + 0.725) \times B8 - B4/B8 + B4 + 0.725$	Huete (1988)
EVI	Enhanced vegetation index:$2.5 * (B8 - B4)/(B8 + 6 \times B4 - 7.5 \times B2 + 1)$	Huete et al.(2002)

2.4 Feature selection

We employed correlation analysis and recursive feature elimination (RFE) algorithm to decrease data dimensionality while preserving the majority of valuable information. The pairwise Pearson's product-moment correlation analysis was applied to assess the relationships between predictors and SOCD observations, as well as to exclude any unnecessary factors for ML algorithms. The magnitude of correlation is assessed by the Pearson correlation coefficient (r). Specifically, r values of 0.00–0.29, 0.30–0.49, 0.50–0.69, 0.70–0.89, and 0.90–1.00 indicate little to no correlation, low correlation, moderate correlation, high correlation, and very high correlation, respectively (Asuero et al. 2006). Besides, r values greater than 0.80 were also employed to indicate the existence of multicollinearity issues among predictor variables from the same source (Chen et al. 2019b; Zhi et al. 2021). Furthermore, prior to implementing ML, the RFE algorithm was utilized for feature selection. This algorithm, combined with fivefold cross-validation, is known for its effectiveness in selecting a limited number of variables while maintaining high prediction accuracy (Shi et al. 2018; Veronesi and Schillaci 2019; Luo et al. 2022). RFE began by evaluating predictors based on the significance criteria of each ML algorithm, systematically removing the least important variable from the model until a single predictor persisted. Subsequently, the feature's optimal subset size was determined, characterized as the number of predictors yielding the largest negative Root Mean Square Error value.

2.5 Predictive models

2.5.1 Geostatistics

Kriging, described as the best linear unbiased predictor, employs the semivariogram to apply weights that assist in estimating the values of unsampled data (Goovaerts 1999; Oliver and Webster 2014). The semivariogram is constructed by calculating the average semivariances for all pairs of points within distance bins (Kravchenko and Bullock 1999), as follows:

$$\gamma(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \left\{ Z\left(x_i\right) - Z(x_{i+h}) \right\}^2 \tag{5}$$

where $\gamma(h)$ denotes semi-variance, $m(h)$ denotes the count of point pairs, x_i and x_{i+h} are sampling points separated by a distance h, $Z\left(x_i\right)$ and $Z(x_{i+h})$ represent measured values of variable Z at the respective points.

We examined spherical, Gaussian, and exponential theoretical models for fitting the sample semivariogram, and the three main parameters of nugget, sill, and range were then determined. Positive sill serves as an indicator of the positive substrate effects (Yao et al. 2019). The nugget illustrates the extent of spatial variability resulting from random elements (Goovaerts 1999). The nugget-to-sill (N/S) ratio is employed to quantify the intensity of spatial dependence, with a lower N/S value indicating a more pronounced spatial autocorrelation (Sun et al. 2019). A variable exhibits strong spatial dependence when the N/S is below 0.25, moderate spatial dependence when the N/S ranges from 0.25 to 0.75, and weak spatial dependence when the N/S exceeds 0.75 (Cambardella et al. 1994). In addition, the ratio of sample spacing over correlation range (SS/CR) was utilized as a measure for the effectiveness of a certain training set in capturing the spatial structure, with a smaller SS/CR indicating a better capturing of spatial structure (Zhu and Lin 2010). To achieve a high kriging interpolation accuracy, the assumptions of stationary and normality need to be met (Webster and Oliver 2007). Here, we used the intrinsic hypothesis for the stationarity assumption and the Kolmogorov–Smirnov test for the data normality assumption. In this study, three forms of kriging, namely OK, UK, and KED, were used to model SOCD.

OK is the predominant method used in soil science (Goovaerts 1999). Once the semivariogram is modeled, the value of a random variable at unsampled points $\widehat{Z}\left(x_0\right)$ is approximated through a linear weighted sum of n observations, as shown in Eq. (6) (Mishra et al. 2009).

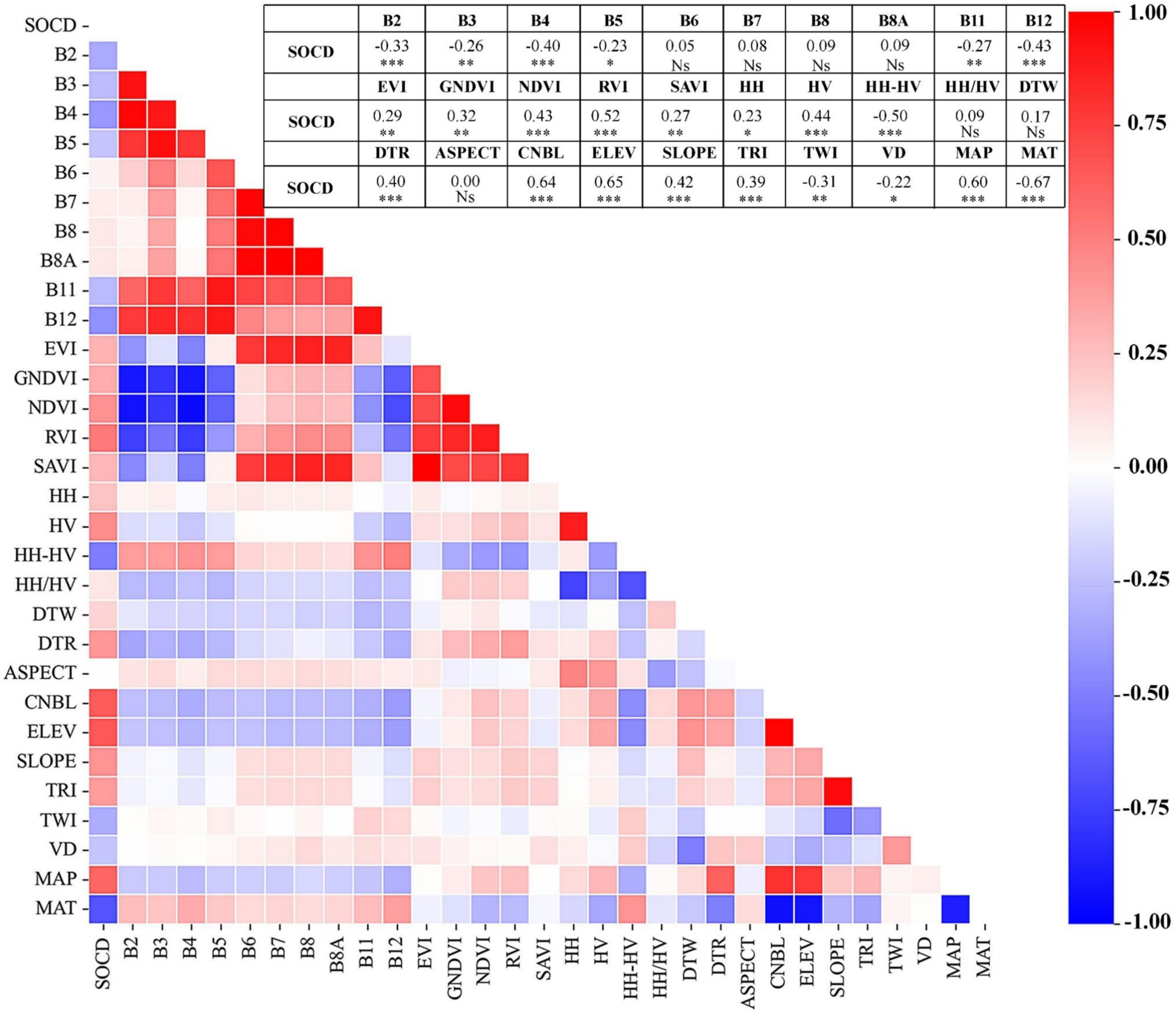
$$\widehat{Z}\left(x_0\right) = \sum_{i=1}^{n} \lambda_i Z\left(x_i\right),\ with \sum_{i=1}^{n} \lambda_i = 1 \tag{6}$$

Where λ_i represents the weights and $z(x_i)$ is the known value of variable Z at the sample location x_i.

OK presupposes that the mean of the variable of interest remains unchanged across the whole research site. However, in practical situations, the local mean may vary, making it necessary to use UK and KED. UK and KED, known as kriging with the trend, involve estimating both the deterministic trend component $m\left(x_i\right)$ and the parameters of the semivariogram model of the residual component $R(x_i)$, as depicted in Eq. (7) (Asa et al. 2012). The main difference between them is that UK accommodates a nonstationary mean where the expected value $\widehat{Z}\left(x_0\right)$ is a deterministic function of the spatial coordinates, whereas KED accommodates a nonstationary mean where $\widehat{Z}\left(x_0\right)$ is a deterministic function of external drift variables (Webster and Oliver 2007).

$$\widehat{Z}\left(x_0\right) = m\left(x_i\right) + R\left(x_i\right),\ with\ m\left(x_i\right) = \sum_{k=0}^{n} \beta_k y_k\left(x_i\right) \tag{7}$$

Where $y_k(x_i)$ are known deterministic functions, and β_k are unknown coefficients to be determined. In this study, the external drift variable was MAT, which had the strongest correlation to SOCD (Fig. 3). A first-order polynomial (which can mitigate erratic behavior at the data set's outer margins) model and linear regression model were used to set



	B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12
SOCD	-0.33 ***	-0.26 **	-0.40 ***	-0.23 *	0.05 Ns	0.08 Ns	0.09 Ns	0.09 Ns	-0.27 **	-0.43 ***
	EVI	GNDVI	NDVI	RVI	SAVI	HH	HV	HH-HV	HH/HV	DTW
SOCD	0.29 **	0.32 **	0.43 ***	0.52 ***	0.27 **	0.23 *	0.44 ***	-0.50 ***	0.09 Ns	0.17 Ns
	DTR	ASPECT	CNBL	ELEV	SLOPE	TRI	TWI	VD	MAP	MAT
SOCD	0.40 ***	0.00 Ns	0.64 ***	0.65 ***	0.42 ***	0.39 ***	-0.31 **	-0.22 *	0.60 ***	-0.67 ***

Fig. 3 Pearson correlation analysis for the relationship between forest topsoil organic carbon density and predictor covariates used in this study ('***', '**', '*', 'Ns' mean that p-values were below 0.001, 0.01, 0.05, and non-significant, respectively)

the trend in UK and KED, respectively. Geostatistical models were implemented by the gstat package in the R software.

