High-Resolution, Multi-Depth Mapping of Soil Bulk Density and pH in China's Forests Using Machine Learning

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 forest soil bulk density (BD) and pH 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 high-resolution (90-m), forest-specific maps of soil 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 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.

1 Introduction

Soil bulk density (BD) and pH are fundamental properties that govern the physical and chemical environment of forest soils. They critically influence key ecosystem processes such as water infiltration, nutrient cycling, and microbial activity, and serve as essential parameters for quantifying forest carbon storage and assessing soil acidification dynamics at regional to global scales (Dai et al., 2019; Kleber et al., 2021). Consequently, accurate, high-resolution spatial information on these properties is

indispensable for advancing our understanding of forest ecosystem functions and their responses to environmental change (Zhu et al., 2016; Huang et al., 2022b; Xu et al., 2015).

Digital Soil Mapping (DSM) addresses the need by serving as an innovative alternative to traditional soil surveys, aiming to generate continuous, high-resolution soil property maps through the quantitative modeling of relationships between soil observations and environmental covariates (McBratney et al., 2003; Minasny et al., 2013; Padarian et al., 2019). Recently, the widespread application of machine learning (ML) algorithms has advanced the field forward, making it central to acquiring soil information from regional to global scales (Chen et al., 2022; Westhuizen et al., 2024). ML models such as Random Forest (RF), Cubist, and Support Vector Machines (SVM) are highly favoured for their ability to capture complex nonlinear relationships between soil and environmental factors (Khaledian and Miller, 2020; Wadoux et al., 2020). A common paradigm in ML-based DSM is the integration of multi-source environmental data within the SCORPAN (Soil, Climate, Organisms, Relief, Parent material, Age, and Space) framework (McBratney et al., 2003). However, this integration often leads to high-dimensional feature spaces, introducing computational and statistical challenges known as the "curse of dimensionality." Beyond this, even models with high predictive accuracy often lack robust mechanisms for quantifying prediction uncertainty, which is a critical aspect for reliable spatial decision-making (Heuvelink et al., 2016).

To navigate these challenges, a dual-focused strategy is increasingly employed. First, feature selection methods, such as the recently advanced Forward Recursive Feature Selection (FRFS), are essential to identify the most informative covariates, thereby enhancing model efficiency and accuracy (Xiao et al., 2022; Xue et al., 2025). Second, algorithms like Quantile Regression Forest (QRF) have gained prominence for their ability to provide reliable prediction intervals alongside accurate predictions, effectively quantifying spatial uncertainty (Vaysse and Lagacherie, 2017). QRF has been successfully applied to map the spatial distribution of BD and pH at national and global scales, effectively quantifying prediction uncertainty (Poggio et al., 2021; Liu et al., 2022a; Xue et al., 2025). Despite its advantages, QRF, similar to many machine learning models, operates as a 'black box,' offering limited insight into how individual environmental drivers contribute to the final predictions. This lack of interpretability can hinder a pedological understanding of the spatial patterns derived from the maps. Given the demonstrated capability of SHAP (SHapley Additive exPlanations) in interpreting complex model predictions, we integrate it into our DSM framework to decompose and explain the outputs of the QRF model, thereby elucidating the role of key environmental drivers (Lundberg and Lee, 2017).

The methodological developments have enabled the production of high-resolution digital soil maps at unprecedented spatial scales. In recent years, significant progress has been made in developing high-resolution digital soil maps for key properties like BD and pH at national and global scales. Prominent examples include the SoilGrids 2.0 global product (Poggio et al., 2021), ChinaSoilInfoGrids (Liu et al., 2022a), and the China Soil Data Library version 2 (CSDLv2) (Shi et al., 2025). However, these existing datasets, while valuable for broad-scale applications, are primarily derived from soil profiles sampled across mixed land cover types (e.g., cropland, grassland, and forest). For dedicated forest ecosystem applications, these general-purpose products exhibit critical limitations. First, they rely on legacy soil profiles from national surveys (e.g., the Second National Soil Survey of China) and global databases where forest soils constitute only a minor fraction (typically less

