https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Unveiling China's Forest Soil properties: High-Resolution, Multi-Depth Mapping of Soil Bulk Density and pH Using Machine Learning Methods

Jizhen Chen[1,2] Xin Zhang[1,2], Zihao Fan[1,2], Tao Liu[3], Wenfa Xiao[1,2], Qiwu Sun[4], Xiangyang Sun[5], Zilin Huang[1,2]

[1] Key Laboratory of Forest Ecology and Environment of National Forestry and Grassland Administration, Ecology and Nature Conservation Institute, Chinese Academy of Forestry, Beijing 100091, China
[2] Hubei Zigui Three Gorges Reservoir Forest Ecosystem Observation and Research Station, Zigui 443600, China
[3] Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing 100084, China
[4] Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China
[5] College of Forestry, Beijing Forestry University, Beijing 100083, China

Correspondence to: Zilin Huang (hzlin66@caf.ac.cn)

Abstract. Precise monitoring of key forest soil properties is crucial for addressing global challenges like carbon sequestration and soil acidification. However, existing national soil maps, primarily derived from comprehensive ecosystem samples, inadequately represent the distinct characteristics and high spatial heterogeneity of China's vast and diverse forest ecosystems. To bridge this gap, we present the first high-resolution (90-m), forest-specific maps of soil bulk density (BD) and pH across China. Leveraging 4,356 forest soil profiles collected through extensive field surveys and 41 environmental covariates within an optimized Quantile Regression Forests (QRF) framework incorporating forward recursive feature selection (FRFS), we generated wall-to-wall predictions for five standardized depth intervals (0–5, 5–15, 15–30, 30–60, 60–100 cm). Model performance, assessed through 10-fold cross-validation (CV) and independent validation (IV), achieved model efficiency coefficients (MEC) ranging from 0.78 to 0.89 (CV) and 0.60 to 0.66 (IV) for bulk density (BD), and from 0.83 to 0.87 (CV) and 0.71 to 0.81 (IV) for pH, indicating the product's strong capability to capture the spatial variability of forest soil properties across China. The 90-m resolution BD and pH maps contribute to the GlobalSoilMap initiative and provide forest-specific inputs for regional Earth system and land surface models. These products advance the quantification of soil acidification processes and provide critical baseline data for estimating forest soil carbon stocks across China. The dataset is available at https://doi.org/10.57760/sciencedb.25375.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

1 Introduction

Forest soils are defined as soils that have developed under forest cover, influenced by long-term vegetation–soil interactions, and distinguished by unique physical, chemical, and biological properties (Binkley and Fisher, 2013; Osman, 2013). As key regulators of carbon storage, water cycling, and nutrient availability, forest soils are vital to forest sustainability and policy (Dai et al., 2019; Kleber et al., 2021). China's forest ecosystems span 209 million hectares across diverse climatic zones and complex topographies, encompassing 452 vegetation types to form one of Earth's most ecologically varied forest spectra (Chen et al., 2016; Patton et al., 2019; Zhang et al., 2024). Revealing the spatial distribution of forest soils is fundamental for estimating forest carbon stocks and evaluating forest soil acidification (Zhu et al., 2016; Huang et al., 2022b; Xu et al., 2015). However, forest soils are highly heterogeneous across geographical space, shaped by long-term climatic gradients, vegetation succession, and topographic variation (Zhao et al., 2019; Chen et al., 2022a; Liu et al., 2024). Consequently, accurately revealing the spatial distribution of key forest soil attributes presents a significant challenge.

Digital Soil Mapping (DSM), which integrates machine learning and environmental covariates to predict soil properties across complex landscapes while significantly enhancing spatial soil representation in areas of varied terrain and vegetation, has become a pivotal methodology for acquiring high-resolution spatial soil information (McBratney et al., 2003; Minasny et al., 2013; Padarian et al., 2019). Consequently, numerous countries globally and transnational initiatives have invested substantial resources in using DSM to build national-scale, high-accuracy digital soil databases. These national initiatives typically target resolutions of 90 meters or finer, predicting the spatial distribution of multiple soil attributes across globally standardized depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm) as established by GlobalSoilMap.net (Arrouays et al., 2014; Hempel et al., 2014). Exemplary national efforts include the Soil and Landscape Grid of Australia (SLGA; Grundy et al., 2015), France's Soil Data Inventory and Management System (DIGSOL; Mulder et al., 2016), the gSSURGO database in the United States (Ramcharan et al., 2018; Thompson et al., 2020), and the high resolution National Soil Information Grids of China (Liu et al., 2022a; Shi et al., 2025). Concurrently, global-scale initiatives such as SoilGrids provide open-access soil predictions at 250m resolution across all continents using the same GlobalSoilMap standards (Hengl et al., 2017; Poggio et al., 2021). Collectively, these national efforts have substantially advanced our understanding of spatial distribution of multiple soil attributes within their respective coverage areas.

However, a significant limitation persists in characterizing forest soils specifically. Current national-scale soil products in China and globally primarily derive from samples located in comprehensive ecosystem (Poggio et al., 2021; Liu et al., 2022b; Shi et al., 2025). Consequently, they fail to adequately capture the unique physical structures (e.g., higher aggregation, root effects) and biogeochemical processes (e.g., greater susceptibility to acidification driven by vegetation inputs) inherent to forest ecosystems (Widyati et al., 2022; Liu et al., 2024). This creates a critical gap between available soil data products and the urgent need for forest-specific soil information to support accurate carbon stock estimation and acidification risk assessment in these vital ecosystems. Further complicating this gap is the methodological challenge of selecting predictive covariates for forest soil mapping that adequately capture the complex vegetation-soil interactions (Wu et al., 2023).

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

While the SCORPAN (Soil, Climate, Organisms, Relief, Parent material, Age, and Space) framework underpins digital soil mapping (McBratney et al., 2003; Chen et al., 2022b), optimal covariate selection for forest soils remains challenging due to complex vegetation-soil interactions (Chen et al., 2021; Xue et al., 2025). Model-based feature importance methods, particularly random forest (RF) metrics, have become primary solutions to address dimensionality traps from extensive predictors (Song et al., 2020; Liu et al., 2022a). Subsequent studies have successfully leveraged RF importance to identify key drivers of soil attributes, such as soil-environment relationships via OOB error (Jeune et al., 2018) and critical covariates for soil hydraulic properties (Santos et al., 2023). However, RF-based approaches frequently fail to identify minimal optimal subsets due to variable redundancy. To overcome this, Recursive Feature Elimination (RFE) was developed, which iteratively prunes low-importance features using RF. It distilled high-dimensional covariate sets into parsimonious subsets for soil organic carbon stocks (Hounkpatin et al., 2021) and identified key topographic-vegetation predictors for soil nutrients in heterogeneous (Helfenstein et al., 2024; Shi et al., 2025). Yet RFE's sequential removal risks discarding combinatorially significant variables and incurs high computational costs. More recently, the forward recursive feature selection (FRFS) method has emerged as a superior alternative, excelling at capturing nonlinear relationships while reducing computational costs (Xiao et al., 2022). Xue et al. (2025) successfully applied this method to map the spatial heterogeneity of complex soil attributes across diverse landscapes in China, demonstrating its promising potential for addressing the specific challenges of mapping heterogeneous forest soils. Therefore, our study leverages the FRFS method to tackle the critical covariate selection challenge inherent to mapping China's diverse forest soils.

To address the critical limitations of legacy soil data in representing China's complex and heterogeneous forest ecosystems, we conducted a systematic nationwide forest soil survey. Leveraging machine learning, this study aims to: (1) construct the first nationwide forest-specific soil profile database; (2) develop and apply an optimized DSM framework integrating QRF and FRFS; and (3) pioneer high-resolution (90-meter) digital maps of two fundamental forest soil properties, bulk density (BD) and pH, across China's entire forest domain. Spanning five standardized depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, and 60–100 cm), these forest-specific maps provide the first continuous, wall-to-wall spatial characterization of BD and pH at 90m resolution, conforming to GlobalSoilMap standards. This unprecedented dataset provides the essential spatial baseline for accurately quantifying forest carbon stocks and assessing soil acidification risks.

2 Materials and Methods

We developed 90-m resolution forest soil BD and pH grids for China (0–100 cm) using an optimized QRF model, a machine learning algorithm effective for both spatial prediction and uncertainty quantification (Szatmári et al., 2024). This framework integrated 4,356 georeferenced forest soil profiles, combining historical inventory data (2018–2023). Sampling efforts were designed to ensure ecological and spatial representativeness across major climatic zones and forest types. Soil profiles were harmonized into standardized depth intervals (0–5, 5–15, 15–30, 30–60, and 60–100 cm) using an adaptive equal-area spline method (Bishop et al., 1999; Liu et al., 2022a) and randomly partitioned into training (80%) and independent validation (20%)

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

subsets. A set of 41 environmental covariates, aligned with soil-forming factors (Jenny, 1941), were resampled to a 90-m grid via bilinear interpolation. Feature selection and hyperparameter tuning were implemented to optimize model performance. Predictive accuracy was evaluated using 10-fold cross-validation and independent validation based on a withheld dataset. A summary of the modelling framework is shown in Figure 1.