2.5.2 Machine learning algorithms

RF, derived from CART, is a tree-based ML technique (Breiman 2001) that divide the training dataset using a sequence of if–then rules to determine probabilities for different classes (Veronesi and Schillaci 2019). During the training process, RF generates multiple trees without pruning, utilizing a distinct bootstrap sample from the original training data set (Wang et al. 2018). RF necessitates the usage of user-defined parameters in order to enhance their predictive accuracy (Yang et al. 2016). We implemented hyperparameter tuning by using grid search to identify three optimal parameters for RF, including the number of trees in the forest (*n_estimators*), the maximum number of features permitted for node splitting (*max_features*), and the minimum amount of data points eligible for in a leaf node (*min_samples_leaf*). In addition, RF utilizes certain observational values that are not employed in tree construction as the out-of-bag sample to calculate the associated relative error (Yang et al. 2016).

BRT is a sophisticated technique that combines the boosting technique with regression trees to enhance predicting accuracy (Zhou et al. 2020b). Specifically, the boosting strategy involves iteratively fitting regression trees by employing recursive binary splits to detect inadequately modeled observations in existing trees until a low model deviation is attained (Yang et al. 2016). Similar to RF, BRT also needs to ascertain the appropriate values for performing (Wang et al. 2018). In the BRT algorithm, we applied grid search for three parameters: 1. Number of trees (*n_estimators*), 2. Learning rate (*learning_rate*), and 3. Maximum depth of the individual regression estimators (*max_depth*).

For this study, hyperparameter tuning and ML algorithms were conducted via the scikit-learn package.

2.5.3 Hybrid approaches of machine learning and geostatistics

RFRK and BRTRK combined the strength of RF and BRT in prediction with the capability of OK in spatial interpolation, which was used with the expectation of enhancing the prediction accuracy for forest SOCD in the study area. Their implementations included two steps. First, RF and BRT were generated predicted SOCD values $\widehat{Z}_{RF/BRT}(x_i)$. Secondly, OK was applied to the residuals from RF and BRT $\widehat{\varepsilon}_{OK}$ and then added to $\widehat{Z}_{RF/BRT}(x_i)$ at the x_i locations. The SOCD values of RFRK and BRTRK ($\widehat{Z}_{RFRK/BRTRK}(x_i)$) were estimated according to Eq. (8).

$$\widehat{Z}_{RFRK/BRTRK}(x_i) = \widehat{Z}_{RF/BRT}(x_i) + \widehat{\varepsilon}_{OK} \tag{8}$$

2.6 Model validation

The performance of OK, UK, KED, RF, BRT, RFRK, and BRTRK was evaluated using a leave-one-out cross-validation (LOOCV) procedure. Three metrics, including mean absolute prediction error (MAE), root mean square error (RMSE), and coefficient of determination (R^2), were calculated for model accuracy evaluation. The smaller the MSE and MAE and the larger the R^2, the superior the predictive capability. Besides, we used mean error (ME) to identify whether the model tends to systematically over- or under-predict, with ME close to zero indicating an unbiased model.

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|P_i - O_i\right| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(P_i - O_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(O_i - \widehat{P}_i\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2} \tag{11}$$

$$ME = \frac{1}{n}\sum\nolimits_{i=1}^{n}(P_i - O_i) \tag{12}$$

where P_i is the predicted SOCD, O_i is the observed SOCD at the location i, O is the mean of the observed values, and n is the number of sampling locations.

3 Results

3.1 Descriptive statistics of forest soil organic carbon density

The descriptive statistics of SOCD at 30-cm topsoil in forests are shown in Table 3. The SOCD ranged from 43.40 t.ha^{-1} to 149.07 t.ha^{-1}, with a mean value and standard deviation of 96.97 t.ha^{-1} and 22.68 t.ha^{-1}, respectively. The skewness and kurtosis coefficients were − 0.15 and − 0.28, respectively, indicating that the field-based SOCD observations exhibited slight left skewness with a light tail.

3.2 Feature selection

Linear correlations between SOCD in forest topsoil and thirty quantitative predictors were calculated as shown in Fig. 3. We found correlations of twenty-three covariates (p-value < 0.05), including B2, B3, B4, B5, B11, B12, EVI, NDVI, GNDVI, RVI, SAVI, HH, HV, HH-HV, DTR, CNBL, ELEV, SLOPE, TRI, TWI, VD, MAP, and MAT with SOCD in the study site. According to Pearson correlation coefficients, it is worth noting that most of the variables derived from climate factors and topographic attributes had stronger correlations with SOCD than those

Table 3 Descriptive statistics of SOCD and residuals derived from UK, KED, RF, and BRT

Variables	Mean (t.ha^{-1})	Std. Deviatio n (t.ha^{-1})	Value Range (t.ha^{-1})	Skewness	Kurtosis	Kolmogorov–Smirnov	
						Statistic	**p-value**
SOCD	96.97	22.68	43.40—149.07	− 0.15	− 0.28	0.08	0.11
R$_{UK}$	0.00	21.15	− 48.37—44.11	− 0.13	− 0.29	0.06	0.20
R$_{KED}$	0.00	16.93	− 39.77—40.86	− 0.07	− 0.23	0.05	0.20
R$_{RF}$	0.07	13.50	− 26.79 – 34.90	0.23	− 0.47	0.08	0.08
R$_{BRT}$	− 0.80	13.28	− 36.61 – 31.62	0.02	0.00	0.06	0.20

extracted from RS and geographical position data. In particular, MAT and ELEV exhibited the highest correlations, with r values of − 0.67 and 0.65, respectively. Besides, absolute r values of these twenty-three environmental covariates ranged from 0.22 to 0.67, indicating the presence of little to moderate correlation with forest SOCD. It also suggested that non-linear modeling was essential to leverage the relationship between SOCD and its environmental covariates.

The RFE method was employed to identify the most critical predictors from significantly correlated variables, subsequently used as input variables for RF and BRT. Figure 4 shows the change in negative RMSE with the increasing variable number. The highest negative RMSE values of variable groups were − 14.70 and − 14.83 for RF and BRT, respectively. The results indicated that the optimal variable number for RF was eleven, while BRT achieved the most efficient performance with twenty variables. The retained variables for RF and BRT are listed in Fig. 4.

To avoid multicollinearity in the ML models, pairs of variables from the same data source with absolute r values greater than 0.8 were identified. For each pair, the variable with lower relative importance in Fig. 6a was withdrawn. As a result, two variables (CNBL and MAP) were eliminated from the set of eleven optimal variables for RF, and eight variables (TRI, CNBL, B12, NDVI, GNDVI, EVI, MAT, and HH) were removed from the set of twenty optimal variables for BRT.

3.3 Deterministic component modeling for SOCD estimation

Following feature selection through correlation analysis and RFE, the deterministic components of forest SOCD in our study were generated by UK (Fig. 5a), KED (Fig. 5b), RF (Fig. 9d), and BRT (Fig. 9e). Regarding geostatistical models, the UK employed a first-order polynomial model to derive trend component between field-based SOCD observations and their corresponding coordinates (X, Y). According to the implementation of the polynomial model, the coefficients of intercept and Y variable were omitted as their p-values exceeded 0.05 (Fig. 5a). Also, in order to derive trend components between field-based SOCD observations and the MAT variable, KED utilized a linear regression model. Once this model was implemented, the coefficients of both the intercept and MAT variables were found to be less than 0.05 (Fig. 5b), and hence, they were retained in the prediction model.

In the case of ML algorithms, nine optimal environmental covariates for RF and twelve optimal environmental covariates for BRT were used to generate the trend components of forest SOCD. The hyperparameter tuning process provided optimal external configuration parameters for both models. For RF, the best performance was achieved with *n_estimators* value = 300, *max_features* value = 4, and *min_samples_leaf* value = 1, whereas BRT gained optimal performance with *n_estimators* value = 400, *learning_rate* value = 0.35, and max_depth value = 3. For SOCD estimation using RF