than 20%) of the total samples (Zhao et al., 2019; Chen et al., 2022; Liu et al., 2024b). This inadequate representation fails to capture the distinct pedogenic processes and spatial heterogeneity inherent to diverse forest ecosystems. Second, the predictive models and environmental covariates within these datasets are optimized for broad-scale mapping across all ecosystems. Consequently, they may overlook forest-specific controlling factors (e.g., forest type) that are crucial for accurately predicting forest soil BD and pH (Zhao et al., 2019; Chen et al., 2022; Liu et al., 2024b).

China's forest ecosystems cover 209 million hectares, spanning diverse climatic zones and complex topographies, and encompass 452 vegetation types, making it one of the most ecologically diverse forest regions on Earth (Chen et al., 2016; Patton et al., 2019; Zhang et al., 2024). Forest soils exhibit considerable heterogeneity across geographical space, influenced by long-term climatic gradients, vegetation succession, and topographic variation (Zhao et al., 2019; Chen et al., 2022; Liu et al., 2024b). Revealing the spatial distribution of forest BD and pH is essential for estimating carbon storage and assessing soil acidification risks in forests (Zhu et al., 2016; Huang et al., 2022b; Xu et al., 2015). However, to the best of our knowledge, no previous study has estimated these properties at high resolution across China's forests. To address this gap, we developed a comprehensive forest soil database for China, comprising 8,709 soil profiles and 18,193 samples across five depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, and 60–100 cm), ensuring broad representativeness of China's diverse forest ecosystems. Our study leverages the uncertainty-handling capabilities of the QRF model, combined with an optimized variable selection process using the FRFS. The key objectives of this study are: (1) to establish a nationwide forest soil profile database; (2) to develop and apply an optimized DSM framework integrating QRF and FRFS; and (3) to create high-resolution (90-meter) digital maps for two critical forest soil properties, BD and pH, across the entire forest domain of China. These maps provide continuous, wall-to-wall spatial characterization of BD and pH at a 90-meter resolution, in compliance with GlobalSoilMap standards. The dataset provides a critical spatial baseline for accurately quantifying forest carbon stocks and assessing soil acidification risks across the country.

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%) subsets The spatial distributions of the training and validation samples were examined to ensure that both subsets maintain comparable spatial coverage across regions and soil depths (Fig.S2; Tables S2–S4). 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.