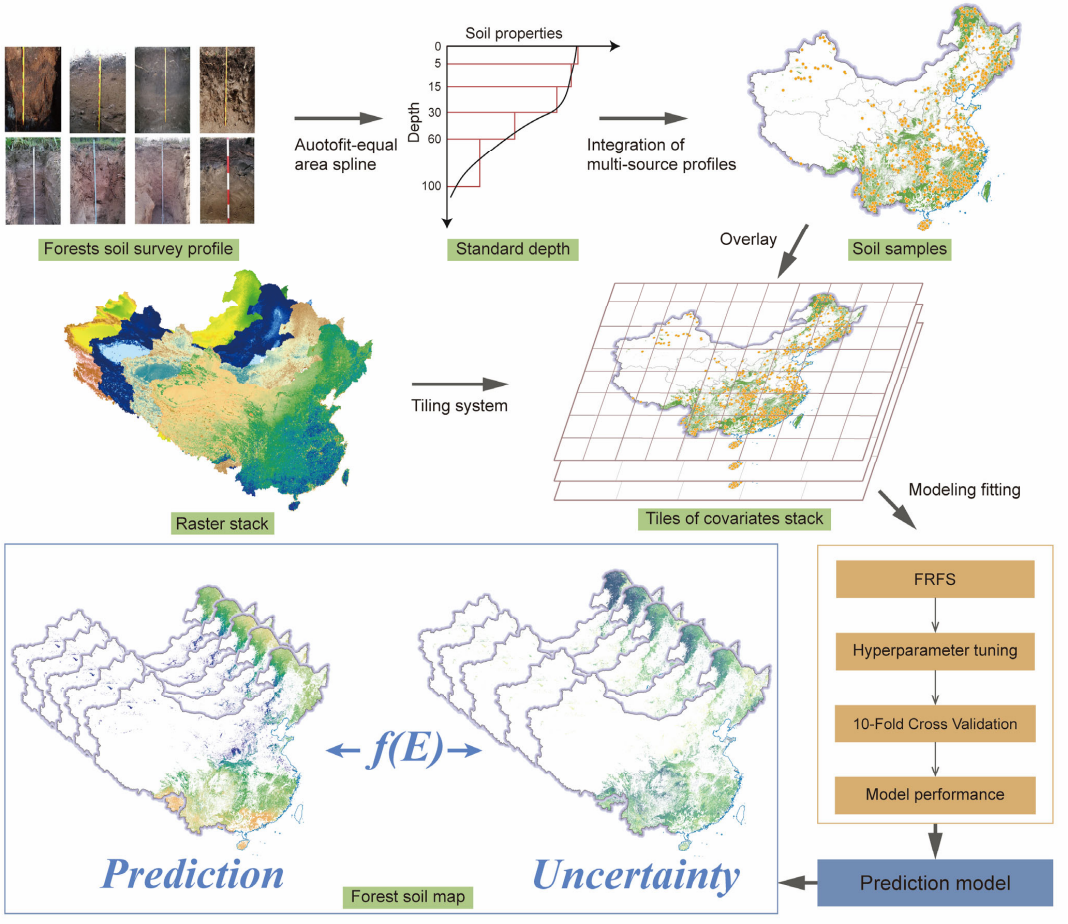


Figure 1. Workflow diagram for forest soil mapping.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

2.1 Data compilation

2.1.1 Soil database

We developed a comprehensive forest soil property database for China, representing the most extensive and up-to-date collection of forest soil data to date. The data were compiled from two major nationwide forest soil surveys conducted in 2018 and 2023, complemented by independently conducted regional forest soil surveys during the intervening years to enhance spatial and ecological representation. These surveys employed a stratified sampling design to ensure broad representativeness across China's major forest ecosystems, covering diverse climate zones, forest types, and topographic gradients. In addition to these national efforts, data from independently conducted regional forest soil surveys during the intervening years were also incorporated to enhance spatial and ecological representation. After rigorous quality control and data harmonization, the final integrated dataset comprises 8,709 soil profiles and 18,193 soil samples. Of these, 4,356 profiles and 11,873 samples contain both BD and pH values, forming the core dataset used in this study. The spatial distribution of sampling plots and forest coverage is displayed in Figure 2.

To ensure data comparability and minimize measurement errors, all samples were processed under identical conditions. Soil sampling and analysis followed standardized protocols to ensure data consistency. Soil samples were collected using a soil auger, air-dried at room temperature, homogenized, and passed through a 2 mm sieve to remove coarse fragments and roots for physicochemical analyses. Undisturbed soil cores were collected from each horizon using a cutting ring sampler to determine BD. Soil pH was measured using a pH meter following the potentiometric method, with a soil-to-water ratio of 1:2.5 (w/v). Reference materials were used throughout the analytical process to ensure measurement accuracy and control data quality.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

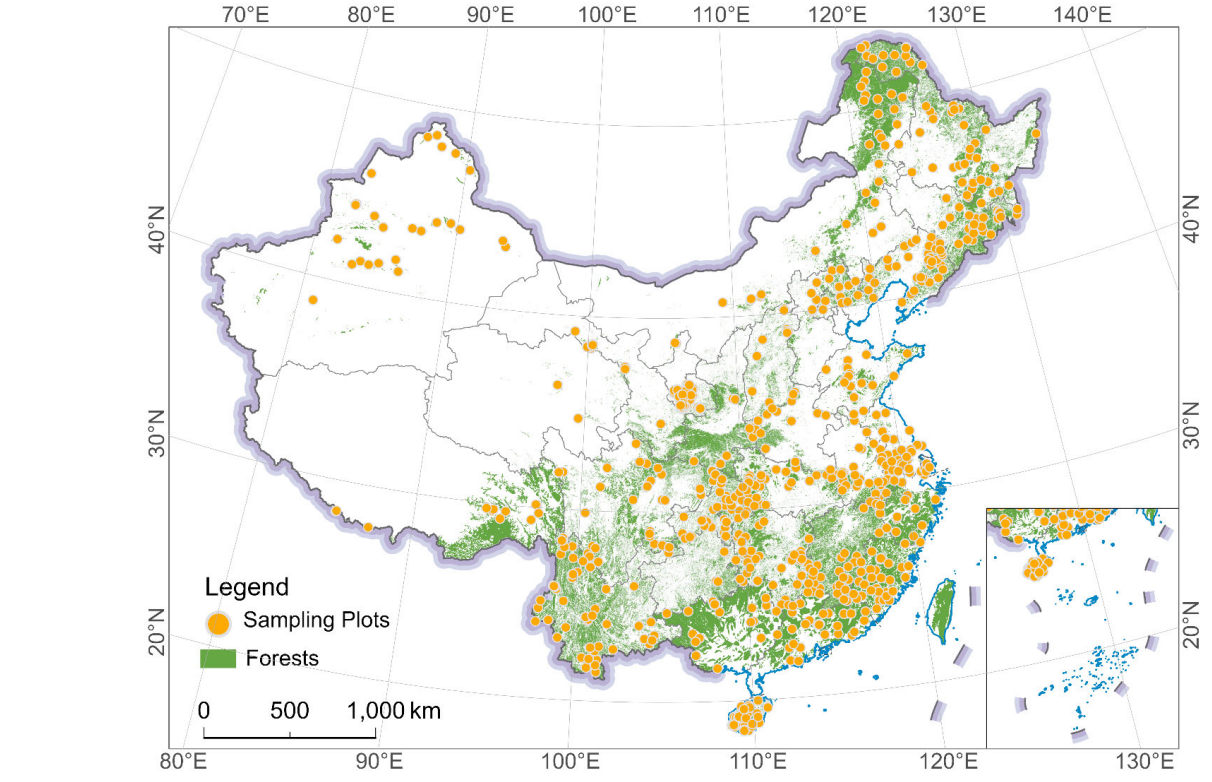


Figure 2. Spatial distribution of soil sampling plots and forest coverage. Publisher's remark: please note that the above figure contains disputed territories.

125

2.1.2 Standard soil depths

Following GlobalSoilMap specifications (Arrouays et al., 2014), soil samples are typically standardized to fixed depth intervals of 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. To model continuous depth functions from soil property measurements recorded by genetic horizons, equal-area quadratic spline interpolation is commonly used (Bishop et al., 1999). However, 130 natural soil profiles often contain abrupt changes in properties between adjacent horizons, leading to inconsistencies with these standardized depth layers. To address this issue and reduce fitting errors, we applied an adaptive equal-area spline method (Liu et al., 2022a). This method detects abrupt transitions by calculating the ratio of property values between adjacent horizons and applying a predefined threshold. When such discontinuities are identified, a 1 cm transitional layer is inserted between the affected horizons before spline fitting. This adjustment ensures improved consistency with the observed morphological

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

structure of each soil profile. While the GlobalSoilMap framework includes the 100–200 cm interval, our study focused on the upper five layers (0–5, 5–15, 15–30, 30–60, and 60–100 cm) due to the limited number of forest soil profiles extending beyond 100 cm in depth.

2.1.3 Environmental Covariates

Soil formation is governed by the combined effects of climate, parent material, topography, vegetation, and human activities. In this study, 41 environmental covariates were selected based on the soil-forming factor framework (Jenny, 1941; Minasny et al., 2013) and categorized into five groups: parent material, climate, organisms, topography, and intrinsic soil properties (Table S1). To reduce multicollinearity, a variance inflation factor (VIF) threshold of less than 10 was applied through iterative variable exclusion.

All covariate layers were projected using the Albers Equal Area coordinate system (EPSG:4326, WGS84 datum) and resampled to a unified 90-m spatial resolution via bilinear interpolation. For multi-year variables, long-term annual means and growing season (May to September) averages were calculated from monthly records spanning 2003 to 2023, thereby capturing both historical trends and contemporary environmental conditions relevant to forest soil development.

Climate-related covariates included temperature, precipitation, potential evapotranspiration, and solar radiation, derived from the National Tibetan Plateau Data Center (https://data.tpdc.ac.cn) and the TerraClimate dataset. Parent material characteristics were obtained from Sentinel-2 imagery using the shortwave infrared band (B12) and the B8/B12 band ratio to estimate clay content. Depth to Bedrock (DTB) data were incorporated to represent weathering intensity, and lithological context was supplemented using the Geological Map of China. Topographic attributes were extracted from the NASADEM digital elevation model (https://lpdaac.usgs.gov/products/nasadem_hgtv001/) and computed using SAGA GIS (http://www.saga-gis.org). Vegetation indicators were sourced from MODIS products, including NDVI, NDWI, LAI, and NPP, while forest type classifications were based on the National Atlas of Forest Vegetation in China.