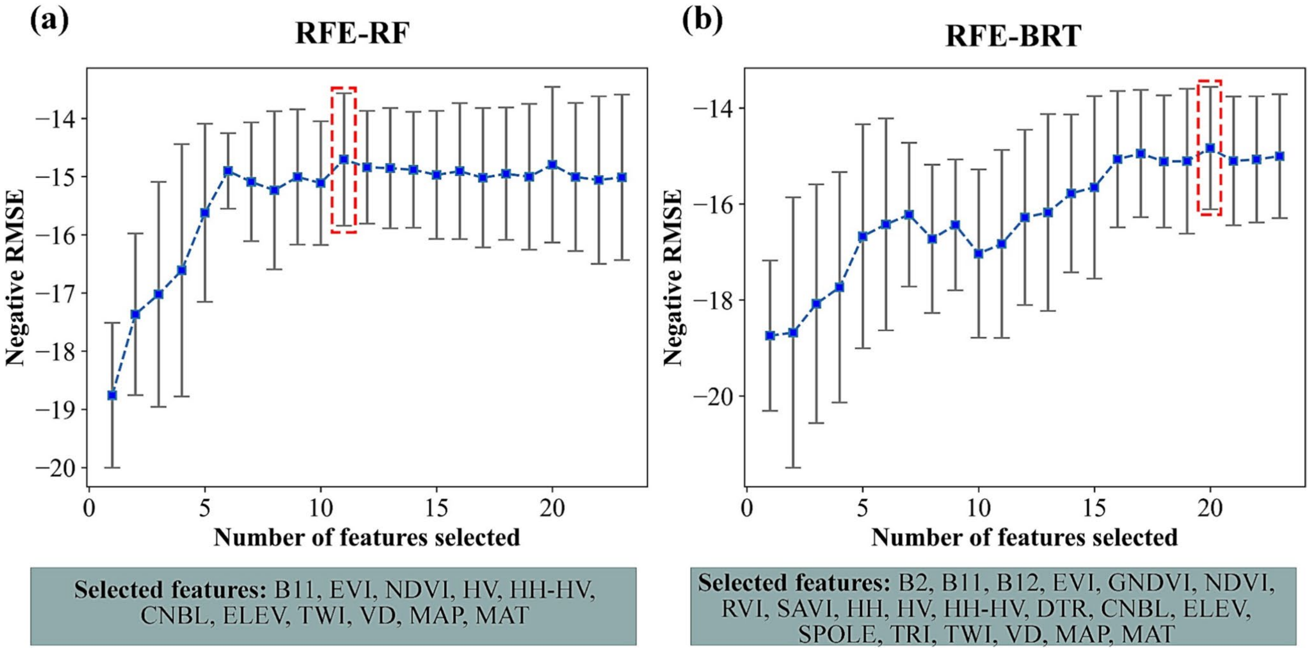
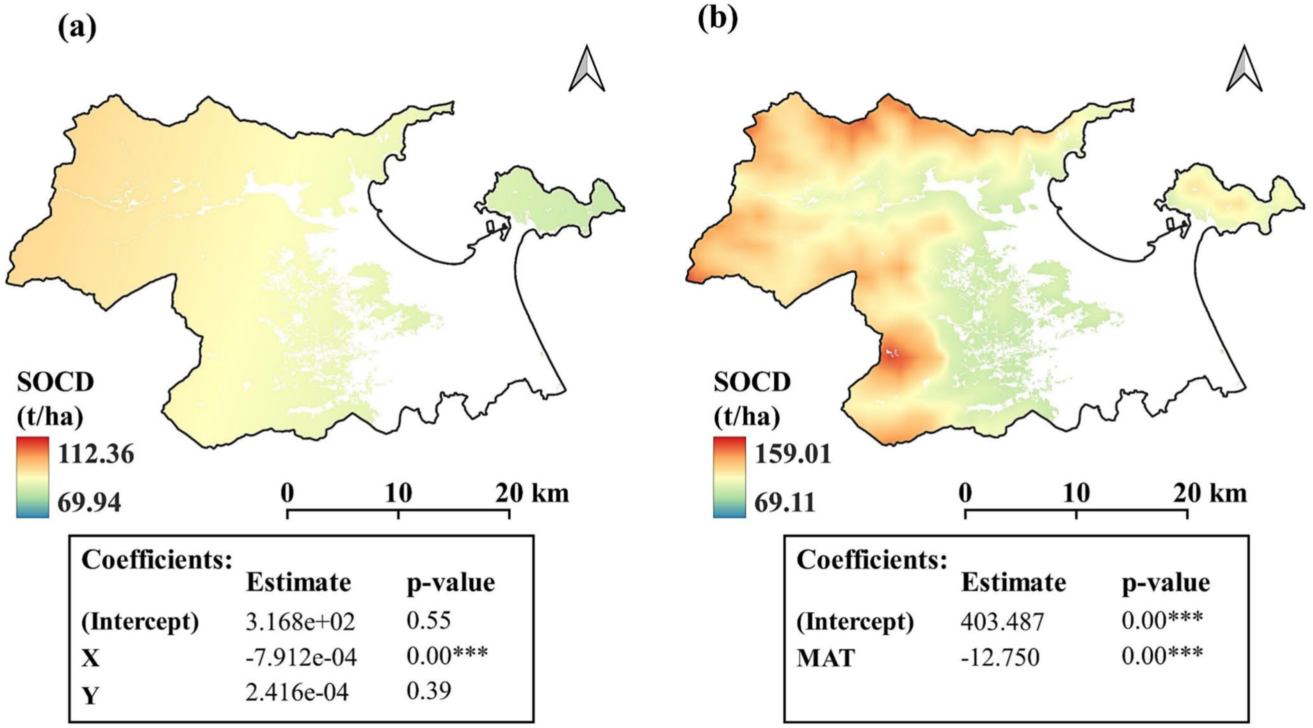


Fig. 4 RFE implementation for RF (**a**) and BRT (**b**) in selecting variables



Coefficients:

	Estimate	p-value
(Intercept)	3.168e+02	0.55
X	-7.912e-04	0.00***
Y	2.416e-04	0.39

Coefficients:

	Estimate	p-value
(Intercept)	403.487	0.00***
MAT	-12.750	0.00***

Fig. 5 Spatial distribution and model summary of the deterministic components of UK (**a**) and KED (**b**)..*** indicates that p-values were below 0.001

and BRT, the ranking of environmental covariates after multicollinearity analysis ordered by relative importance is illustrated in Fig. 6b. The top three variables for RF were ELEV, MAT, and NDVI, while for BRT, they were ELEV, MAP, and HV. Notably, ELEV was consistently the most crucial predictor in both models to estimate forest SOCD within our study site.

3.4 Semivariogram analysis of SOCD and residuals from predictive models

The study used field-based SOCD data and residuals derived from the deterministic functions of UK, KED, RF, and BRT for the semivariogram analysis. The outcomes of the Kolmogorov–Smirnov test in Table 3 and histograms in Fig. 7a, b, c, d, e show that the field-based SOCD data (skewness = − 0.15, kurtosis = − 0.28), as well as the residuals from UK (R_{UK}, skewness = − 0.13, kurtosis = − 0.29), KED (R_{KED}, skewness = − 0.07, kurtosis = − 0.23), RF (R_{RF}, skewness = 0.23, kurtosis = − 0.47), and BRT (R_{BRT}, skewness = 0.02, kurtosis = 0.00), all had p-values greater than 0.05 and displayed bell-shaped curves, indicating all mentioned variables possessed a normal distribution. Additionally, the semivariogram cloud of these five variables (Fig. 7f, g, h, i, j) exhibited no significant trends in semivariance (the mean value of semivariance was constant over the research area), suggesting that the intrinsic stationarity assumption was satisfied.

After confirming normality and stationarity assumptions, these variables were directly utilized in the computation of experimental semivariograms. From the visual analysis, there were spatial autocorrelation structures in the SOCD, UK residuals, KED residuals, RF residuals, and BRT residuals (Fig. 7k, l, m, n, o). Table 4 shows the basic parameters of the semivariogram models. The function exhibiting the maximum R^2 value alongside the minimum MAE and RMSE values was chosen for the best-fit experimental semivariogram models. As a result, the semivariogram of the ML residuals used the Gaussian function, while that of SOCD and residuals from geostatistical models used the exponential function, except for the UK, which used the spherical function. The sill values were all positive, ranging from 189.77 to 595.67, suggesting a variety of positive substrate effects. The nugget values of R_{RF} and R_{BRT} (101.81 and 111.36, respectively) were higher than those of SOCD, R_{UK}, and R_{KED} (0.00, 23.09, and 0.00), revealing that residuals generated by ML algorithms had a stronger spatial variability attributed to random elements (e.g., undetectable and inherent variability), in comparison to SOCD and residuals from geostatistical models. SOCD, R_{UK}, and R_{KED} had N/S ratios of 0.00, 0.05, and 0.00, respectively, signifying the presence of strong spatial dependence, while R_{RF} and R_{BRT} had N/S ratios of 0.51 and 0.59, indicating moderate spatial dependence. The N/S value for SOCD was lower than that for most residuals derived from the regression algorithms,

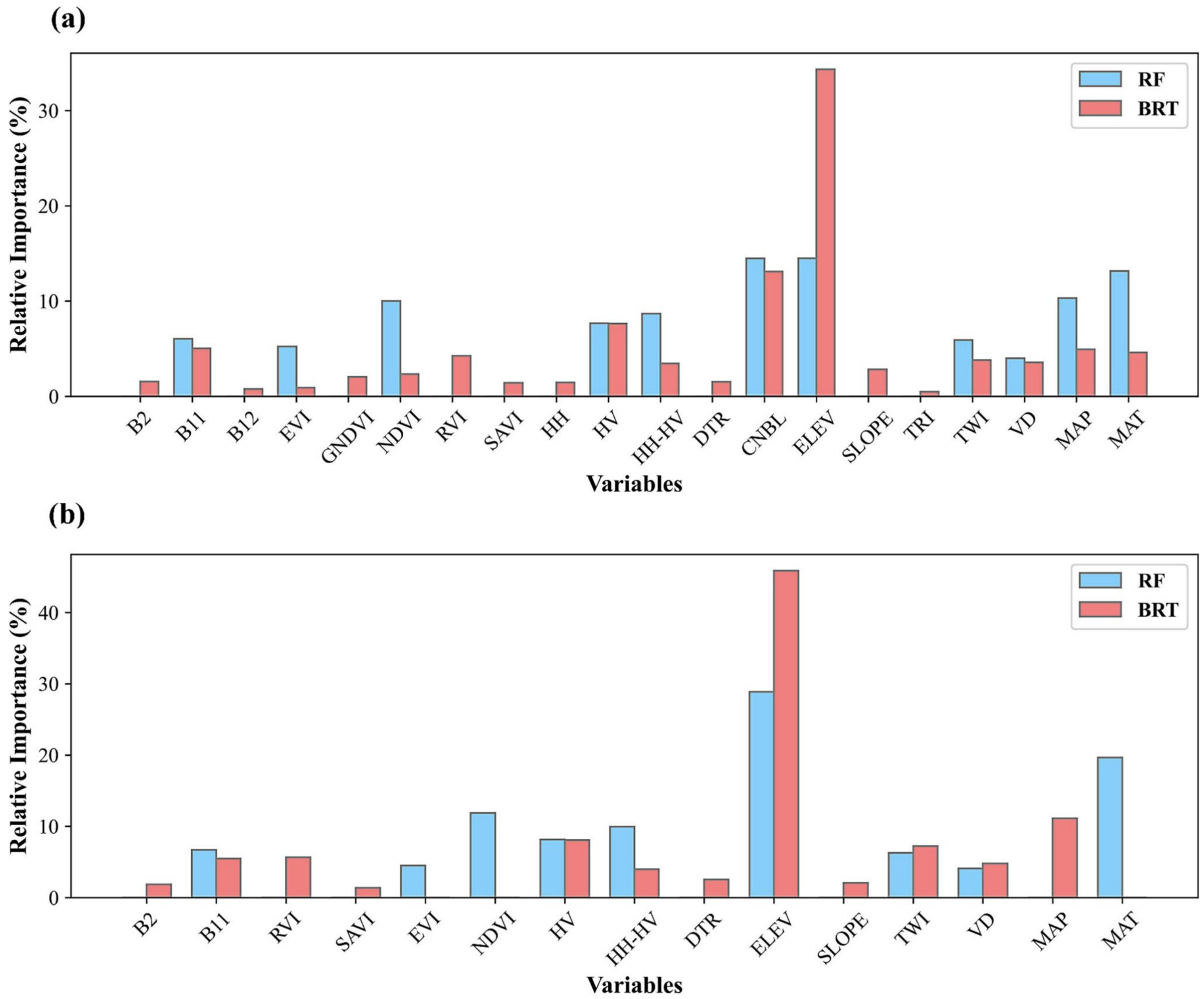


Fig. 6 The relative importance of the environmental covariates in ML models before multicollinearity variables elimination (**a**) and after multicollinearity variables elimination (**b**)

indicating a reduction in the degree of spatial dependence after accounting for deterministic trends. Among residual components, N/S values in R_{UK} and R_{KED} were smaller than those in R_{RF} and R_{BRT}, suggesting that residuals from geostatistical models had stronger spatial dependence than those from ML models.