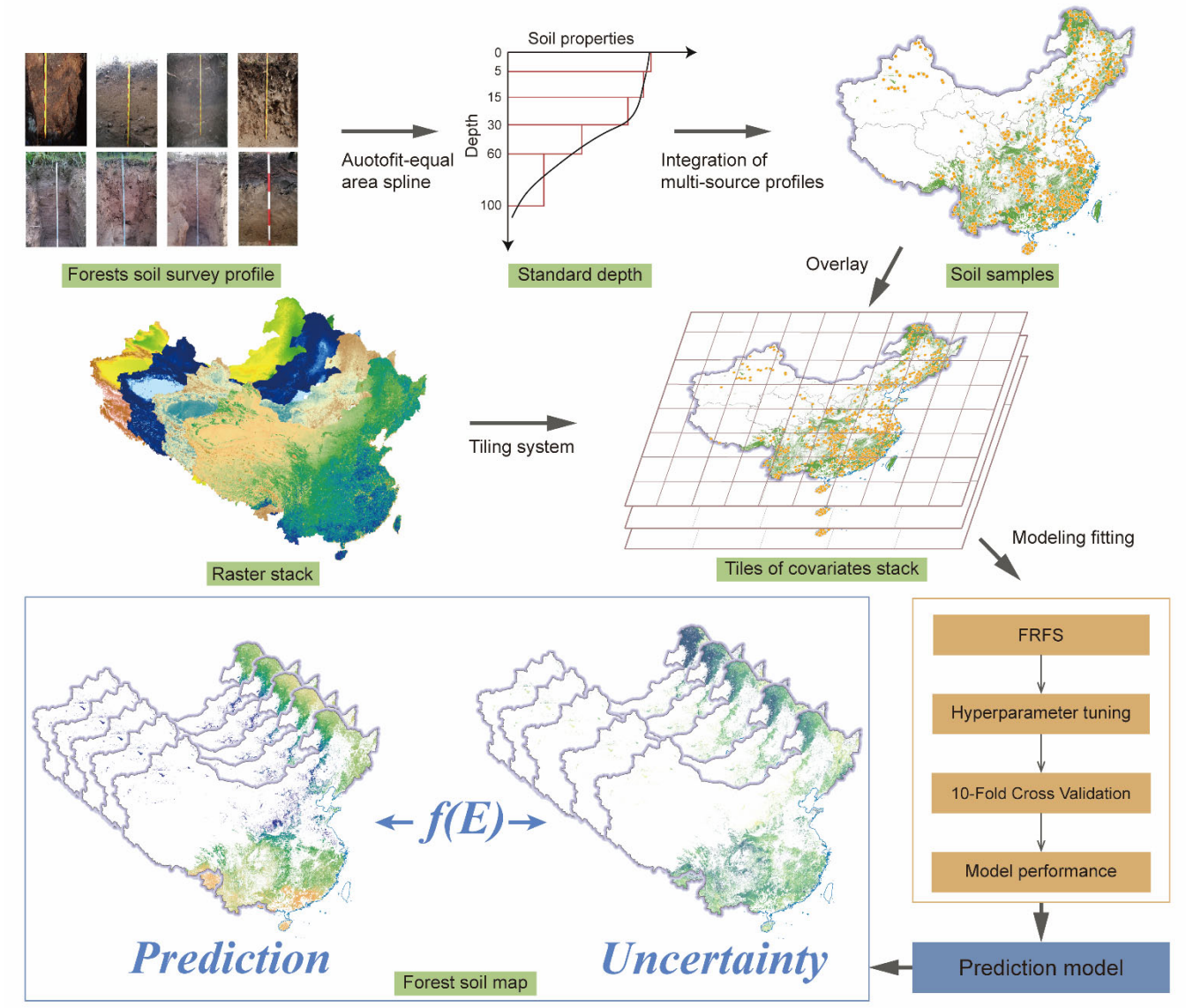


Figure 1. Workflow diagram for forest soil mapping.

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 (Figure 2). 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, with detailed profile counts for each category summarized in Tables S2–S5. Quality control procedures included the removal of duplicated records, consistency checks of geographic coordinates, and screening for physically implausible values of soil BD and pH based on established reference ranges.

Data harmonization involved the standardization of measurement units, alignment of soil depth intervals to the five target layers using mass-preserving spline functions, and the reconciliation of metadata across different survey sources. 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. These profiles were collected between 2018 and 2023, with profile counts of 2,045 in 2018, 157 in 2019, 530 in 2021, 621 in 2022, and 1,003 in 2023, reflecting the timing and scope of different survey campaigns. 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.