2.2 Modelling

2.2.1 Covariate selection

To balance model parsimony with biogeochemical interpretability, we adapted the Forward Recursive Feature Selection (FRFS) approach proposed by Xiao et al. (2022) into a depth-specific selection framework, applied independently to four standardized soil layers. The procedure comprised three sequential steps. First, the covariate with the highest predictive importance, as assessed by permutation-based Random Forest analysis, was selected to initiate the model. Subsequently, additional variables were iteratively added based on two criteria: a reduction of more than 5% in five-fold cross-validated root mean square error (RMSE) and a variance inflation factor (VIF) below 10. The selection process was automatically terminated when five consecutive iterations failed to achieve an RMSE improvement of at least 1%, thereby avoiding model overfitting. This

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

hierarchical strategy ensured effective dimensionality reduction while maintaining predictive performance across all soil depths. The framework was applied across four distinct soil horizons.

2.2.2 Predictive models

Quantile Regression Forests (QRF), a nonparametric ensemble learning method extending the Random Forest framework, were used to model the relationships between environmental covariates and soil properties, while explicitly quantifying predictive uncertainty (Meinshausen, 2006). As a state-of-the-art algorithm in DSM (Liu et al., 2022a; Poggio et al., 2021; Pouladi et al., 2019), QRF leverages both bootstrap aggregation of regression trees and randomized feature subset selection at each node, enabling robust handling of high-dimensional, non-stationary data.

Unlike standard Random Forests, QRF retains the full conditional distribution $F(y \parallel X = x)$ At each prediction node, allowing estimation of both point predictions and confidence intervals. This is achieved via kernel-based empirical distribution construction:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) I(y_i \leq y) \quad (1)$$

where $w_i(x)$ Is the weight assigned to each training observation based on terminal node proximity. The conditional quantile function is derived as:

$$\hat{q}_\alpha(x) = \inf\{y: \hat{F}(y \mid X = x) \geq \alpha\} \quad (2)$$

for a given quantile level $\alpha \in (0,1)$. This allows the derivation of the median estimate $\hat{q}_{0.5}(x)$, prediction intervals $[\hat{q}_{a/2}(x), \hat{q}_{1-a/2}(x)]$, and the full uncertainty distribution, enhancing both interpretability and decision support in forest soil assessments.

where specifies the quantile level (e.g., $\alpha = 0.5$ or median prediction). This formulation generates three interconnected outputs: the median predictor as a robust central tendency estimate, prediction intervals for heteroscedastic uncertainty quantification, and the complete conditional distribution through parametric evaluation of $\hat{q}_\alpha(x)$ across the α continuum.

To implement QRF across China's forested regions, we adopted a spatially parallel computing framework. The study area was divided into 461 contiguous grid tiles (10 × 10 km) using the Albers Equal Area projection. Model execution was carried out using the quantregForest package in R 4.2.1, running on 24 logical cores of a high-performance computing node. Spatial continuity was preserved across grid boundaries using a Gaussian kernel-based edge matching algorithm, enabling seamless 90-m resolution prediction without artifacts.

2.2.3 Hyperparameter tuning

Hyperparameter optimization was conducted for three parameters critical to model performance: mtry (number of variables randomly sampled at each split), num.trees (number of trees), and min.node.size (minimum samples per terminal node). The randomized search strategy was employed, guided by 10-fold cross-validation and using RMSE as the evaluation metric. Fifty

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

iterations of parameter space sampling were performed to identify the optimal combination. Final hyperparameter values were selected based on configurations that yielded the highest prediction accuracy on the validation dataset. A summary of optimized parameters for each soil property and depth interval is provided in Table S2.

2.3 Model validation

To comprehensively evaluate model performance, we applied two complementary validation strategies: 10-fold cross-validation on the training dataset (80%) and independent validation using a held-out test set (20%). These schemes were implemented across the entire study region to assess the predictive accuracy of forest soil BD and pH.

In 10-fold cross-validation, the training set was randomly partitioned into ten equal subsets. In each iteration, nine subsets were used to train the model, and the remaining one was used for validation. This procedure was repeated ten times, ensuring each subset served as validation data exactly once. Model accuracy was assessed by averaging performance metrics across folds, including mean error (ME), root mean square error (RMSE), and the model efficiency coefficient (MEC).

For independent validation, the reserved test set was excluded entirely from model training and hyperparameter tuning, thereby providing an unbiased evaluation of generalizability. The formulas used for calculating the evaluation metrics are as follows:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$MEC = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the observed soil property value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of observed values. ME, also referred to as bias, measures average deviation. RMSE reflects the overall prediction error, with lower values indicating higher accuracy. MEC, equivalent to the coefficient of determination (R^2), ranges from 0 to 1, with higher values indicating better predictive performance.

2.4 Uncertainty Quantification

Quantifying spatial uncertainty is essential in DSM, as prediction errors may arise from input data variability, model structure, and environmental heterogeneity (Arrouays et al., 2014; Poggio et al., 2021; Liu et al., 2022a; Shi et al., 2025). To visualize the spatial distribution of prediction uncertainty, we calculated the Prediction Interval Ratio (PIR), defined as the ratio between the 90% prediction interval width and the median estimate:

$$PIR = \frac{q_{0.95} - q_{0.05}}{q_{0.50}} \tag{6}$$

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

where $q_{0.95}$ and $q_{0.05}$ represent the upper and lower bounds of the 90% prediction interval, respectively, and $q_{0.05}$ denotes the median prediction. PIR is a dimensionless metric that quantifies the relative spread of prediction uncertainty around the central estimate. Higher PIR values indicate greater dispersion and, therefore, higher predictive uncertainty.

To evaluate the calibration of these uncertainty estimates, we used the Prediction Interval Coverage Probability (PICP), computed from the independent validation dataset (Goovaerts, 2001). PICP measures the proportion of observed values that fall within their corresponding prediction intervals at a specified confidence level (e.g., 90%). A well-calibrated model should yield a PICP value close to the nominal coverage. For example, a 90% prediction interval is considered reliable if the empirical PICP also approximates 90%. Systematic deviation from this benchmark can indicate miscalibration: a PICP above the target level suggests that intervals are too narrow (underestimated uncertainty), while a PICP below the target indicates overly wide intervals (overestimated uncertainty) (Poggio et al., 2021; Liang et al., 2019). This diagnostic approach supports the robust interpretation of uncertainty in DSM outputs.

3 Results

3.1 Forest soil database overview

Table 1 presents the harmonized forest soil database comprised 4,356 forest soil profiles distributed across China. Using the equal-area spline method, soil property values were standardized to fixed depth intervals (0–5, 5–15, 15–30, 30–60, and 60–100 cm), resulting in 15,845 horizons for BD and 15,978 horizons for pH. BD showed low skewness across depths (0.16–0.42), while pH closely followed a normal distribution (skewness 0.05–0.19). Mean values of both BD and pH increased gradually with depth, from 1.206 to 1.342 g/cm³ for BD and from 6.07 to 6.47 for pH. The standard deviation of BD increased gradually from 0.261 in the shallowest layer (0–5 cm) to 0.308 in the lowest depth interval (60–100 cm), while pH showed a more pronounced rise in variability, with its standard deviation increasing from 0.909 to 1.327 across the same range. Both parameters showed wide value ranges across all depths (BD: 0.15–2.30 g/cm³; pH: 4.00–8.70).

Table 1. Statistical summary of BD and pH at five depth intervals. Refer to Table 1 for the abbreviations and units of the soil properties interested.

Property	Depth (cm)	N[a)]	Mean	SD	Min	Max	Skewness
BD	0–5	4356	1.206	0.261	0.152	2.057	0.162
	5–15	3522	1.209	0.288	0.284	2.291	0.317
	15–30	3488	1.287	0.269	0.301	2.271	0.422
	30–60	2973	1.340	0.269	0.257	2.215	0.393
	60–100	1506	1.342	0.308	0.534	2.291	0.367
pH	0–5	3963	6.066	0.909	4.000	8.440	0.045
	5–15	3962	6.131	0.991	4.000	8.515	0.111

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

	15–30	3816	6.172	1.005	4.030	8.420	0.087
	30–60	2783	6.458	1.198	4.040	8.640	0.194
	60–100	1454	6.466	1.327	4.040	8.704	0.119

[a)] N varies slightly between properties and depths due to sample availability for specific analyses.

3.2 Model performance

The performance of the QRF models was evaluated after training and optimisation. Table 2 lists the 10-fold cross-validation (CV) and independent validation (IV) results for BD and pH predictions of our study across five soil depth intervals. Model performance varied with specific soil properties. For BD, 10-fold CV achieved high accuracy (MEC = 0.782–0.889, RMSE = 0.079–0.090 g/cm³), explaining 78.2–88.9% of variability. IV yielded robust but reduced performance (MEC = 0.598–0.657, RMSE = 0.155–0.181 g/cm³), retaining 59.8–65.7% explanatory power. Conversely, pH predictions demonstrated superior accuracy: CV maintained strong performance across depths (MEC = 0.834–0.868, RMSE = 0.214–0.254), peaking at 60–100 cm (MEC = 0.868, RMSE = 0.238; 86.8% variability explained). IV confirmed generalizability (MEC = 0.705–0.812, RMSE = 0.432–0.515).

Model performance also varied with depth. BD prediction accuracy exhibited non-monotonic depth dependence during independent validation, peaking at intermediate depths (15–30 cm: MEC = 0.657) with lower accuracy in surface layers (0–5 cm: MEC = 0.598) and deep layers (60–100 cm: MEC = 0.656). In contrast, pH prediction accuracy systematically increased with soil depth under both validation frameworks. CV showed MEC progression from 0.844 (0–5 cm) to 0.868 (60–100 cm), while IV demonstrated improvement from 0.705 (0–5 cm) to 0.812 (60–100 cm). Optimal pH performance consistently occurred at the deepest interval (60–100 cm) for both validation methods. In addition, all predictions maintained negligible bias ($|ME| \leqslant 0.019$) across depth intervals.

Table 2. Predictive performance of BD and pH predictions.