Based on the parameters of experimental semivariogram models (Table 4), we interpolated the spatial distribution of R_{UK} (Fig. 7p), R_{KED} (Fig. 7q), R_{RF} (Fig. 7r), and R_{BRT} (Fig. 7s). The ranges of R_{UK} (from − 45.36 to 40.94 $t.ha^{-1}$) and R_{KED} (from − 38.32 to 37.19 $t.ha^{-1}$) were considerably larger than R_{RF} (from − 16.14 to 23.35 $t.ha^{-1}$) and R_{BRT} (from − 19.66 to 14.07 $t.ha^{-1}$), indicating ML performed better than geostatistics in forest SOCD estimation with respect to accuracy in our research site.

3.5 Accuracy assessment and comparison of geostatistics, ML algorithms, and their hybrid approaches

The results of LOOCV for accuracy evaluation are presented in the numeric format in Table 5 via three metrics, namely MAE, RMSE, and R^2. Overall, in our study site, the hybrid approaches combining ML with geostatistics performed best for forest SOCD estimation, followed by ML algorithms, with geostatistical models performing the least effectively. Among the used seven predictive models, RFRK achieved the greatest accuracy and the low errors (R^2= 0.71, MAE = 10.28 $t.ha^{-1}$, and RMSE = 12.23 $t.ha^{-1}$), whereas the opposite pattern was observed in OK (R^2= 0.53, MAE = 12.69 $t.ha^{-1}$, and RMSE = 15.54 $t.ha^{-1}$).

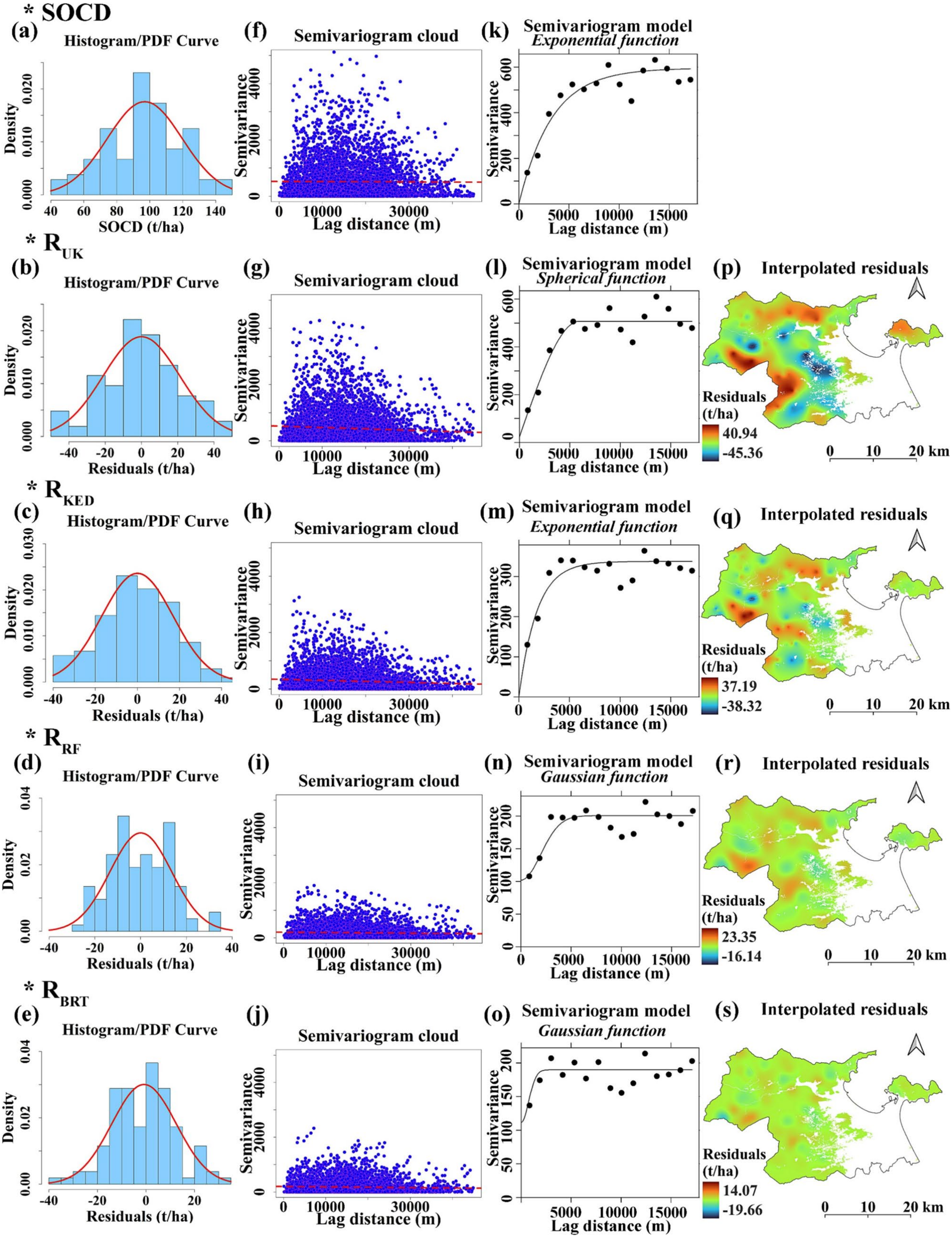


Fig. 7 Histograms/Probability density functions (**a, b, c, d, e**), semivariogram clouds (**f, g, h, i, j**), and semivariogram analysis (**k, l m, n, o**) of SOCD, R_{UK}, R_{KED}, R_{RF}, and R_{BRT}; Spatial distribution of R_{UK}, R_{KED}, R_{RF}, and R_{BRT} (**p, q, r, s**)

Table 4 Parameter estimations for semivariogram analysis of SOCD and residuals from UK, KED, RF, and BRT

Parameters	Theoretical models	Nugget	Sill	N/S	Range (m)	MAE	RMSE	R^2
SOCD	Exponential	0.00	595.67	0.00	3269.15	12.69	15.54	0.53
	Gaussian	88.63	552.69	0.16	2873.67	12.74	16.12	0.49
	Spherical	25.07	544.14	0.05	6358.38	12.56	15.70	0.52
R_{UK}	Exponential	0.00	549.70	0.00	2996.03	12.66	15.42	0.46
	Gaussian	109.39	515.15	0.21	3210.44	12.66	15.69	0.44
	Spherical	23.09	507.41	0.05	5940.31	12.40	15.30	0.47
R_{KED}	Exponential	0.00	337.68	0.00	1765.85	12.04	14.60	0.25
	Gaussian	93.07	331.75	0.28	1715.24	12.22	14.83	0.23
	Spherical	48.75	328.68	0.15	4505.97	11.97	14.65	0.24
R_{RF}	Exponential	41.90	201.45	0.21	1650.27	10.40	12.34	0.16
	Gaussian	101.81	200.47	0.51	2746.65	10.28	12.23	0.17
	Spherical	68.13	197.52	0.34	4329.88	10.32	12.26	0.17
R_{BRT}	Exponential	6.10	189.30	0.03	671.51	10.61	12.96	0.04
	Gaussian	111.36	189.77	0.59	1045.04	10.67	12.88	0.05
	Spherical	98.26	190.35	0.52	2919.34	10.70	12.90	0.05

Table 5 Evaluation metrics for geostatistics, machine learning algorithms, and hybrid approaches

Modeling techniques	ME ($t.ha^{-1}$)	MAE ($t.ha^{-1}$)	RMSE ($t.ha^{-1}$)	R^2
Geostatistics				
OK	0.17	12.69	15.54	0.53
UK	0.12	12.40	15.30	0.54
KED	− 0.10	12.04	14.60	0.58
ML algorithms				
RF	− 0.07	11.36	13.44	0.65
BRT	0.80	10.69	13.24	0.66
Hybrid approaches				
RFRK	− 0.10	10.28	12.23	0.71
BRTRK	− 0.07	10.67	12.88	0.67

Although being the least accurate in the forest SOCD compared to ML algorithms and the hybrid approaches, three geostatistical models achieved acceptable prediction accuracy. Among these, KED that used an external drift variable (MAT) instead of an internal drift variable (spatial coordinate) as in OK and UK to build the deterministic function, was superior over others, with $R^2 = 0.58$, MAE = 12.04 $t.ha^{-1}$, and RMSE = 14.60 $t.ha^{-1}$. When it comes to ML algorithms, the findings indicate that both RF and BRT had strong predictive capabilities, evidenced by the R^2 values ranging from 0.65 to 0.66, the MAE values ranging from 10.69 to 11.36 $t.ha^{-1}$, the RMSE values ranging from 13.24 to 13.44 $t.ha^{-1}$. After additionally interpolating residuals by OK as the spatial autocorrelation, ML modeling plus residual kriging methods enhanced the prediction accuracy in the comparison of ML models. Specifically, RFRK and BRTRK reduced prediction errors (MAE and RMSE) and increased R^2. MAE, RMSE, and R^2 were 10.28 $t.ha^{-1}$, 12.23 $t.ha^{-1}$, and 0.71 for the former and were 10.67 $t.ha^{-1}$, 12.88 $t.ha^{-1}$, and 0.67 for the latter, respectively. The prediction accuracy order of the seven models, from highest to lowest, was RFRK, BRTRK, BRT, RF, KED, UK, and OK.

Regarding ME, the OK, UK, and BRT models had positive ME values, supporting the fact that these predictive models tended to overestimate the observed values on average. The opposite pattern was recorded in the KED, RF, RFRK, and BRTRK models when their ME values were negative, suggesting that these models had a tendency to underestimate the observed values on average. Notably, the highest absolute ME value was found in BRT (0.80). Consequently, although BRT provided better accuracy than its ML counterpart (RF) and geostatistical models (OK, UK, and KED) in terms of MAE, RMSE, and R^2, it had a greater tendency to bias. However, incorporating interpolated residuals into BRT's deterministic trend reduced the absolute ME values to 0.07. Overall, the ME values of geostatistics, ML algorithms, and hybrid approaches were all relatively close to zero, illustrating the absence of significant problems with the models' biases.