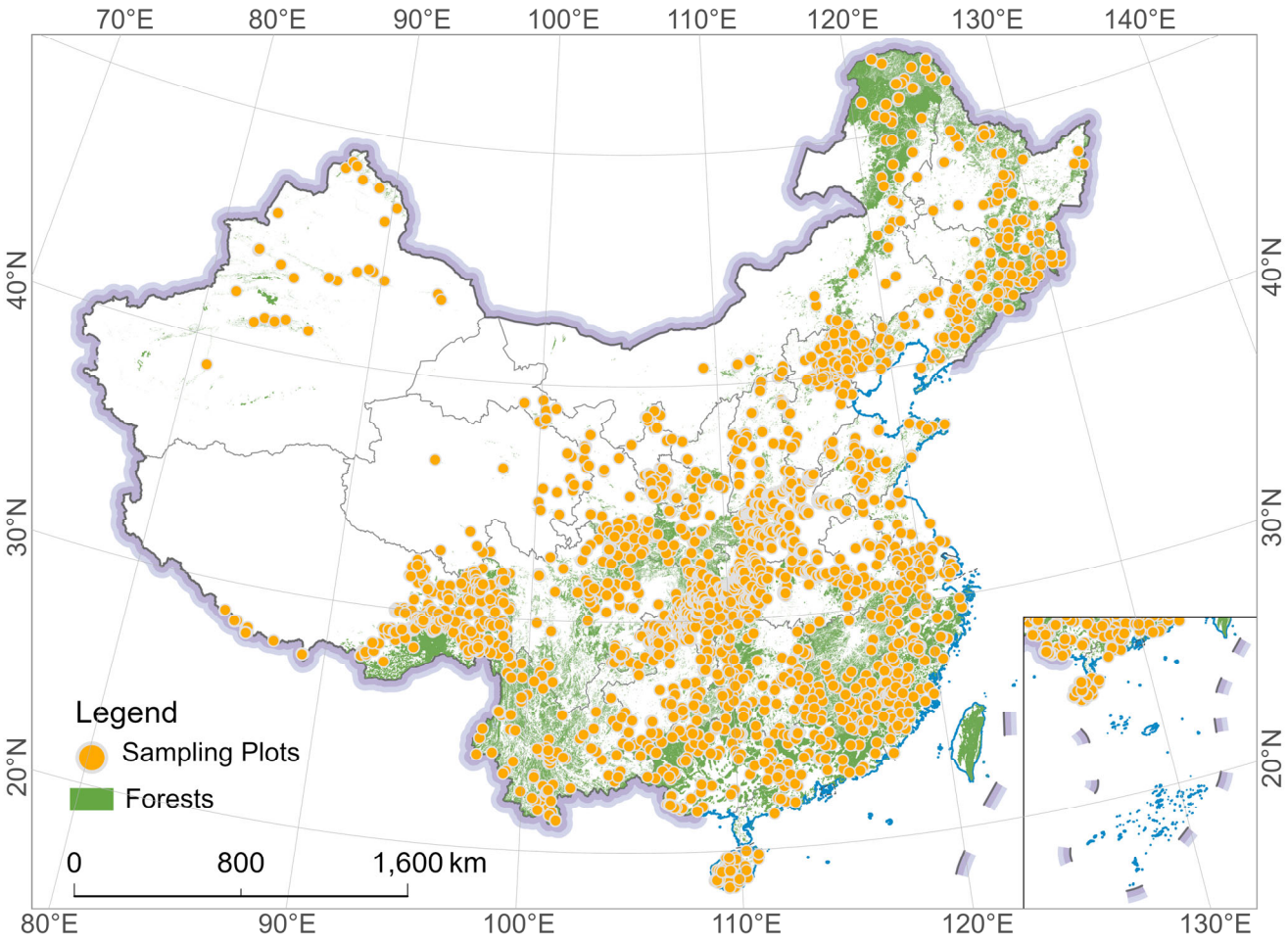


Figure 2. Spatial distribution of soil sampling plots and forest coverage.

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, 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 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 and Fig.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). Organism indicators were sourced from MODIS products, including NDVI, 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 five 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 hierarchical strategy ensured effective dimensionality reduction while maintaining predictive performance across all soil

depths. The framework was applied across five distinct soil horizons, enabled depth-specific dimensionality reduction, retaining 7 to 16 covariates per soil layer while eliminating 60.98% to 77.93% of the initial predictor set (Table S8).

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:

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

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

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

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

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 and Interpretability

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.

To enhance the interpretability of our predictive models and quantify the contribution of each environmental covariate to the predictions of soil BD and pH across the five soil layers, we utilized SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017), a game-theoretic approach for calculating Shapley values that fairly distributes the prediction's outcome among each feature, quantifying its individual contribution. SHAP provides a unified framework for evaluating feature importance and the directionality of each feature's effect on model predictions. This allows us to determine not only how each environmental factor influences the model's predictions for soil BD and pH but also revealing the direction (positive or negative) and magnitude of the feature's impact on that specific prediction. SHAP values were computed using TreeSHAP based on the training dataset.