Validation	Depth (cm)	10-fold CV			IV		
		MEC	RMSE	ME	MEC	RMSE	ME
BD	0–5	0.782	0.090	0.000	0.598	0.164	-0.01
	5–15	0.815	0.084	0.000	0.611	0.181	-0.017
	15–30	0.828	0.081	-0.000	0.657	0.155	0.006
	30–60	0.874	0.079	-0.000	0.614	0.166	0.005
	60–100	0.889	0.087	0.000	0.656	0.166	-0.019
pH	0–5	0.844	0.215	0.000	0.705	0.432	-0.003
	5–15	0.834	0.254	0.000	0.726	0.480	-0.001
	15–30	0.854	0.214	0.000	0.742	0.448	-0.007
	30–60	0.854	0.256	0.001	0.760	0.515	-0.002
	60–100	0.868	0.238	0.001	0.812	0.492	0.014

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

3.3 Spatial patterns

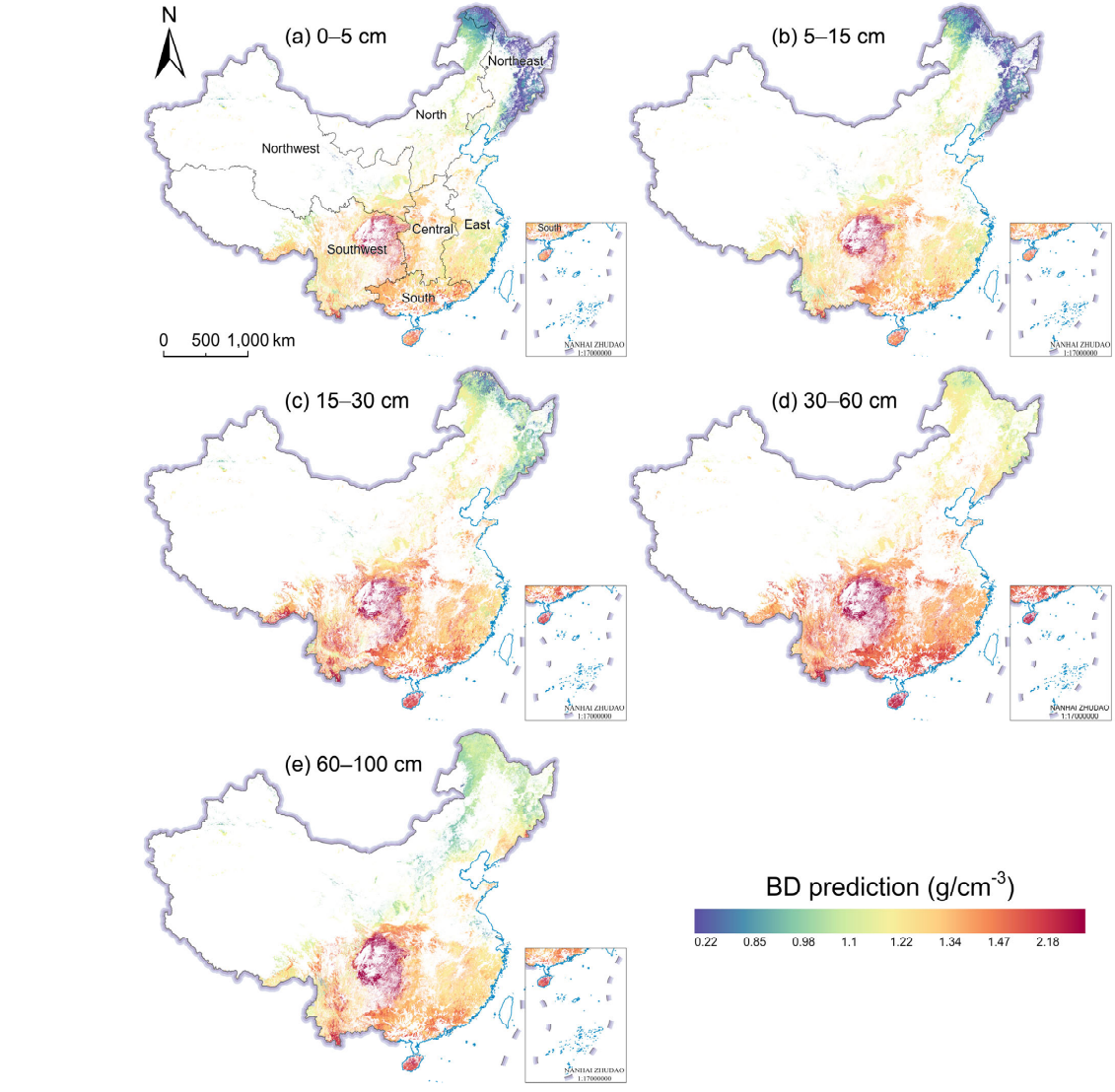
3.3.1 Spatial patterns of BD

265 The BD maps predicted by QRF (Figure 3) show consistent mean values ranging from 1.16 to 1.34 g/cm³ and standard deviations of 0.15–0.21 g/cm³ across all soil depths (Tables S3 and Fig S3). Macroscale patterns align with CSDLv2, ChinaSoilInfoGrids, and SoilGrids 2.0 (Fig. S1).

Spatially, BD exhibits distinct regional variation. The highest values occur in southwestern China (mean BD = 1.45 g/cm³) and the lowest values in northeastern China (mean BD = 0.79 g/cm³). Southwestern China consistently forms the highest-value 270 area across all soil layers. Northeastern China constitutes the lowest-value zone. In eastern China, BD values increase from the coast inland. Across the eastern coastal and southern regions, BD gradients occur from south to north and from coast to inland.

Vertically, BD increases with depth. Surface layers (0–5 cm) show the lowest BD, with minimal values concentrated in northeastern and southeastern coastal regions (mean values 0.79 g/cm³ and 1.19 g/cm³, respectively). High-BD zones in the 275 southwest expand slightly with depth. Within the middle soil depths (5–15 cm and 15–30 cm), spatial variability intensifies: low-BD zones extend from the northeast into North China, alongside distinct high-BD cores in the southwest. The 30–60 cm layer reaches the highest mean BD (1.32 g/cm³). The deep soil layer (60–100 cm) has a mean BD of 1.23 g/cm³, featuring extensive high-BD areas in southwest China and reduced low-BD coverage in northeastern areas.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025



280 **Figure 3. The predicted maps of predicted BD at 0–5 cm (a), 5–15 cm (b), 15–30 cm (c), 30–60 cm (d) and 60–100cm (e) depths.**
Publisher's remark: please note that the above figure contains disputed territories.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

3.3.2 Spatial patterns of pH

The predicted pH maps based on QRF are illustrated in Figure 4.. These maps showed mean values ranging from 5.70 to 6.06 and standard deviations ranging from 0.65 to 0.81 across all depths (Tables S3 and Fig.S4). Macroscale patterns align with those of CSDLv2, ChinaSoilInfoGrids, and SoilGrids 2.0 (Fig. S2).

Spatially, pH shows regional differentiation. Forest soils in South and Southwest China exhibit lower pH values (pH $\leqslant$ 5.76). Forest soils in North and Northwest China exhibit higher pH values (pH $\geqslant$ 6.50). Northeast China shows intermediate pH values, ranging between those of the southern and northern regions.

Vertically, pH patterns show both consistency with surface layers and changes with depth. Surface layers (0–5 cm, 5–15 cm) in the South and Southwest show the lowest pH values. With increasing soil depth (15–30 cm, 30–60 cm, and 60–100 cm), pH values in the southern regions increase. pH values in the northern regions become more stable with depth, showing reduced range. Overall, spatial variability in pH decreases in deeper soil layers.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

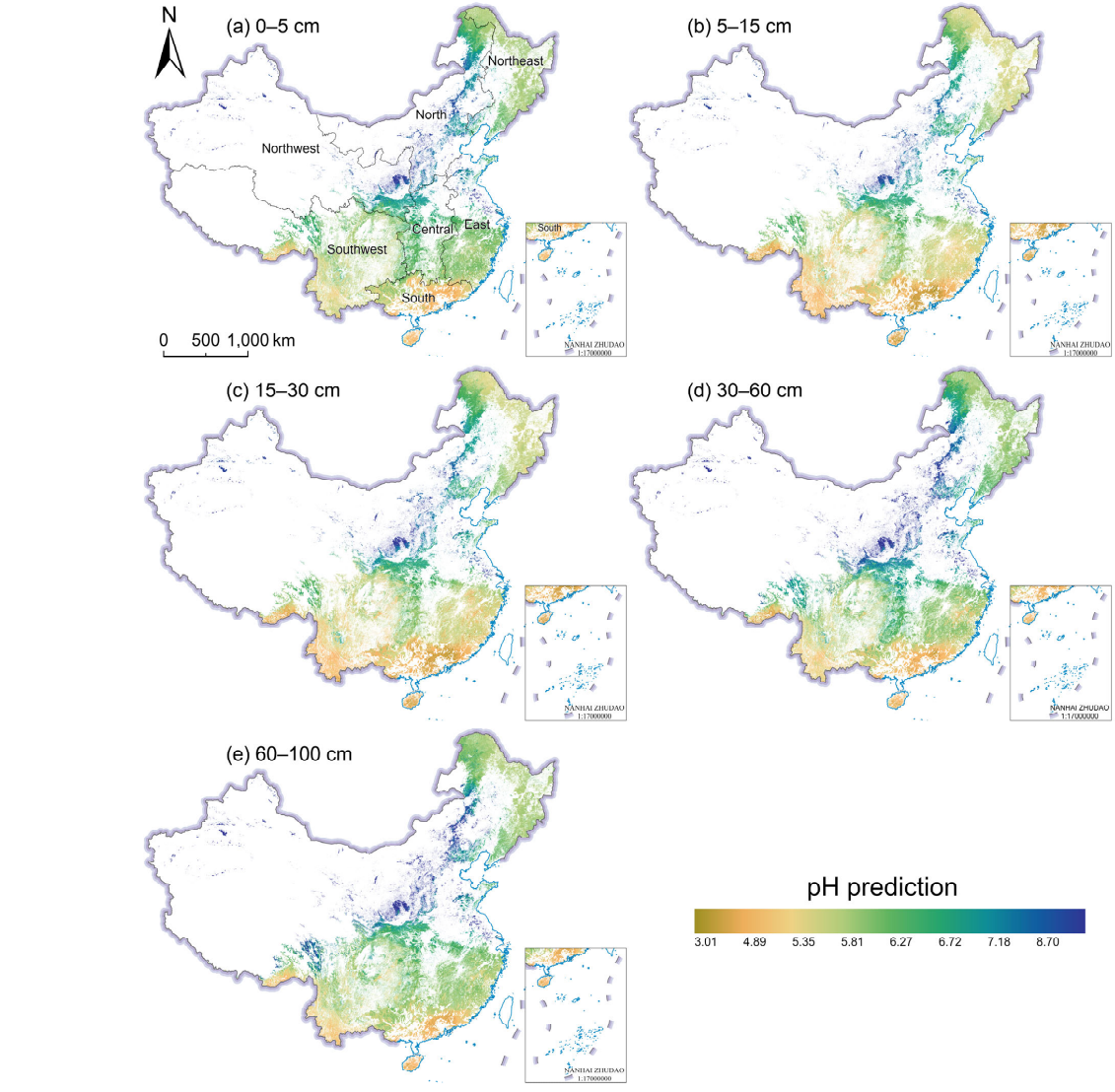


Figure 4. The predicted maps of predicted pH at 0–5 cm (a), 5–15 cm (b), 15–30 cm (c), 30–60 cm (d) and 60–100cm (e) depths.
Publisher's remark: please note that the above figure contains disputed territories.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