The performance of predictive models was also illustrated through a graphical format (Fig. 8). The predicted-observed fit lines of ML algorithms along with residual kriging were closer to 1:1 line than that of both ML algorithms and geostatistical models. Additionally, compared to geostatistics, the slope of the fit lines for hybrid approaches increased significantly, indicating a substantial improvement in prediction accuracy. In contrast, the slope of the fit lines for ML methods changed less compared to geostatistics, reflecting an improvement in forest SOCD estimation, though not as high as that observed for the hybrid approaches, yet still outperforming geostatistics.

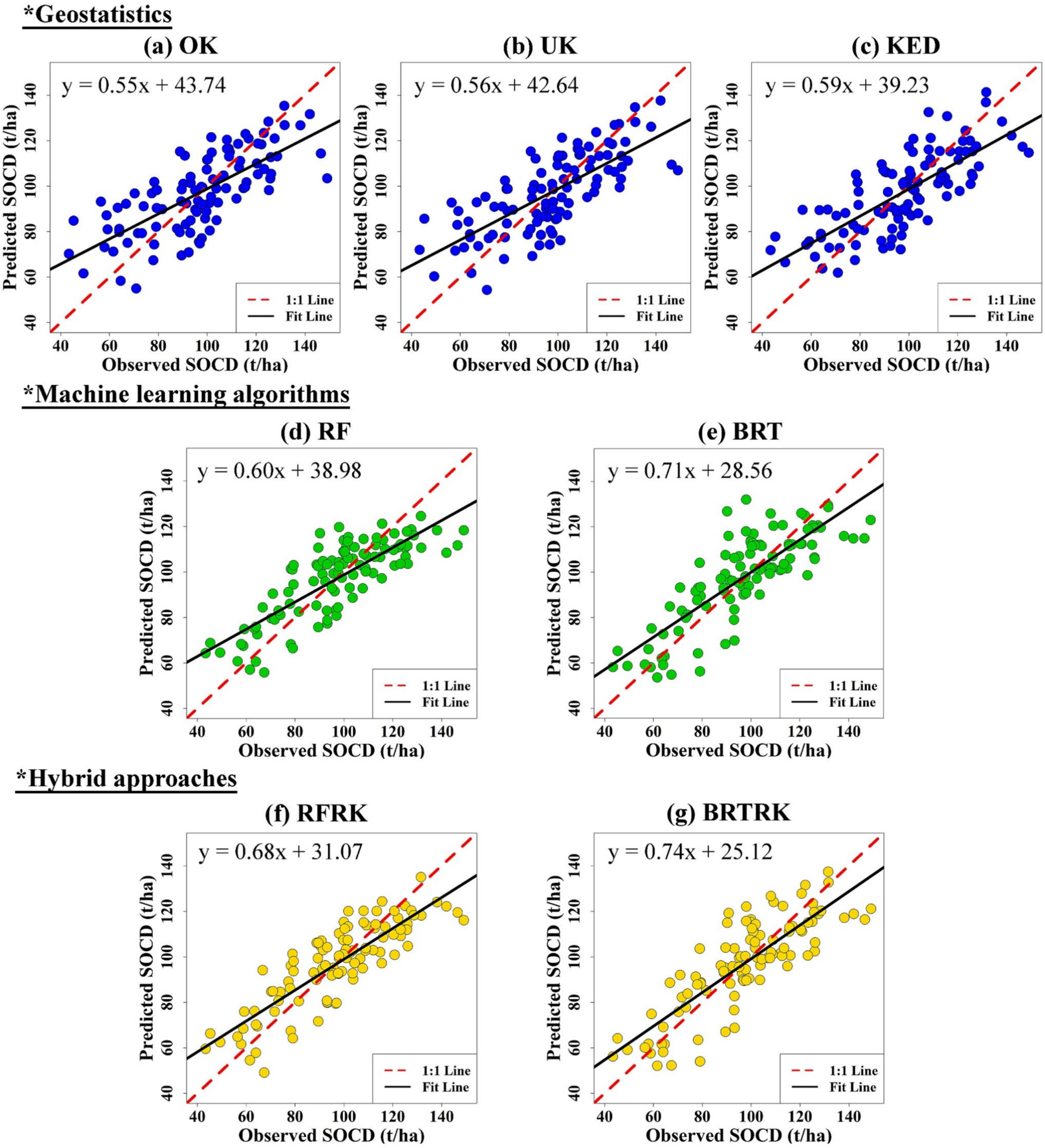


Fig. 8 Observed vs. predicted values for soil organic carbon density based on OK (**a**), UK (**b**), KED (**c**), RF (**d**), BRT (**e**), RFRK (**f**), and BRTRK (**g**). Blue dots, green dots, and yellow dots present values of geostatistics, ML algorithms, and hybrid approaches, respectively

3.6 Spatial distribution of forest SOCD

The spatial pattern of forest SOCD using OK, RF, and BRT was generated directly after semivariogram analysis or trend modeling, whereas spatial distributions of forest SOCD based on UK, KED, RFRK, and BRTRK were obtained by combining Fig. 5a with Fig. 7p, Fig. 5b with Fig. 7q, Fig. 9d with Fig. 7r, and Fig. 9e with Fig. 7s, respectively. Significantly, there were no considerable discrepancies in the spatial distribution of forest SOCD across the seven utilized

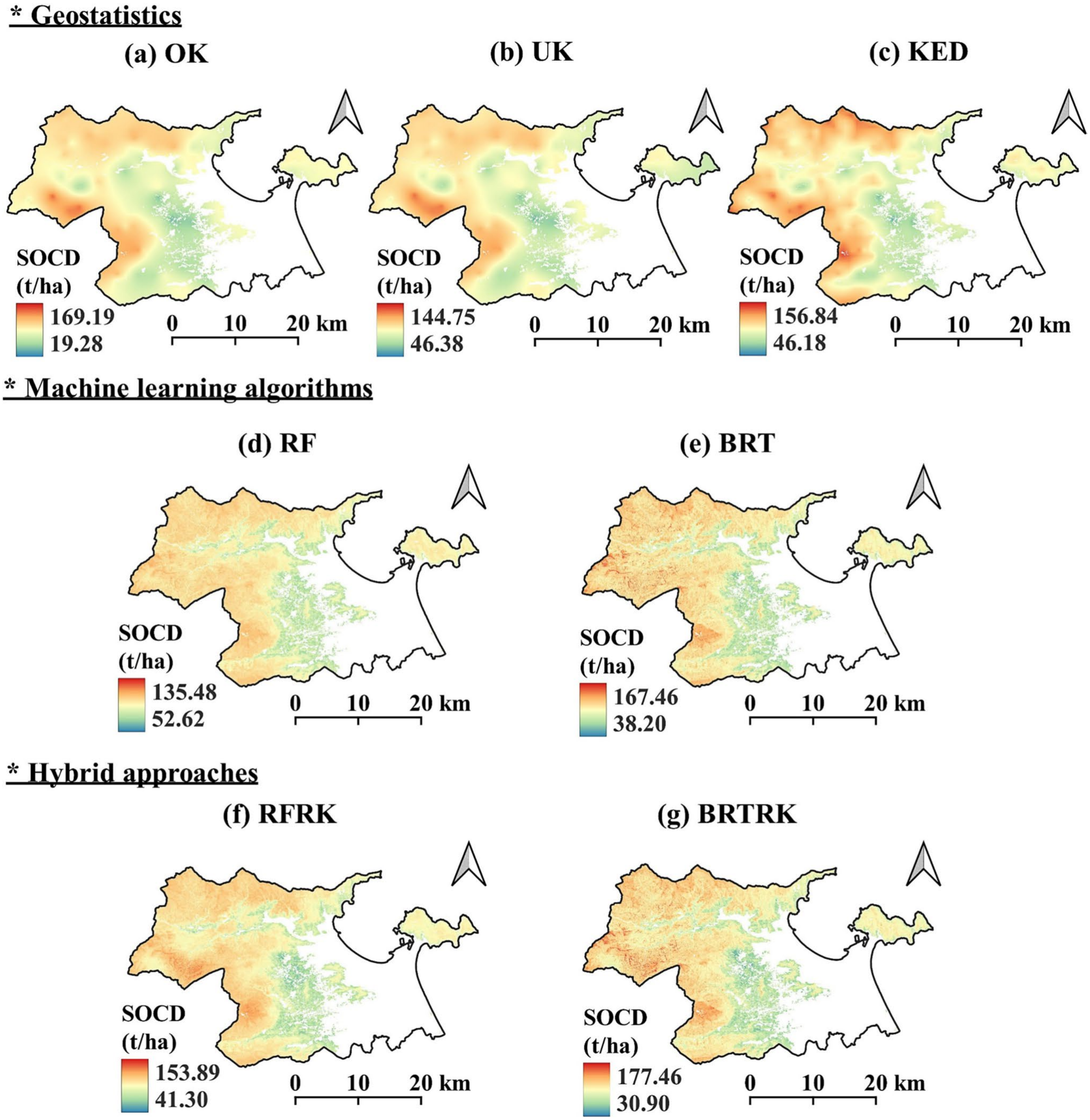


Fig. 9 Spatial distribution of the estimated SOCD in the top 30 cm of forest surface using OK (**a**), UK (**b**), KED (**c**), RF (**d**), BRT (**e**), RFRK (**f**), and BRTRK (**g**)

predictive models (Fig. 9). Nevertheless, while these maps exhibited a comparable spatial distribution pattern, the forest SOCD values exhibited considerable dissimilarity.

It is evident that the forest SOCD exhibited a clear and consistent decline from west to east directions, which closely aligned with the topography characteristics of the research site. The largest SOCD was recorded in the western and northwestern zones, encompassing a significant expanse of predominantly forested areas managed by the Danang Department of Forest Protection. Contradictorily, SOCD values were lowest in the middle of Danang city where forested and non-forested areas intersect, likely due to lower cover of vegetation and the impact of intensive anthropogenic activities. Besides, based on the residual maps (Fig. 7), SOCD maps (Fig. 9), and DEM (Fig. 1), it can be observed that all models used for estimating SOCD underestimated high SOCD values in high-altitude mountainous areas and overestimated low SOCD values in even terrain areas.

4 Discussion

The primary findings in our study indicate that, regardless of the influence of environmental covariates, the selection of the predictive models could substantially affect the accuracy of predicting SOCD in tropical forest ecosystems. From the study results, it is worth noting that selecting one suboptimal model could result in diminished predictive efficacy by approximately 25.35%, 23.44%, and 27.06% in terms of R^2, MAE, and RMSE, respectively, when comparing the OK model and the RFRK model (Table 5).