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:

$$PIR = \frac{q_{0.95} - q_{0.05}}{q_{0.50}} \quad (6)$$

where $q_{0.95}$ and $q_{0.05}$ represent the upper and lower bounds of the 90% prediction interval, respectively, and $q_{0.50}$ 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

As shown in Fig. 3 and Table S6, the dataset comprises 4,356 forest soil profiles distributed nationwide. Soil properties were standardized to fixed depth intervals (0–5, 5–15, 15–30, 30–60, and 60–100 cm) using the equal-area spline method, resulting in 15,845 horizons for BD and 15,978 horizons for pH, which were visualized using violin plots to illustrate their distribution across soil depths. Differences among soil depth intervals were assessed using the Kruskal–Wallis test, followed by Dunn's post hoc pairwise comparisons with Bonferroni correction. Statistical significance was determined at $p < 0.05$ based on adjusted p-values.

Forest soil BD exhibited a non-linear vertical distribution pattern, characterized by an increase from surface soils to mid-depth layers followed by a decrease in deeper layers (Fig. 3a; Table S6). Specifically, mean BD values were similar in the two surface layers (1.206 g/cm³ at 0–5 cm and 1.209 g/cm³ at 5–15 cm), increased in the 15–30 cm layer (1.287 g/cm³), and reached the maximum value in the 30–60 cm layer (1.340 g/cm³). In the deepest layer (60–100 cm), mean BD declined to 1.242 g/cm³. The violin plots revealed a relatively compact distribution of BD across all soil layers, with no pronounced extreme outliers

(Fig. 3a). The left and right contours of the violin plots were well balanced, corresponding to low skewness values (0.16–0.42), indicating nearly symmetrical distributions. BD differed significantly among soil depth intervals (Fig. 3a), with statistically significant differences observed between layers ($p < 0.05$).

Forest soil pH exhibited a monotonic vertical distribution pattern, with values increasing progressively with soil depth (Fig. 3b; Table S6). Mean pH increased from 6.066 in the 0–5 cm layer to 6.131 at 5–15 cm and 6.172 at 15–30 cm, followed by a more pronounced increase in deeper layers, reaching 6.458 at 30–60 cm and 6.466 at 60–100 cm. The violin plots for the upper layers (0–5 cm, 5–15 cm, and 15–30 cm) showed more extended contours, whereas the deeper layers (30–60 cm and 60–100 cm) exhibited wider upper contours, indicating a higher density of pH values toward the higher end of the scale (Fig. 3b). The violin plot profiles for pH were nearly symmetric, with low skewness values (0.05–0.19), indicating a distribution close to normal. Pairwise comparison results revealed that soil pH formed two distinct depth-related levels, with no significant differences within the 0–30 cm and 30–100 cm intervals, but significant differences between these two groups (Fig. 3b).