3.4 Prediction uncertainty

Visualization of prediction uncertainty using the PIR highlighted clear regional variations in uncertainty for BD and pH across China. Higher uncertainty for BD was concentrated in the northeastern and southwestern regions, while lower uncertainty characterized the southeastern coastal areas (Fig. S5). Conversely, pH uncertainty was more pronounced in northern China and parts of the southwest, with relatively lower levels observed in the northeast and the central-eastern coastal zone (Fig. S6). Overall, areas of elevated uncertainty predominantly coincided with southwestern China, where complex soil-landscape interactions likely contribute to increased model uncertainty. Additionally, regions with sparse data coverage, such as high-altitude areas, exhibited amplified extrapolation uncertainty due to limited representation in the training dataset, further challenging model reliability in these environments. For both BD and pH, prediction uncertainty generally increased with soil depth, a pattern potentially attributable to the reduced availability of soil observations at deeper intervals.

To ensure that biased uncertainty estimates do not compromise practical applications of the model, we further employed the PICP to perform this critical validation step. Five predictive accuracy plots were generated to evaluate the alignment of predicted intervals with actual observations for BD and pH (Figure 5). The QRF-based digital soil mapping model showed close adherence to the 1:1 reference line across both properties, indicating strong consistency in local uncertainty estimation. However, for pH, a slight overestimation of uncertainty was detected at intermediate probability levels (60%–90%) within subsurface layers (0–60 cm), suggesting minor deviations from optimal calibration. In contrast, uncertainty quantification for BD remained well-calibrated across all depth intervals and probability thresholds.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

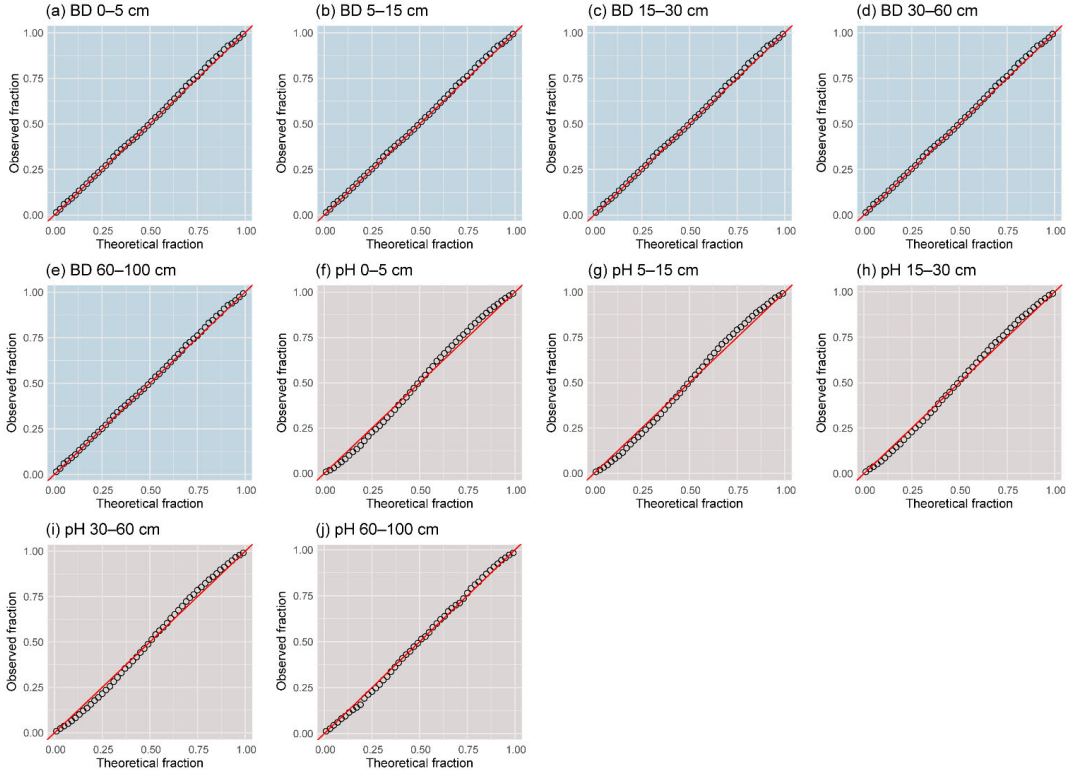


Figure 5. Validation of uncertainty quantifications.

3.5 Covariate importance

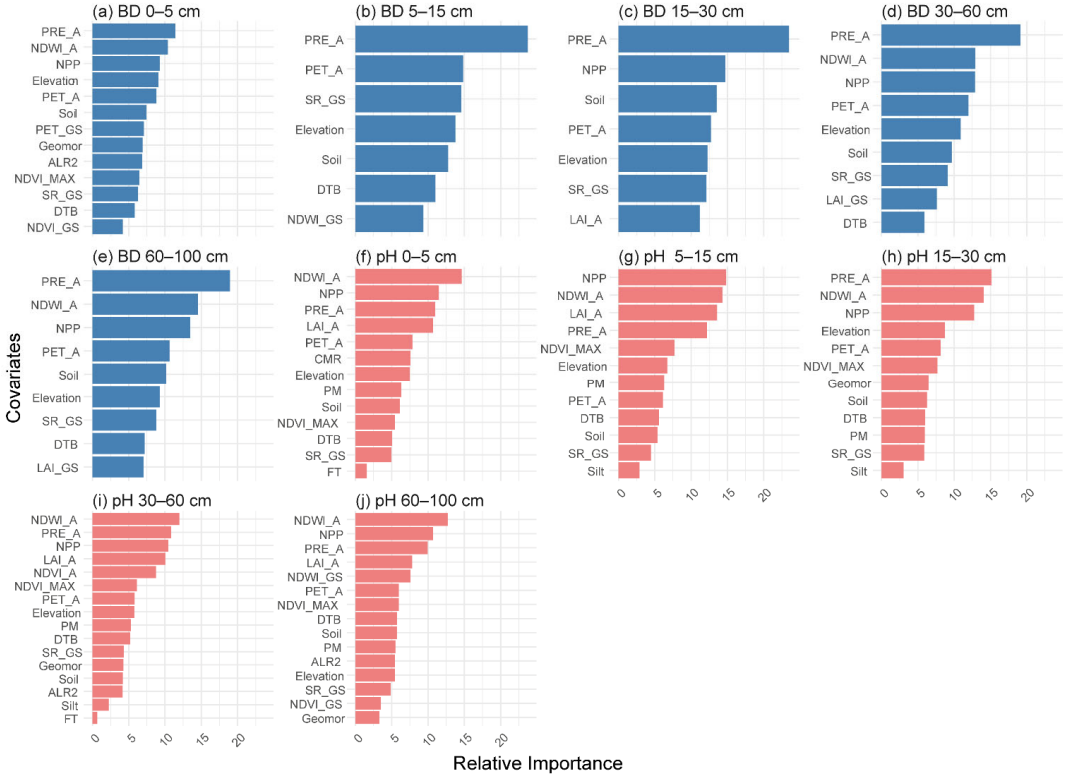
Relative importance of the environmental covariates used in the soil spatial predictions is shown in Figure 6. The FRFS framework enabled depth-specific dimensionality reduction, retaining 7 to 16 covariates per soil layer while eliminating 60.98% to 77.93% of the initial set (Table S4). Considerable variation in covariate relative importance was observed both between soil properties and across depths.

For forest soil BD prediction in the ensemble model, PRE_A showed the highest relative importance across all depths (0-100 cm), ranging from 11.41% to 23.73%. Other predictors exhibited clear depth-dependent variations in relative importance. In surface soils (0–15 cm), NDWI_A, NPP, Elevation, and Soil had notable relative importance. In middle layers (15–60 cm), the relative importance of NPP and Soil increased. In deep layers (60-100 cm), PRE_A, NDWI_A, and NPP maintained high relative importance, while parent material (PM) and bedrock depth (DTB) showed increased relative importance. Analysis by

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

325 factor category (Fig. S7) indicated that climate factors contributed the most to BD prediction across all depths (44.00%–62.47%), significantly exceeding the contributions from other factor categories.

For soil pH prediction in the ensemble model, the synergistic combination of NDWI_A, PRE_A, and NPP showed the highest relative importance across the entire soil profile (29.11–36.14%). Secondary factors exhibited depth-dependent variations in their relative importance. In surface soils (0–15 cm), vegetation indicators (LAI_A, NDVI_MAX) had the 330 strongest secondary relative importance. In middle layers (15–60 cm), the relative importance of topographic factors (Elevation, Geomor) and parent material (PM) increased. In the deep soil layer (60–100 cm), NDWI_GS and ALR2 showed increased relative importance, while DTB maintained consistently high relative importance. Analysis by factor category (Fig. S7) indicated that organism factors (primarily vegetation-related) contributed the most to the prediction (up to 36.14%), followed by varying relative contributions from climate and parent material factors across depths.