Our results show that RF and BRT had greater accuracy and lower errors than OK, UK, and KED, indicating that ML algorithms comparatively outperformed geostatistics in forest SOCD estimation. This conclusion contradicted the results of Beguin et al. (2017), who evaluated the prediction effectiveness of four ML systems (including BRT, RF, Cubist, and weighted k-nearest neighbors) and OK in Canadian forests, reported that the kriging method exhibited greater accuracy than all ML-based models. The success of ML over geostatistics in our study can be explained by two main reasons. Firstly, given that the geographical location of sampling points was suboptimal for geostatistics, it was unsurprising that kriging did not perform perfectly. The prerequisites regarding sampling for kriging and ML differ significantly: geostatistics necessitates a sampling strategy that is intended for accurately accommodating variogram fitting, while ML involves a design to ensure thorough coverage of all predictors (Veronesi and Schillaci 2019; Suleymanov et al. 2024). In spatial soil sampling for geostatistics, the sampling design is a crucial determinant in achieving effective outcomes in soil mapping endeavors (van Groenigen et al. 1999). Given knowledge about the variogram of the property being studied, it is feasible to enhance the sampling technique to minimize an objective function that is associated with the error in the survey. However, the knowledge of the variogram is seldom available before sampling unless it has been calculated based on a preliminary survey or for the similarity in soil characteristics under comparable circumstances (Wadoux et al. 2019), which was not the situation in this investigation. For this reason, we chose to employ a regular grid as a spatial coverage scheme instead, which can take account of the spatial dependence of soil parameters in specific sampling regions and minimize the maximum kriging error (Di et al. 1989; Webster and Oliver 2007). However, implementing a sampling strategy based on geostatistics in tropical forest landscapes was problematic due to access issues such as complex terrain and dense vegetation (Barker and Pinard 2001; Clough and Green 2013). Some identified locations from the initial sampling design could not be reached and were either eliminated or substituted with the farthest accessible one toward these locations, resulting in inflated sampling bias (the sampling locations were densely clustered in some regions and sparsely distributed across the study site) and increased sample spacing. Consequently, the kriging standard error rose, and the spatial structure was not accurately represented due to the elevated SS/CR (Di et al. 1989; Goidts et al. 2009; Zhu and Lin 2010; Clough and Green 2013). In contrast to geostatistics, RF and BRT models have proven useful in predicting soil properties in challenging mountainous terrains with limited accessibility for soil sampling and places with very low soil sample density (Lamichhane et al. 2019). The possible explanation for these ML algorithms' advantage is that, like standard regression, they typically emphasize optimization within feature space rather than considering the geographic distribution of observations (Brus and Heuvelink 2007). Secondly, from the literature, it is proven that when a non-linear relationship exists between the target variable and predictor covariates, and the target variable presents spatial correlation, ML algorithms outperform geostatistics in predictive accuracy (Fouedjio and Klump 2019). In our study, the predictor variables used for modeling were mostly derived from RS data, which have been demonstrated to have a non-linear relationship to SOCD (Guo et al. 2015; Akbari et al. 2021; Abdoli et al. 2023). The little to moderate correlations between field-based SOCD measurements and environmental variables, as shown in Fig. 3, further confirmed non-linear relationships among them. Additionally, SOCD is considered a regionalized variable and often tends to be spatially correlated (Di et al. 1989; Bergstrom et al. 2001; Dai et al. 2018).

Despite the differing methods, the forest SOCD maps produced by geostatistics and ML algorithms were generally analogous (Fig. 9). However, the maps of predicted forest SOCD based on ML, particularly decision tree-based algorithms like RF and BRT in this study, showed that ML methods were inclined to produce more detailed patterns and the abrupt changes of SOCD level in the forested area. It is attributed to the ML's key feature of mainly learning from data to make the prediction, making it sensitive to the used predictor covariates such as elevation and climate factors (Erdogan Erten et al. 2022). Meanwhile, the kriged SOCD generated from OK, UK, and KED had a smoother pattern since kriging techniques are able to minimize the variability of the estimates relative to that of the observations (Oliver and Webster 1990; Veronesi and Schillaci 2019).

It is clear from Table 5 and Fig. 8 that the performance of RF and BRT appeared inferior to the performance of RFRK and BRTRK, suggesting a greater predictive efficacy of the hybrid approaches relative to ML algorithms in tropical forest SOCD estimation. This finding is in accordance with the previous research (Guo et al. 2015; Silatsa et al. 2020; Suleymanov et al. 2023), which was also conducted under forest vegetation. Theoretically, the application of ML techniques for spatial estimation is problematic, because they presuppose the data are spatial uncorrelatedness (Erdogan Erten et al. 2022). As a result, the spatial autocorrelation structures are overlooked and persist in the residuals (Liu et al. 2022; Zhu et al. 2022). From Table 4 and Fig. 7, it is noticeable that spatial autocorrelation structures still existed in the RF and BRT's residual components, which means that RF and BRT could not properly extract completely structured information of forest SOCD. Guo et al. (2015) demonstrated that when ML residuals exhibit spatial autocorrelation, the accuracy of ML algorithms can be improved by applying kriging to interpolate the residuals and subsequently joining the interpolated residuals into the ML estimates. Therefore, the enhancement of hybridization in our study came from the successful combination of the strength of ML algorithms in predicting deterministic parts and the capability of OK in handling the remaining spatial autocorrelation structure of stochastic errors that may provide challenges when relying solely on either geostatistics or ML algorithms.

The improvement in the accuracy of the hybrid approaches in tropical forest SOCD estimation was still limited because of the poor spatial dependence of residual components extracted from ML, and the enhancement of RFRK was somewhat superior to that of BRTRK. The strength of spatial dependence of the RFRK and BRTRK models was illustrated by the values of N/S, as shown in Table 4. A weak and strong spatial dependence suggests that spatial variability is predominantly driven by external soil factors and soil internal (e.g., vegetation and soil parent material) factors, respectively, while a moderate spatial dependence results from a combination of both these factors (Yao et al. 2019). From the result in Table 4, N/S values of residuals from both ML algorithms, including RF and BRT, were 0.51 and 0.59, respectively, revealing a moderate spatial dependence of residuals controlled by both external and internal soil factors. The moderate spatial dependence of R_{RF} and R_{BRT} is probably attributed to the effectiveness of tree-based ensemble ML models and the competing demands regarding the suitable scheme of sampling points (Brus and Heuvelink 2007). More specifically, there was a considerable reduction in the strength of spatial dependence from strong spatial dependence in the SOCD variable to moderate spatial dependence in residuals after implementing ML models (Table 4). Meanwhile, the conflict regarding the optimal arrangement of sample locations arises from the fact that in hybrid techniques, the observed data was used twice: first, to fit the ML regression models, and second, to perform kriging on the residuals. Our sampling scheme, however, appeared to be more appropriate for ML models, leading to a worse capacity to capture the degree of spatial autocorrelation structure in residuals. Besides, Guo et al. (2015) pointed out that the ML plus residuals kriging approach encountered difficulty interpreting the relationship between the target variable and predictor variables. To address this limitation, Pearson correlation analysis and RFE were integrated with the hybrid models, making it possible to enhance the predictive accuracy.

The maps of forest SOCD's spatial distribution produced by the ML algorithms and the hybrid techniques exhibited little difference despite their distinct methodologies (Fig. 9). However, forest SOCD of hybrid approaches showed a clumped distribution compared to that of ML algorithms. In other words, when integrating interpolated residuals with predicted SOCD of ML, the spatial variability of forest SOCD became smoother, and the occurrence of spatial randomness decreased. Besides, all these models appropriately estimated SOCD on average, yet underestimation occurred with higher observed SOCD values and overestimation appeared with lower observed SOCD values (Fig. 8). Since the incorporation of the kriged residuals can broaden the upper and lower bounds of the predicted values, the ranges of forest SOCD in RFRK and BRTRK exhibited significantly greater compared to those of RF and BRT. The more extensive range partly suggested that the hybrid methods can handle better the underestimation of high SOCD values in high-elevation mountain areas and the overestimation of low SOCD values in low-lying terrain areas compared to corresponding ML models.

5 Conclusions

In this study, three geostatistical models (OK, UK, KED), two ML algorithms (RF and BRT), and two combined models of ML and geostatistics (RFRK and BRTRK) were utilized to model SOCD in tropical forest ecosystems of Central Vietnam on the basis of the correlated predictor covariates, including RS data, topographic attributes, climatic factors, and geographic position variables. Under the similarity in the quantity and quality of geo-referenced data, along with the restricted availability of observational information due to access issues in tropical forest landscapes, all applied approaches yielded promising results in SOCD estimation, as revealed via high R^2 and low error metrics. Nevertheless, the hybrid methods consistently exhibited the greatest accuracy, with the ML algorithms surpassing the geostatistical models. Regarding visualization, the prediction maps produced by the hybridization showed the most realistic SOCD pattern for the forests. Meanwhile, the prediction maps generated by geostatistics were prone to have smoother patterns and those produced by ML tended to possess more detailed patterns. Last but not least, by adding the kriged residuals for capturing spatial autocorrelation structures, the combined models showed superiority in alleviating over- and underestimation of the extreme values over the ML algorithms. Our findings suggest that the use of hybrid models should be taken into consideration for accurately modeling SOCD in tropical forest ecosystems.

Acknowledgements The authors are grateful to our colleagues at Hue University of Agriculture and Forestry and the staff of the Danang Forest Protection Department who joined the field surveys and data collection. Specifically, we give many thanks to Trung Quoc Tran, Quang Trung Nguyen, Phuoc Han Truong, and Trung Tin Nguyen. We acknowledge Prof. Dr. Morihiro Maeda and Prof. Dr. Kazuto Shima for their guidance and for allowing us to use the laboratories at Okayama University, Japan to analyze soil samples.

Author contributions V.H.H.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing—original draft, Visualization. H.M.: Conceptualization, Methodology, Resources, Formal analysis, Writing-review and editing, Supervision. T.H.H.: Methodology, Investigation, Writing-review and editing. F.B.: Methodology, Resources, Writing-review and editing. T.T.N.: Investigation, Writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding provided by Okayama University. The authors did not receive support from any organization for the submitted work.