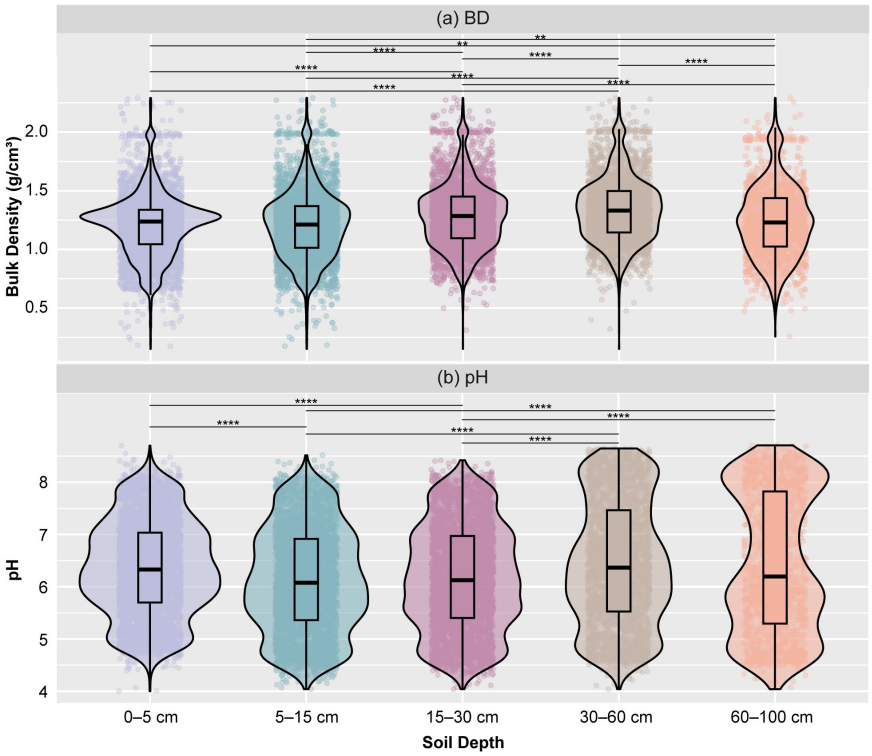


Figure 3. Variation of bulk density (BD) and pH across soil depths. Violin plots illustrate the distribution of observations, with boxplots showing medians and interquartile ranges. Horizontal bars indicate significant pairwise differences among soil depths (Dunn's test with Bonferroni correction; *, *, *** denote $p < 0.05$, 0.01, and 0.001, respectively).

3.2 Model performance

Table 1 presents the evaluation results of model predictive performance based on 10-fold cross-validation (CV) and independent validation (IV). Overall, model performance varied between soil properties and validation strategies, yet consistently indicated reliable predictive capability across all depth intervals.

For soil BD, the MEC values obtained from 10-fold CV range from 0.782 to 0.889, with RMSE values between 0.079 and 0.090 g/cm³, while the mean error (ME) remains close to zero. Under the independent validation scheme, MEC values are slightly lower (0.598–0.657), accompanied by higher RMSE values (0.155–0.181 g/cm³); however, the predictive performance remains acceptable, and ME values are likewise close to zero, indicating the absence of systematic bias.

Model performance for soil pH is consistently higher. This result is consistent with previous studies indicating pH as a more stable and predictable property at regional scales (Abdullah et al., 2025). The MEC values derived from 10-fold CV range from 0.834 to 0.868, with RMSE values of 0.214–0.256, whereas the independent validation yields MEC values of 0.705–0.812 and RMSE values of 0.432–0.515. Across all soil depths, ME values are consistently close to zero, further confirming the unbiased nature of the predictions. The comparable performance between the CV and IV results suggests good model generalizability across soil depths.

Table 1. Predictive performance of bulk density (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

3.3 Spatial patterns

We applied the QRF model to generate spatially explicit maps of soil BD and pH across five depth intervals in forested regions of China. To characterize their spatial patterns, we calculated depth-specific mean values of BD and pH along both latitudinal and longitudinal gradients. In addition, the prediction results were statistically summarized and visualized using boxplots to illustrate regional differences (Fig. 4 and Fig. 5).

3.3.1 Spatial patterns of BD

As shown in Fig. 4, the predicted spatial distribution of BD closely matches the statistical characteristics of the observations, indicating that the model effectively captures the underlying spatial patterns of soil BD. The mapping results show that the mean BD values across soil layers range from 1.16 to 1.34 g/cm³, with standard deviations between 0.15 and 0.21 g/cm³. Along the vertical profile, forest soil BD exhibits an increasing trend with depth followed by stabilization: values increase from 1.16 g/cm³ in the surface layer (0–5 cm) to a maximum of 1.34 g/cm³ at 30–60 cm, before slightly decreasing to 1.25 g/cm³ in the 60–100 cm layer. This vertical pattern is consistent with that observed in the field measurements, further supporting the model's ability to reproduce the vertical gradient of BD.