335

Figure 6. Variable importance for model training at different soil depths. The abbreviations of the predictors are defined in Table S1.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

4 Discussion

4.1 Model performance improvement

China has established valuable national soil datasets, such as the comprehensive CSDLv2 (Shi et al., 2025) and the high-resolution ChinaSoilInfoGrids (Liu et al., 2022a), which provide robust insights into soil properties across diverse ecosystems (Table S5). Building upon these foundations and focusing specifically on forest ecosystems, our work developed a model to simulate the spatial distribution of BD and pH within China's forests. As reported in Sect.3.2, the forest-optimized model demonstrated reliable performance for this targeted application under both CV and IV frameworks. Critically, the IV results, based solely on external forest samples withheld from model development, confirmed the model's robustness and generalizability within the specific context of China's forest ecosystems. These outcomes underscore the effectiveness of our approach in capturing forest-specific soil patterns.

Our methodology integrates two key innovations specifically designed to address the unique challenges of DSM in forest ecosystems. First, recognizing that forest soil sites are often underrepresented in large-scale national databases (which typically prioritize agricultural land) (Liu et al., 2022a; Shi et al., 2025), we constructed China's first dedicated, systematic forest soil dataset. This dataset, encompassing dominant forest types across major ecoregions, provides the essential foundation for characterizing forest-specific paedogenic processes and deriving accurate soil-environment relationships reflecting forest biogeochemistry.

Second, forest covariate optimization contributed significantly to the improved accuracy (Table S5). Recent DSM studies employing the SCORPAN framework have highlighted that relying solely on universal predictors may overlook critical ecosystem-specific variations, particularly in heterogeneous regions (Sun et al., 2022; Zhang et al., 2025). Building on this insight, our study explicitly incorporates forest-specific drivers to enhance mapping accuracy across China's complex landscapes. By tailoring covariate selection to forest ecosystems, we overcome the cross-ecosystem extrapolation bias prevalent in current national datasets. For instance, forest soil BD in China is mainly influenced by root-mediated aggregation (Liu et al., 2019; Zheng et al., 2023), while agricultural BD reflects region-specific tillage practices, such as the puddling effects in rice-wheat rotations (Hou et al., 2012). Similarly, forest soil pH is controlled by litterfall chemistry with distinct stoichiometry in Chinese subtropical forests (Zhou et al., 2016; Farooq et al., 2022), whereas agricultural pH is dominated by long-term nitrogen fertilization (Wang et al., 2019; Jia et al., 2022). Consequently, moving beyond universal predictors to select covariates specific to forest ecosystems was fundamental to the improved mapping accuracy.

Consequently, our forest-optimized framework, underpinned by the purpose-built dataset and ecologically informed covariates, generates a specialized and complementary perspective on soil property mapping for forest ecosystems. It delivers China's first comprehensive, 90-m resolution wall-to-wall maps characterizing forest BD and pH spatial patterns. These results validate the critical importance of incorporating ecosystem context into digital soil mapping (Padarian et al., 2019).

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

4.2 Potential applications

The high-resolution spatial dataset of forest soil BD and pH developed in this study represents a national-scale digital soil mapping product that captures the spatial variability of key soil physical and chemical properties across forested regions in China. Accurate knowledge of BD and pH is fundamental for estimating soil carbon stocks (Batjes, 2016), and monitoring forest ecosystem responses to land-use and climate change (Pan et al., 2011). Bulk density is critical for accurate estimation of soil carbon stocks, while pH governs nutrient availability, microbial activity, and forest productivity (Liu et al., 2024), and is significant for understanding forest soil acidification (Farooq et al., 2022). The dataset fills longstanding gaps in forest soil data coverage in China, and supports applications in ecosystem assessment and long-term soil monitoring. Beyond its scientific value, this product contributes to national strategies on carbon neutrality and ecological restoration and aligns with international environmental commitments including the UN Decade on Ecosystem Restoration and the Sustainable Development Goals (UNEP, 2021; IPCC, 2022).

4.3 Limitations and outlook

Our study advances high-resolution DSM in forest ecosystems, yet several methodological limitations remain and merit further investigation, particularly regarding the predictive reliability of machine learning approaches. Machine learning, while significantly enhancing DSM through capturing nonlinear soil–environment relationships, are constrained by limitations in spatial coverage and feature-space representativeness (Yang et al., 2013; Chen et al., 2019). Forests exhibit pronounced landscape heterogeneity, complicating sampling design and frequently resulting in imbalanced training datasets (Huang et al., 2022a; Liu et al., 2022b; Shao et al., 2022). As demonstrated by Westhuizen et al. (2024), models trained on such datasets yield biased predictions in undersampled regions. Although ensemble methods manage uncertainty in sparse data settings, they may prioritize statistical regularities over mechanistic soil formation processes (Sylvain et al., 2021; Liu et al., 2022b). Emerging hybrid frameworks integrating environmental similarity metrics with pedological expertise show promise in addressing these challenges (Zhao et al., 2024), though their scalability requires further validation (Miranda et al., 2023; Potash et al., 2023; Rodrigues et al., 2025). Specifically, strategic sampling designs incorporating stratified and adaptive approaches across diverse forest landscapes and soil types are crucial to mitigate dataset imbalance and capture underlying heterogeneity (Brus et al., 2011). Concurrently, exploring novel covariates derived from multi-source remote sensing (e.g., hyperspectral, LiDAR, radar) and proximal sensing (Xue et al., 2025), alongside improved representations of depth-dependent properties and long-term environmental legacies, could substantially enrich the feature space and better characterize the complex soil-forming factors operating in forest ecosystems (Vaysse and Lagacherie, 2017; Wadoux et al., 2020). Integrating such refined datasets within hybrid modeling frameworks holds considerable potential for improving the accuracy and reliability of forest DSM predictions.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

5 Data and code availability

The soil property maps generated in this study include soil pH and BD for five depth intervals (0–5 cm, 5–15 cm, 15–30 cm, All resources for the ensemble machine learning model, including training and testing code, are publicly available at https://github.com/cjz-ux/China_forest_DSM/tree/main. The soil property maps generated in this study include soil pH and BD for five depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, and 60–100 cm), with a spatial resolution of 90 meters. These maps are openly accessible via the platform link: https://doi.org/10.57760/sciencedb.25375 (last access: 19 September 2025) (Chen et al., 2025). Users can download the datasets efficiently using the provided FTP credentials and any standard FTP client.

6. Usage note

It is important to highlight that uncertainties associated with the spatial predictions of soil pH and BD have been not only quantified but also explicitly embedded in the corresponding maps. These uncertainty estimates offer critical insights into the reliability of predictions. Users are strongly encouraged to interpret the pH and BD maps alongside their respective uncertainty layers to ensure scientific rigor in downstream analyses and to support evidence-based decision-making and policy formulation. The inclusion of uncertainty information should not be regarded as a drawback. In fact, the adoption of standardized protocols for uncertainty quantification and reporting, which are now commonly used in DSM, enhances the transparency and applicability of the dataset. Users should also be aware that no spatial map represents a perfect depiction of reality. Interpreting these predictions without considering uncertainty introduces scientific and practical risks. The uncertainty layers serve as a guide for context-sensitive interpretation.

The current version of the China forest soil pH and BD grids is based on soil sampling limited to mainland China. Data from Hong Kong, Macau, and Taiwan are not included due to availability constraints. While this spatial extent reflects the existing sampling framework, future updates will aim to incorporate broader geographic coverage. Users should clearly acknowledge this limitation when applying the dataset in regional-scale modelling or policy-oriented analyses.
In addition, the environmental covariates used in the DSM workflow exhibit spatially heterogeneous coverage, with localized data gaps in certain regions (e.g., areas with steep elevation gradients or low-quality remote sensing input). To ensure model reliability, soil property predictions were restricted to areas where all covariates are fully available. Consequently, regions with missing covariate data were excluded from the final maps. Users should check the alignment of their study area with the covariate intersection mask, which is provided as ancillary metadata, to confirm the spatial applicability of the dataset.

7 Conclusions

Our study developed the first high-resolution mapping of forest soil BD and pH across China, leveraging forest soil profiles from the latest national forest soil survey. We achieved this detailed characterization across complex forest soil landscapes by

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

integrating the predictive soil mapping paradigm with FRFS, QRF, and a detailed suite of forest-specific soil-forming environmental factors within a high-performance parallel computing environment. This integrated approach not only effectively reduced errors and training time but also enhanced the performance of the final predictive models. The resultant multilayer maps delineate pronounced regional gradients and fine-scale forest soil heterogeneity across depths, outperforming existing products in accuracy, spatial detail, and provision of local uncertainty metrics. These high-resolution forest soil property maps represent a contribution to the GlobalSoilMap.net project and provide critical baseline data for China's forest carbon accounting and understanding of soil acidification processes.

Author contributions.

Conceptualization: JZC; Data curation: QWS, XYS, JZC, ZHF, XZ, and ZLH; Formal analysis: JZC; Funding acquisition: ZLH and WFX; Methodology: JZC and XZ; Supervision: JZC, ZLH, and WFX; Validation: JZC; Writing – original draft preparation: JZC, ZLH, and TL; Writing – review & editing: JZC, ZLH, and TL.

Competing interests.

The contact author has declared that none of the authors has any competing interests.

Acknowledgements.

This work was supported by the National Key Research and Development Program of China (No.2021FY100800).
We would like to express our gratitude to Professor Feng Liu at the Institute of Soil Science, Chinese Academy of Sciences (Nanjing, China), for his valuable suggestions that contributed to this study.

References

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M. d.L., Minasny, B., Montanarella, L., Odeh, I. O. A., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: GlobalSoilMap, in: Advances in Agronomy, vol. 125, Elsevier, 93–134, https://doi.org/10.1016/B978-0-12-800137-0.00003-0, 2014.