Data availability Data will be made available on request.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abdoli P, Khanmirzaei A, Hamzeh S et al (2023) Use of remote sensing data to predict soil organic carbon in some agricultural soils of Iran. Remote Sens Appl Soc Environ 30:100969. https://doi.org/10.1016/j.rsase.2023.100969

Akbari M, Goudarzi I, Tahmoures M et al (2021) Predicting soil organic carbon by integrating Landsat 8 OLI, GIS and data mining techniques in semi-arid region. Earth Sci Informatics 14:2113–2122. https://doi.org/10.1007/s12145-021-00673-8

Anderson-Teixeira KJ, Herrmann V, Banbury Morgan R et al (2021) Carbon cycling in mature and regrowth forests globally. Environ Res Lett 16:053009. https://doi.org/10.1088/1748-9326/abed01

Asa E, Saafi M, Membah J, Billa A (2012) Comparison of Linear and Nonlinear Kriging Methods for Characterization and Interpolation of Soil Data. J Comput Civ Eng 26:11–18. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000118

Asuero AG, Sayago A, González AG (2006) The Correlation Coefficient: An Overview. Crit Rev Anal Chem 36:41–59. https://doi.org/10.1080/10408340500526766

Barker MG, Pinard MA (2001) Forest canopy research: sampling problems, and some solutions. In: Linsenmair KE, Davis AJ, Fiala B, Speight MR (eds). Springer Netherlands, Dordrecht, pp 23–38

Beguin J, Fuglstad G-A, Mansuy N, Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306:195–205. https://doi.org/10.1016/j.geoderma.2017.06.016

Bergstrom DW, Monreal CM, St. Jacques E, (2001) Spatial dependence of soil organic carbon mass and its relationship to soil series and topography. Can J Soil Sci 81:53–62. https://doi.org/10.4141/S00-016

Borůvka L, Vašát R, Šrámek V, et al (2022) Predictors for digital mapping of forest soil organic carbon stocks in different types of landscape. Soil Water Res 17:69–79. https://doi.org/10.17221/4/2022-SWR

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Broge N, Leblanc E (2001) Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sens Environ 76:156–172. https://doi.org/10.1016/S0034-4257(00)00197-8

Brus DJ, Heuvelink GBM (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138:86–95. https://doi.org/10.1016/j.geoderma.2006.10.016

Cambardella CA, Moorman TB, Novak JM et al (1994) Field-Scale Variability of Soil Properties in Central Iowa Soils. Soil Sci Soc Am J 58:1501–1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Ceddia M, Gomes A, Vasques G, Pinheiro É (2017) Soil Carbon Stock and Particle Size Fractions in the Central Amazon Predicted from Remotely Sensed Relief. Multispectral and Radar Data Remote Sens 9:124. https://doi.org/10.3390/rs9020124

Chen L, Ren C, Li L et al (2019a) A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content. ISPRS Int J Geo-Information 8:174. https://doi.org/10.3390/ijgi8040174

Chen L, Wang Y, Ren C et al (2019b) Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging. For Ecol Manage 447:12–25. https://doi.org/10.1016/j.foreco.2019.05.057

Cleveland CC, Townsend AR, Taylor P et al (2011) Relationships among net primary productivity, nutrients and climate in tropical rain forest: a pan-tropical analysis. Ecol Lett 14:939–947. https://doi.org/10.1111/j.1461-0248.2011.01658.x

Clough BJ, Green EJ (2013) Comparing spatial and non-spatial approaches for predicting forest soil organic carbon at unsampled locations. Math Comput for Nat Sci 5:115–125

Dai W, Zhao K, Fu W et al (2018) Spatial variation of organic carbon density in topsoils of a typical subtropical forest, southeastern China. CATENA 167:181–189. https://doi.org/10.1016/j.catena.2018.04.040

Dash PK, Panigrahi N, Mishra A (2022) Identifying opportunities to improve digital soil mapping in India: A systematic review. Geoderma Reg 28:e00478. https://doi.org/10.1016/j.geodrs.2021.e00478

Delmelle EM (2021) Spatial Sampling. Handbook of Regional Science. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 1829–1844

Di HJ, Kemp RA, Trangmar BB (1989) Use of Geostatistics in Designing Sampling Strategies for Soil Survey. Soil Sci Soc Am J 53:1163–1167. https://doi.org/10.2136/sssaj1989.03615995005300040028x

Duarte E, Zagal E, Barrera JA et al (2022) Digital mapping of soil organic carbon stocks in the forest lands of Dominican Republic. Eur J Remote Sens 55:213–231. https://doi.org/10.1080/22797254.2022.2045226

Emadi M, Taghizadeh-Mehrjardi R, Cherati A et al (2020) Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. Remote Sens 12:2234. https://doi.org/10.3390/rs12142234

Erdogan Erten G, Yavuz M, Deutsch CV (2022) Combination of Machine Learning and Kriging for Spatial Estimation of Geological Attributes. Nat Resour Res 31:191–213. https://doi.org/10.1007/s11053-021-10003-w

FAO (2015) Knowledge reference for national forest assessments. Italy, Rome

Farooq I, Bangroo SA, Bashir O et al (2022) Comparison of Random Forest and Kriging Models for Soil Organic Carbon Mapping in the Himalayan Region of Kashmir. Land 11:2180. https://doi.org/10.3390/land11122180

Fathololoumi S, Vaezi AR, Alavipanah SK et al (2020) Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: A case study in Iran. Sci Total Environ 721:137703. https://doi.org/10.1016/j.scitotenv.2020.137703

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37:4302–4315. https://doi.org/10.1002/joc.5086

Fouedjio F, Klump J (2019) Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. Environ Earth Sci 78:38. https://doi.org/10.1007/s12665-018-8032-z

Garsia A, Moinet A, Vazquez C et al (2023) The challenge of selecting an appropriate soil organic carbon simulation model: A comprehensive global review and validation assessment. Glob Chang Biol 29:5760–5774. https://doi.org/10.1111/gcb.16896

Gia Pham T, Kappas M, Van Huynh C, Hoang Khanh Nguyen L (2019) Application of Ordinary Kriging and Regression Kriging Method for Soil Properties Mapping in Hilly Region of Central Vietnam. ISPRS Int J Geo-Information 8:147. https://doi.org/10.3390/ijgi8030147

Gitelson AA, Merzlyak MN (1998) Remote sensing of chlorophyll concentration in higher plant leaves. Adv Sp Res 22:689–692. https://doi.org/10.1016/S0273-1177(97)01133-2

Goidts E, Van Wesemael B, Crucifix M (2009) Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. Eur J Soil Sci 60:723–739. https://doi.org/10.1111/j.1365-2389.2009.01157.x

Goovaerts P (1999) Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89:1–45. https://doi.org/10.1016/S0016-7061(98)00078-0

Gu H, Wang J, Ma L et al (2019) Insights into the BRT (Boosted Regression Trees) Method in the Study of the Climate-Growth Relationship of Masson Pine in Subtropical China. Forests 10:228. https://doi.org/10.3390/f10030228

Guo P-T, Li M-F, Luo W et al (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237–238:49–59. https://doi.org/10.1016/j.geoderma.2014.08.009

Ho VH, Morita H, Bachofer F, Ho TH (2024) Random forest regression kriging modeling for soil organic carbon density estimation using multi-source environmental data in central Vietnamese forests. Model Earth Syst Environ. https://doi.org/10.1007/s40808-024-02158-1

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

Huete A, Didan K, Miura T et al (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 83:195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huy B, Poudel KP, Temesgen H (2016) Aboveground biomass equations for evergreen broadleaf forests in South Central Coastal ecoregion of Viet Nam: Selection of eco-regional or pantropical models. For Ecol Manage 376:276–283. https://doi.org/10.1016/j.foreco.2016.06.031

Jenny H (1994) Factors of soil formation: a system of quantitative pedology. Dover Publications

Ji Y, Wang L, Zhang W et al (2024) Forest above-ground biomass estimation using X, C, L, and P band SAR polarimetric observations and different inversion models. Int J Digit Earth 17. https://doi.org/10.1080/17538947.2024.2310730

Kravchenko A, Bullock DG (1999) A Comparative Study of Interpolation Methods for Mapping Soil Properties. Agron J 91:393–400. https://doi.org/10.2134/agronj1999.00021962009100030007x

Lamichhane S, Kumar L, Wilson B (2019) Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352:395–413. https://doi.org/10.1016/j.geoderma.2019.05.031

Leff JW, Wieder WR, Taylor PG et al (2012) Experimental litterfall manipulation drives large and rapid changes in soil carbon cycling in a wet tropical forest. Glob Chang Biol 18:2969–2979. https://doi.org/10.1111/j.1365-2486.2012.02749.x

Li Y, Brando PM, Morton DC et al (2022) Deforestation-induced climate change reduces carbon storage in remaining tropical forests. Nat Commun 13:1964. https://doi.org/10.1038/s41467-022-29601-0

Li R, Tan S, Zhang M et al (2024a) Geological Disaster Susceptibility Evaluation Using a Random Forest Empowerment Information Quantity Model. Sustainability 16:765. https://doi.org/10.3390/su16020765

Li T, Cui L, Kuhnert M et al (2024b) A comprehensive review of soil organic carbon estimates: Integrating remote sensing and machine learning technologies. J Soils Sediments 24:3556–3571. https://doi.org/10.1007/s11368-024-03913-8

Liu Y, Guo L, Jiang Q et al (2015) Comparing geospatial techniques to predict SOC stocks. Soil Tillage Res 148:46–58. https://doi.org/10.1016/j.still.2014.12.002

Liu X, Kounadi O, Zurita-Milla R (2022) Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. ISPRS Int J Geo-Information 11:242. https://doi.org/10.3390/ijgi11040242

Liyanage TDP, Maeda M, Somura H et al (2022) Nitrous oxide and carbon dioxide emissions from two types of soil amended with manure compost at different ammonium nitrogen rates. Soil Sci Plant Nutr 68:473–490. https://doi.org/10.1080/00380768.2022.2087198

Luo K, Wei Y, Du J et al (2022) Machine learning-based estimates of aboveground biomass of subalpine forests using Landsat 8 OLI and Sentinel-2B images in the Jiuzhaigou National Nature Reserve, Eastern Tibet Plateau. J for Res 33:1329–1340. https://doi.org/10.1007/s11676-021-01421-w

Mackey B, Kormos CF, Keith H et al (2020) Understanding the importance of primary tropical forest protection as a mitigation strategy. Mitig Adapt Strateg Glob Chang 25:763–787. https://doi.org/10.1007/s11027-019-09891-4

Marchant BP (2018) Model-Based Soil Geostatistics. pp 341–371

Meliho M, Boulmane M, Khattabi A et al (2023) Spatial Prediction of Soil Organic Carbon Stock in the Moroccan High Atlas Using