As shown in Fig. 4A, forest soil BD in China exhibits a clear decreasing trend along the latitudinal gradient and a unimodal pattern along the longitudinal gradient. Specifically, mean BD values decrease systematically from south to north with increasing latitude across all soil layers. Along the longitudinal gradient, BD follows a unimodal pattern, increasing from west to east, peaking at approximately 105°E, and subsequently decreasing. This peak approximately corresponds to the transition zone between China's second and first topographical steps, suggesting a potential influence of terrain–climate interactions on soil BD. At the pixel scale, the simulations reveal pronounced regional heterogeneity in forest soil BD. Together, the latitudinal and longitudinal gradients define a broad spatial pattern characterized by higher BD values in southwestern regions and lower values in northeastern regions (Fig. 4). Southwestern China consistently exhibits higher BD values across all soil layers (mean of 2.03 g/cm³), whereas northeastern regions show comparatively lower values (mean of 1.36 g/cm³). Notably, the magnitude of vertical BD variation differs substantially among regions. Specifically, northeastern regions show the largest increase in BD from surface to deeper layers (from 1.16 to 1.34 g/cm³), whereas northwestern regions exhibit relatively stable BD values across depths (from 1.95 to 1.96 g/cm³). Overall, the model successfully reproduces both the spatial and vertical variations of forest soil BD, offering a comprehensive depiction of its distribution across China.