Batjes, N. H.: Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks, Geoderma, 269, 61–68, https://doi.org/10.1016/j.geoderma.2016.01.034, 2016.

Binkley, D. and Fisher, R. F.: Ecology and management of forest soils, 4th ed., Wiley, Hoboken, NJ, 347 pp., 2013.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Bishop, T. F. A., McBratney, A. B., and Laslett, G. M.: Modelling soil attribute depth functions with equal-area quadratic sm oothing splines, Geoderma, 91, 27–45, https://doi.org/10.1016/S0016-7061(99)00003-8, 1999.

Brus, D. J., Kempen, B., and Heuvelink, G. B. M.: Sampling for validation of digital soil maps, Eur. J. Soil Sci., 62, 394–407, https://doi.org/10.1111/j.1365-2389.2011.01364.x, 2011.

Chen, J. Z.; Huang, Z. L. (2025). High-resolution maps of forest soil bulk density and pH across China at 90-m resolution [DS/OL]. V1. Science Data Bank. https://doi.org/10.57760/sciencedb.25375.

Chen, J., Deng, Z., Jiang, Z., Sun, J., Meng, F., Zuo, X., Wu, L., Cao, G., and Cao, S.: Variations of rhizosphere and bulk soi l microbial community in successive planting of chinese fir (cunninghamia lanceolata), Front. Plant Sci., 13, 954777, https:// doi.org/10.3389/fpls.2022.954777, 2022a.

Chen, L. F., He, Z. B., Du, J., Yang, J. J., and Zhu, X.: Patterns and environmental controls of soil organic carbon and total ni trogen in alpine ecosystems of northwestern China, CATENA, 137, 37–43, https://doi.org/10.1016/j.catena.2015.08.017, 201 6.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hann am, J., Meersmans, J., Richer-de-Forges, A. C., and Walter, C.: Digital mapping of GlobalSoilMap soil properties at a broad scale: a review, Geoderma, 409, 115567, https://doi.org/10.1016/j.geoderma.2021.115567, 2022b.

Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de-Forges, A. C., Saby, N. P. A., Loiseau, T., Hu, B., and Arrouays, D.: Probability mapping of soil thickness by random survival forest at a national scale, Geoderma, 344, 184–1 94, https://doi.org/10.1016/j.geoderma.2019.03.016, 2019.

Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Hu, B., Zhou, Y., Wang, N., Arrouays, D., and Shi, Z.: Evaluating validati on strategies on the performance of soil property prediction from regional to continental spectral data, Geoderma, 400, 11515 9, https://doi.org/10.1016/j.geoderma.2021.115159, 2021.

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., and Yan, F.: A review of the globa l soil property maps for Earth system models, SOIL, 5, 137–158, https://doi.org/10.5194/soil-5-137-2019, 2019.

FAO and IIASA: Harmonized World Soil Database version 2.0, FAO, International Institute for Applied Systems Analysis (I IASA) [data set], https://doi.org/10.4060/cc3823en, 2023.

Farooq, T. H., Li, Z., Yan, W., Shakoor, A., Kumar, U., Shabbir, R., Peng, Y., Gayathiri, E., Alotaibi, S. S., Wróbel, J., Kalaj i, H. M., and Chen, X.: Variations in litterfall dynamics, C:N:P stoichiometry and associated nutrient return in pure and mixe d stands of camphor tree and masson pine forests, Front. Environ. Sci., 10, 903039, https://doi.org/10.3389/fenvs.2022.90303 9, 2022.

Goovaerts, P.: Geostatistical modelling of uncertainty in soil science, Geoderma, 103, 3–26, https://doi.org/10.1016/S0016-7 061(01)00067-2, 2001.

Grundy, M. J., Rossel, R. A. V., Searle, R. D., Wilson, P. L., Chen, C., and Gregory, L. J.: Soil and landscape grid of Austral ia, Soil Res., 53, 835, https://doi.org/10.1071/SR15191, 2015.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Helfenstein, A., Mulder, V. L., Hack-ten Broeke, M. J. D., Van Doorn, M., Teuling, K., Walvoort, D. J. J., and Heuvelink, G. B. M.: BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands, Earth Syst. Sci. Data, 16, 2941–2970, https://doi.org/10.5194/essd-16-2941-2024, 2024.

Hempel, J., McBratney, A., Arrouays, D., McKenzie, N., and Hartemink, A.: GlobalSoilMap project history, in: GlobalSoilMap, edited by: Arrouays, D., McKenzie, N., Hempel, J., De Forges, A., and McBratney, A., CRC Press, 3–8, https://doi.org/10.1201/b16500-3, 2014.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hou, X. Q., Li, R., Jia, Z. K., Han, Q. F., Yang, B. P., and Nie, J. F.: Effects of rotational tillage practices on soil structure, organic carbon concentration and crop yields in semi-arid areas of northwest China, Soil Use Manage., 28, 551–558, https://doi.org/10.1111/j.1475-2743.2012.00429.x, 2012.

Hounkpatin, K. O. L., Stendahl, J., Lundblad, M., and Karltun, E.: Predicting the spatial distribution of soil organic carbon stock in swedish forests using a group of covariates and site-specific data, Soil, 7, 377–398, https://doi.org/10.5194/soil-7-377-2021, 2021.

Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Wu, Q., Cai, Y., Shen, F., and Zhou, C.: A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect, Environ. Res. Lett., 17, 123004, https://doi.org/10.1088/1748-9326/aca41e, 2022a.

Huang, X., Cui, C., Hou, E., Li, F., Liu, W., Jiang, L., Luo, Y., and Xu, X.: Acidification of soil due to forestation at the global scale, For. Ecol. Manage., 505, 119951, https://doi.org/10.1016/j.foreco.2021.119951, 2022b.

IPCC.: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 2022.

Jenny, H.: Factors of soil formation: a system of quantitative pedology, McGraw-Hill, New York, 1941.

Jeune, W., Francelino, M. R., Souza, E. D., Fernandes Filho, E. I., and Rocha, G. C.: Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti, Rev. Bras. Cienc. Solo, 42, https://doi.org/10.1590/18069657rbcs20170133, 2018.

Jia, S., Yuan, D., Li, W., He, W., Raza, S., Kuzyakov, Y., Zamanian, K., and Zhao, X.: Soil Chemical Properties Depending on Fertilization and Management in China: A Meta-Analysis, Agronomy, 12, 2501, https://doi.org/10.3390/agronomy12102501, 2022.

Kleber, M., Bourg, I. C., Coward, E. K., Hansel, C. M., Myneni, S. C. B., and Nunan, N.: Dynamic interactions at the mineral–organic matter interface, Nat. Rev. Earth Environ., 2, 402–421, https://doi.org/10.1038/s43017-021-00162-y, 2021.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., and Viscarra Rossel, R. A.: National digital soil map of organic matter in top soil and its associated uncertainty in 1980's china, Geoderma, 335, 47–56, https://doi.org/10.1016/j.geoderma.2018.08.011, 2019.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J., Song, X., Shi, Z., Zhu, A., and Zhang, G.: Mapping high resolution national soil information grids of China, Sci. Bull., 67, 328–340, https://doi.org/10.1016/j.scib.2021.10.013, 2022a.

Liu, F., Yang, F., Zhao, Y., Zhang, G., and Li, D.: Predicting soil depth in a large and complex area using machine learning and environmental correlations, J. Integr. Agric., 21, 2422–2434, https://doi.org/10.1016/S2095-3119(21)63692-4, 2022b.

Liu, R., Zhou, X., Wang, J., Shao, J., Fu, Y., Liang, C., Yan, E., Chen, X., Wang, X., and Bai, S. H.: Differential magnitude of rhizosphere effects on soil aggregation at three stages of subtropical secondary forest successions, Plant Soil, 436, 365–380, https://doi.org/10.1007/s11104-019-03935-z, 2019.

Liu, Z., Gu, H., Yao, Q., Jiao, F., Hu, X., Liu, J., Jin, J., Liu, X., and Wang, G.: Soil pH and carbon quality index regulate the biogeochemical cycle couplings of carbon, nitrogen and phosphorus in the profiles of isohumosols, Sci. Total Environ., 922, 171269, https://doi.org/10.1016/j.scitotenv.2024.171269, 2024.

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

Meinshausen, N.: Quantile regression forests, J. Mach. Learn. Res., 7, 983–999, 2006.

Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I.: Digital mapping of soil carbon, in: Advances in Agronomy, vol. 118, Elsevier, 1–47, https://doi.org/10.1016/B978-0-12-405942-9.00001-3, 2013.

Miranda, R., Nobrega, R., Silva, E., Silva, J., Araújo Filho, J., Moura, M., Barros, A., Souza, A., Verhoef, A., Yang, W., Shao, H., Srinivasan, R., Ziadat, F., Montenegro, S., Araújo, M., and Galvíncio, J.: Hybrid machine learning for integrating pedological knowledge into digital soil mapping to advance next-generation earth system models, https://doi.org/10.31223/X57P9W, 2023.

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., and Arrouays, D.: GlobalSoilMap france: high-resolution spatial modelling the soils of France up to two meter depth, Sci. Total Environ., 573, 1352–1369, https://doi.org/10.1016/j.scitotenv.2016.07.066, 2016.

Osman, K. T.: Forest soils, in: Soils, Springer Netherlands, Dordrecht, 229–251, https://doi.org/10.1007/978-94-007-5663-2_14, 2013.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning for digital soil mapping, SOIL, 5, 79–89, https://doi.org/10.5194/soil-5-79-2019, 2019.

Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., and Hayes, D.: A large and persistent carbon sink in the world's forests, Science, 333, 988–993, https://doi.org/10.1126/science.1201609, 2011.