Machine Learning. Remote Sens 15:2494. https://doi.org/10.3390/rs15102494

Minasny B, McBratney AB (2016) Digital soil mapping: A brief history and some lessons. Geoderma 264:301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Mishra U, Lal R, Slater B et al (2009) Predicting Soil Organic Carbon Stock Using Profile Depth Distribution Functions and Ordinary Kriging. Soil Sci Soc Am J 73:614–621. https://doi.org/10.2136/sssaj2007.0410

Mutanga O, Masenyama A, Sibanda M (2023) Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects. ISPRS J Photogramm Remote Sens 198:297–309. https://doi.org/10.1016/j.isprsjprs.2023.03.010

National Institute of Agricultural Planning and Projection of Vietnam (2005) Soil map of Danang city. National institute of agricultural planning and projection of Vietnam, Hanoi, Vietnam

Nussbaum M, Papritz A, Baltensweiler A, Walthert L (2014) Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. Geosci Model Dev 7:1197–1210. https://doi.org/10.5194/gmd-7-1197-2014

Odebiri O, Mutanga O, Odindi J et al (2020) Predicting soil organic carbon stocks under commercial forest plantations in KwaZulu-Natal province, South Africa using remotely sensed data. Giscience Remote Sens 57:450–463. https://doi.org/10.1080/15481603.2020.1731108

Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. Int J Geogr Inf Syst 4:313–332. https://doi.org/10.1080/02693799008941549

Oliver MA, Webster R (2014) A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 113:56–69. https://doi.org/10.1016/j.catena.2013.09.006

Padarian J, Stockmann U, Minasny B, McBratney AB (2022) Monitoring changes in global soil organic carbon stocks from space. Remote Sens Environ 281:113260. https://doi.org/10.1016/j.rse.2022.113260

Pan Y, Birdsey RA, Phillips OL et al (2024) The enduring world forest carbon sink. Nature 631:563–569. https://doi.org/10.1038/s41586-024-07602-x

Pearson TRH, Brown SL, Birdsey RA (2007) Measurement guidelines for the sequestration of forest carbon. U.S. Department of Agriculture, Forest Service, Northern Research Station.https://doi.org/10.2737/NRS-GTR-18

Pillay R, Venter M, Aragon-Osejo J et al (2022) Tropical forests are home to over half of the world's vertebrate species. Front Ecol Environ 20:10–15. https://doi.org/10.1002/fee.2420

Qin Q, Wagai R, Aoyagi R et al (2024) Destructive selective logging in tropical forests causes soil carbon loss through forest degradation and soil redox change. For Ecol Manage 551:121555. https://doi.org/10.1016/j.foreco.2023.121555

Scolforo HF, Scolforo JRS, de Mello JM et al (2016) Spatial interpolators for improving the mapping of carbon stock of the arboreal vegetation in Brazilian biomes of Atlantic forest and Savanna. For Ecol Manage 376:24–35. https://doi.org/10.1016/j.foreco.2016.05.047

Shafizadeh-Moghadam H, Minaei F, Talebi-khiyavi H et al (2022) Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. CATENA 212:106077. https://doi.org/10.1016/j.catena.2022.106077

Sharma S, Jain PK, Soloman PE (2023) Carbon Storage Potential of Soil in Diverse Terrestrial Ecosystems. Nat Environ Pollut Technol 22:1809–1819. https://doi.org/10.46488/NEPT.2023.v22i04.009

Shi J, Yang L, Zhu A-X et al (2018) Machine-Learning Variables at Different Scales vs. Knowledge-based Variables for Mapping Multiple Soil Properties. Soil Sci Soc Am J 82:645–656. https://doi.org/10.2136/sssaj2017.11.0392

Shimada M (2011) Model-Based Polarimetric SAR Calibration Method Using Forest and Surface-Scattering Targets. IEEE Trans Geosci Remote Sens 49:1712–1733. https://doi.org/10.1109/TGRS.2010.2090046

Shimada M, Isoguchi O, Tadono T, Isono K (2009) PALSAR Radiometric and Geometric Calibration. IEEE Trans Geosci Remote Sens 47:3915–3932. https://doi.org/10.1109/TGRS.2009.2023909

Md. Shoaibur Rahman, Raihan A, Samanta Islam, et al (2023) Enhancing Soil Carbon Sequestration and Land Restoration through Tropical Forest Management. J Agric Sustain Environ 2:70–85. https://doi.org/10.56556/jase.v2i2.906

Silatsa FBT, Yemefack M, Tabi FO et al (2020) Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon. Geoderma 367:114260. https://doi.org/10.1016/j.geoderma.2020.114260

Song Y-Q, Yang L-A, Li B et al (2017) Spatial Prediction of Soil Organic Matter Using a Hybrid Geostatistical Model of an Extreme Learning Machine and Ordinary Kriging. Sustainability 9:754. https://doi.org/10.3390/su9050754

Suleymanov A, Tuktarova I, Belan L et al (2023) Spatial prediction of soil properties using random forest, k-nearest neighbors and cubist approaches in the foothills of the Ural Mountains, Russia. Model Earth Syst Environ 9:3461–3471. https://doi.org/10.1007/s40808-023-01723-4

Suleymanov A, Abakumov E, Nizamutdinov T et al (2024) Soil organic carbon stock retrieval from Sentinel-2A using a hybrid approach. Environ Monit Assess 196. https://doi.org/10.1007/s10661-023-12172-y

Sun X-L, Yang Q, Wang H-L, Wu Y-J (2019) Can regression determination, nugget-to-sill ratio and sampling spacing determine relative performance of regression kriging over ordinary kriging? CATENA 181:104092. https://doi.org/10.1016/j.catena.2019.104092

Tebeje Y (2020) A Review Paper on the Role of Terrestrial Carbon Stocks for Climate Change Mitigation Mechanisms. J Environ Earth Sci 10:32–50. https://doi.org/10.7176/JEES/10-8-04

Tucker CJ (1979) Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ 8:127–150. https://doi.org/10.1016/0034-4257(79)90013-0

Tziachris P, Aschonitis V, Chatzistathis T, Papadopoulou M (2019) Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. CATENA 174:206–216. https://doi.org/10.1016/j.catena.2018.11.010

van Groenigen JW, Siderius W, Stein A (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87:239–259. https://doi.org/10.1016/S0016-7061(98)00056-1

Veronesi F, Schillaci C (2019) Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol Indic 101:1032–1044. https://doi.org/10.1016/j.ecolind.2019.02.026

Vizzari M (2022) PlanetScope, Sentinel-2, and Sentinel-1 Data Integration for Object-Based Land Cover Classification in Google Earth Engine. Remote Sens 14:2628. https://doi.org/10.3390/rs14112628

Wadoux AMJC, Marchant BP, Lark RM (2019) Efficient sampling for geostatistical surveys. Eur J Soil Sci 70:975–989. https://doi.org/10.1111/ejss.12797

Wang B, Waters C, Orgill S et al (2018) High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Sci Total Environ 630:367–378. https://doi.org/10.1016/j.scitotenv.2018.02.204

Wang S, Gao J, Zhuang Q et al (2020a) Multispectral Remote Sensing Data Are Effective and Robust in Mapping Regional Forest Soil Organic Carbon Stocks in a Northeast Forest Region in China. Remote Sens 12:393. https://doi.org/10.3390/rs12030393

Wang S, Zhuang Q, Jin X et al (2020b) Predicting Soil Organic Carbon and Soil Nitrogen Stocks in Topsoil of Forest Ecosystems in Northeastern China Using Remote Sensing Data. Remote Sens 12:1115. https://doi.org/10.3390/rs12071115

Wang H, Wang J, Zhang Y et al (2023) Spatial distribution of soil organic carbon and its response to forest growth and soil layer in Cunninghamia lanceolata plantations in mid-subtropical China. For Ecol Manage 545:121302. https://doi.org/10.1016/j.foreco.2023.121302

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. John Wiley & Sons

Wiesmeier M, Prietzel J, Barthold F et al (2013) Storage and drivers of organic carbon in forest soils of southeast Germany (Bavaria) – Implications for carbon sequestration. For Ecol Manage 295:162–172. https://doi.org/10.1016/j.foreco.2013.01.025

Xia Y, McSweeney K, Wander MM (2022) Digital Mapping of Agricultural Soil Organic Carbon Using Soil Forming Factors: A Review of Current Efforts at the Regional and National Scales. Front Soil Sci 2:1–19. https://doi.org/10.3389/fsoil.2022.890437

Xu Y, Li B, Bai J et al (2022) Effects of multi-temporal environmental variables on SOC spatial prediction models in coastal wetlands of a Chinese delta. L Degrad Dev 33:3557–3567. https://doi.org/10.1002/ldr.4408

Yang R-M, Zhang G-L, Liu F et al (2016) Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol Indic 60:870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

Yao X, Yu K, Deng Y et al (2019) Spatial distribution of soil organic carbon stocks in Masson pine (Pinus massoniana) forests in subtropical China. CATENA 178:189–198. https://doi.org/10.1016/j.catena.2019.03.004

Zhi J, Zhou Z, Cao X (2021) Exploring the determinants and distribution patterns of soil mattic horizon thickness in a typical alpine environment using boosted regression trees. Ecol Indic 133:108373. https://doi.org/10.1016/j.ecolind.2021.108373

Zhou T, Geng Y, Chen J et al (2020a) Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. Ecol Indic 114:106288. https://doi.org/10.1016/j.ecolind.2020.106288

Zhou T, Geng Y, Chen J et al (2020b) High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Sci Total Environ 729:138244. https://doi.org/10.1016/j.scitotenv.2020.138244

Zhou T, Luo X, Hou Y et al (2020c) Quantifying the effects of road width on roadside vegetation and soil conditions in forests. Landsc Ecol 35:69–81. https://doi.org/10.1007/s10980-019-00930-8

Zhou Y, Zhao X, Guo X, Li Y (2022) Mapping of soil organic carbon using machine learning models: Combination of optical and radar remote sensing data. Soil Sci Soc Am J 86:293–310. https://doi.org/10.1002/saj2.20371

Zhu Q, Lin HS (2010) Comparing Ordinary Kriging and Regression Kriging for Soil Properties in Contrasting Landscapes. Pedosphere 20:594–606. https://doi.org/10.1016/S1002-0160(10)60049-5

Zhu C, Wei Y, Zhu F et al (2022) Digital Mapping of Soil Organic Carbon Based on Machine Learning and Regression Kriging. Sensors 22:8997. https://doi.org/10.3390/s22228997

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.