Patton, N. R., Lohse, K. A., Seyfried, M. S., Godsey, S. E., and Parsons, S. B.: Topographic controls of soil organic carbon on soil-mantled landscapes, Sci Rep, 9, 6390, https://doi.org/10.1038/s41598-019-42556-5, 2019.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: pr
555 oducing soil information for the globe with quantified spatial uncertainty, Soil, 7, 217–240, https://doi.org/10.5194/soil-7-21
7-2021, 2021.

Potash, E., Guan, K., Margenot, A. J., Lee, D., Boe, A., Douglass, M., Heaton, E., Jang, C., Jin, V., Li, N., Mitchell, R., Nam
oi, N., Schmer, M., Wang, S., and Zumpf, C.: Multi-site evaluation of stratified and balanced sampling of soil organic carbon
stocks in agricultural fields, Geoderma, 438, 116587, https://doi.org/10.1016/j.geoderma.2023.116587, 2023.

560 Pouladi, N., Møller, A. B., Tabatabai, S., and Greve, M. H.: Mapping soil organic matter contents at field level with cubist, r
andom forest and kriging, Geoderma, 342, 85–92, https://doi.org/10.1016/j.geoderma.2019.02.019, 2019.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., and Thompson, J.: Soil property and class map
s of the conterminous united states at 100-meter spatial resolution, Soil Sci. Soc. Am. J., 82, 186–201, https://doi.org/10.2136
/sssaj2017.04.0122, 2018.

565 Rodrigues, H., Ceddia, M. B., Vasques, G. M., Grunwald, S., and Babaeian, E.: AutoRA: an innovative algorithm for automa
tic delineation of reference areas in support of smart soil sampling and digital soil twins, Front. Soil Sci., 5, 1557566, https://
doi.org/10.3389/fsoil.2025.1557566, 2025.

Santos, P. A. D., Pinheiro, H. S. K., Carvalho, W. D. C., Silva, I. L. D., Pereira, N. R., Bhering, S. B., and Ceddia, M. B.: Hy
dropedological digital mapping: machine learning applied to spectral VIS-IR and radiometric data dimensionality reduction,

570 Rev. Bras. Cienc. Solo, 47, e0220149, https://doi.org/10.36783/18069657rbcs20220149, 2023.

Shao, S., Su, B., Zhang, Y., Gao, C., Zhang, M., Zhang, H., and Yang, L.: Sample design optimization for soil mapping using
improved artificial neural networks and simulated annealing, Geoderma, 413, 115749, https://doi.org/10.1016/j.geoderma.20
22.115749, 2022.

Shi, G., Sun, W., Shangguan, W., Wei, Z., Yuan, H., Li, L., Sun, X., Zhang, Y., Liang, H., Li, D., Huang, F., Li, Q., and Dai,

575 Y.: A China dataset of soil properties for land surface modelling (version 2, CSDLv2), Earth Syst. Sci. Data, 17, 517–543, ht
tps://doi.org/10.5194/essd-17-517-2025, 2025.

Song, X., Wu, H., Ju, B., Liu, F., Yang, F., Li, D., Zhao, Y., Yang, J., and Zhang, G.: Pedoclimatic zone-based three-dimensi
onal soil organic carbon mapping in China, Geoderma, 363, 114145, https://doi.org/10.1016/j.geoderma.2019.114145, 2020.

Sun, X. L., Lai, Y. Q., Ding, X., Wu, Y. J., Wang, H. L., and Wu, C.: Variability of soil mapping accuracy with sample sizes,

580 modelling methods and landform types in a regional case study, Catena, 213, 106217, https://doi.org/10.1016/j.catena.2022.
106217, 2022.

Sylvain, J.-D., Anctil, F., and Thiffault, É.: Using bias correction and ensemble modelling for predictive mapping and related
uncertainty: a case study in digital soil mapping, Geoderma, 403, 115153, https://doi.org/10.1016/j.geoderma.2021.115153,
2021.

585 Szatmári, G., Laborczi, A., Mészáros, J., Takács, K., Benő, A., Koós, S., Bakacsi, Z., and Pásztor, L.: Gridded, temporally re
ferenced spatial information on soil organic carbon for Hungary, Sci. Data, 11, 1312, https://doi.org/10.1038/s41597-024-041
58-3, 2024.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

Thompson, J. A., Kienast-Brown, S., D'Avello, T., Philippe, J., and Brungard, C.: Soils2026 and digital soil mapping – A fo
undation for the future of soils information in the United States, Geoderma Regional, 22, e00294, https://doi.org/10.1016/j.ge
odrs.2020.e00294, 2020.

UNEP.: Becoming #GenerationRestoration: Ecosystem Restoration for People, Nature and Climate. United Nations Environ
ment Programme, 2021.

Vaysse, K. and Lagacherie, P.: Using quantile regression forest to estimate uncertainty of digital soil mapping products, Geo
derma, 291, 55–64, https://doi.org/10.1016/j.geoderma.2016.12.017, 2017.

Wadoux, A., Minasny, B., and McBratney, A.: Machine learning for digital soil mapping: applications, challenges and sugge
sted solutions, https://doi.org/10.31223/OSF.IO/8EQ6S, 6 February 2020.

Wang, H., Xu, J., Liu, X., Zhang, D., Li, L., Li, W., and Sheng, L.: Effects of long-term application of organic fertilizer on i
mproving organic matter content and retarding acidity in red soil from China, Soil Tillage Res., 195, 104382, https://doi.org/
10.1016/j.still.2019.104382, 2019.

Westhuizen, S. V. D., Heuvelink, G. B. M., Hofmeyr, D. P., Poggio, L., Nussbaum, M., and Brungard, C.: Mapping soil thic
kness by accounting for right-censored data with survival probabilities and machine learning, Eur. J. Soil Sci., 75, e13589, ht
tps://doi.org/10.1111/ejss.13589, 2024.

Widyati, E., Nuroniah, H. S., Tata, H. L., Mindawati, N., Lisnawati, Y., Abdulah, L., Lelana, N. E., Octavia, D., Prameswari,
D., and Rachmat, H. H.: Soil degradation due to conversion from natural to plantation forests in indonesia, Forests, 13, 1913,
https://doi.org/10.3390/f13111913, 2022.

Wu, X., Yuan, Z., Li, D., Zhou, J., and Liu, T.: Geographic variations of pore structure of clayey soils along a climatic gradie
nt, Catena, 222, 106861, https://doi.org/10.1016/j.catena.2022.106861, 2023.

Xiao, Y., Xue, J., Zhang, X., Wang, N., Hong, Y., Jiang, Y., Zhou, Y., Teng, H., Hu, B., Lugato, E., Richer-de-Forges, A. C.,
Arrouays, D., Shi, Z., and Chen, S.: Improving pedotransfer functions for predicting soil mineral associated organic carbon b
y ensemble machine learning, Geoderma, 428, 116208, https://doi.org/10.1016/j.geoderma.2022.116208, 2022.

Xu, L., He, N. P., Yu, G. R., Wen, D., Gao, Y., and He, H. L.: Differences in pedotransfer functions of bulk density lead to hi
gh uncertainty in soil organic carbon estimation at regional scales: evidence from chinese terrestrial ecosystems, J. Geophys.
Res.: Biogeosci., 120, 1567–1575, https://doi.org/10.1002/2015JG002929, 2015.

Xue, J., Zhang, X., Chen, S., Chen, Z., Lu, R., Liu, F., Van Wesemael, B., and Shi, Z.: National-scale mapping topsoil organi
c carbon of cropland in China using multitemporal sentinel-2 images, Geoderma, 456, 117272, https://doi.org/10.1016/j.geod
erma.2025.117272, 2025.

Yang, L., Zhu, A. X., Qi, F., Qin, C. Z., Li, B., and Pei, T.: An integrative hierarchical stepwise sampling strategy for spatial
sampling and its application in digital soil mapping, Int. J. Geogr. Inf. Sci., 27, 1–23, https://doi.org/10.1080/13658816.2012.
658053, 2013.

https://doi.org/10.5194/essd-2025-496
Preprint. Discussion started: 5 November 2025

620 Zhang, L., Yang, L., Ma, Y., Zhu, A.-X., Wei, R., Liu, J., Greve, M. H., and Zhou, C.: Regional-scale soil carbon predictions can be enhanced by transferring global-scale soil–environment relationships, Geoderma, 461, 117466, https://doi.org/10.1016/j.geoderma.2025.117466, 2025.

Zhang, S., Zhou, X., Chen, Y., Du, F., and Zhu, B.: Soil organic carbon fractions in China: spatial distribution, drivers, and future changes, Sci. Total Environ., 919, 170890, https://doi.org/10.1016/j.scitotenv.2024.170890, 2024.

625 Zhao, C., Long, J., Liao, H., Zheng, C., Li, J., Liu, L., and Zhang, M.: Dynamics of soil microbial communities following vegetation succession in a karst mountain ecosystem, southwest China, Sci. Rep., 9, 2160, https://doi.org/10.1038/s41598-018-36886-z, 2019.

Zhao, F., Zhu, A., Zhu, L., and Qin, C.: iSoLIM : a similarity-based spatial prediction software for the big data era, Ann. Gis, 30, 535–549, https://doi.org/10.1080/19475683.2024.2324381, 2024.

630 Zheng, Y., Wang, Y., Zhang, Y., Zhang, J., Wang, Y., and Zhu, J.: Broadleaf trees increase soil aggregate stability in mixed forest stands of southwest China, Forests, 14, 2402, https://doi.org/10.3390/f14122402, 2023.

Zhou, J., Lang, X., Du, B., Zhang, H., Liu, H., Zhang, Y., and Shang, L.: Litterfall and nutrient return in moist evergreen broad-leaved primary forest and mixed subtropical secondary deciduous broad-leaved forest in China, Eur. J. For. Res., 135, 77–86, https://doi.org/10.1007/s10342-015-0918-7, 2016.

635 Zhu, Q., De Vries, W., Liu, X., Zeng, M., Hao, T., Du, E., Zhang, F., and Shen, J.: The contribution of atmospheric deposition and forest harvesting to forest soil acidification in China since 1980, Atmos. Environ., 146, 215–222, https://doi.org/10.1016/j.atmosenv.2016.04.023, 2016.