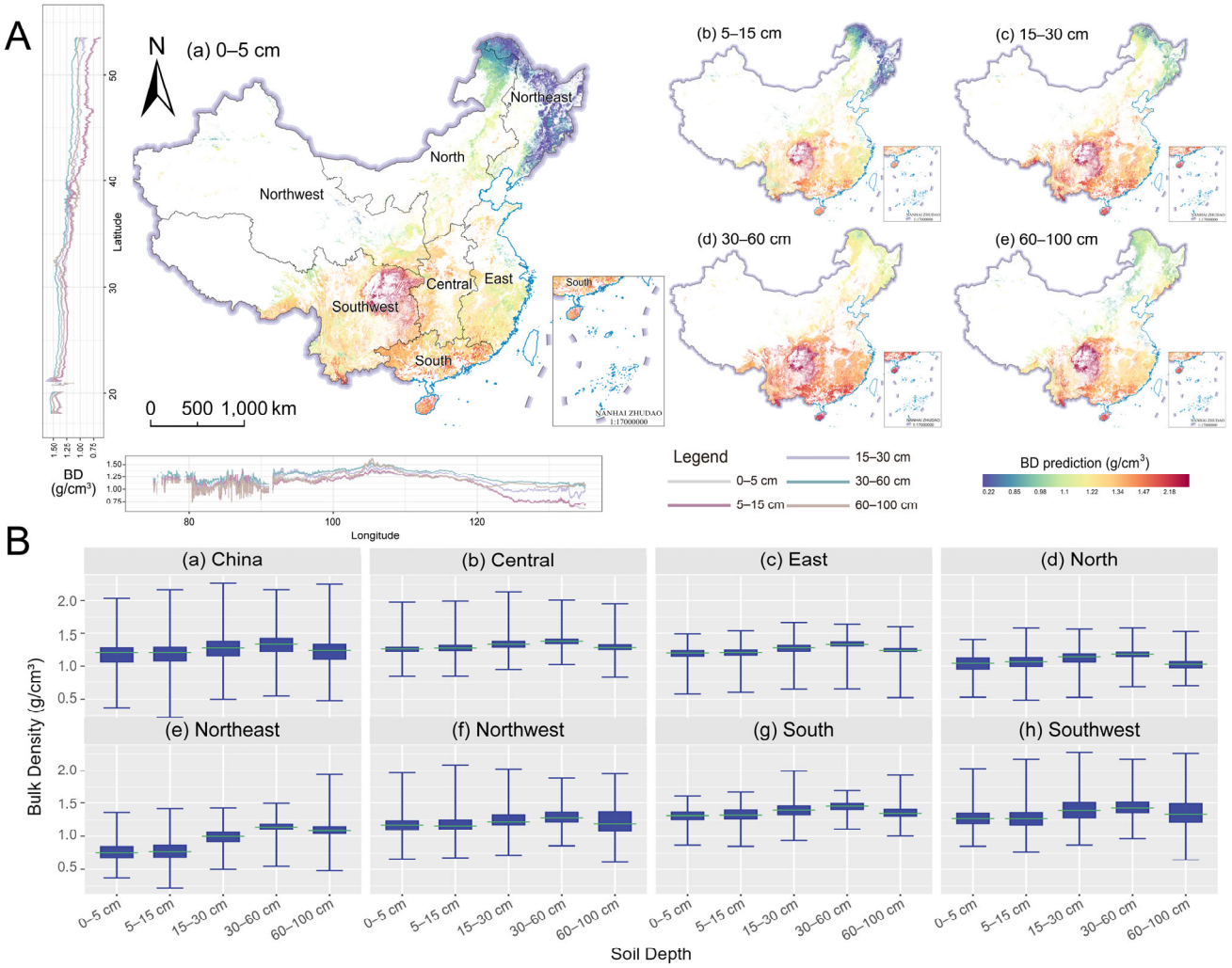


Figure 4. Spatially explicit maps of predicted bulk density (BD) at depths of 0–5 cm (a), 5–15 cm (b), 15–30 cm (c), 30–60 cm (d), and 60–100 cm (e), and their differences across China's seven geographic regions. Subfigure A illustrates the spatial distribution (prediction maps) of forest soil BD across the five soil layers, together with the latitudinal and longitudinal profile-averaged BD values for each layer. Subfigure B presents boxplots of BD for China as a whole and for each of the seven geographic regions under different soil layers. Publisher's remark: please note that the above figure contains disputed territories.

3.5.2 Spatial patterns of pH

We next examine the spatial patterns of soil pH, which are also well reproduced by the model (Fig. 5C, D). The resulting forest soil pH maps show mean values ranging from 5.70 to 6.06 across all soil layers, with standard deviations between 0.65 and 0.81 (Fig. 5A), indicating moderate spatial variability at the national scale. Specifically, soil pH exhibited a generally increasing trend with depth. Mean pH values changed from 5.99 at 0–5 cm to 5.70 at 15–30 cm, increased to 6.06 at 30–60 cm, and then slightly decreased to 6.01 at 60–100 cm. The pattern of pH variation across soil layers observed in the prediction map is consistent with the statistical results from field data, indicating the model's strong performance.

As shown in Fig. 5B, the spatial distribution of soil pH in Chinese forests exhibits a distinct latitudinal unimodal pattern, combined with an overall decreasing trend from west to east and localized sharp declines. Specifically, along the latitudinal gradient, soil pH first increases and then decreases with latitude, peaking at approximately 39°N, which corresponds to the transition between the warm temperate zone and semi-arid regions of China. Along the longitudinal gradient, pH values generally decrease with increasing longitude, with a pronounced decline around 77°E, which may reflect the influence of the major topographic barrier associated with the eastern Tibetan Plateau on soil processes. As a result, forest soil pH exhibits an overall increasing gradient from southwestern China toward northern regions. Northwestern China exhibits the highest pH values (mean range from 6.75 to 7.31), whereas the lowest values occur in southern regions (mean range from 5.01 to 5.35). Regional differences were also evident in the magnitude of depth-related pH variation. Northwestern China exhibits the largest vertical variation in mean pH across soil depths, with a range of 0.14, whereas the smallest variation is observed in northeastern regions, with a range of only 0.01. In summary, the predicted maps successfully reproduce the pronounced spatial patterns of forest soil pH across China, characterized by higher values in the northwest and lower values in southern regions.

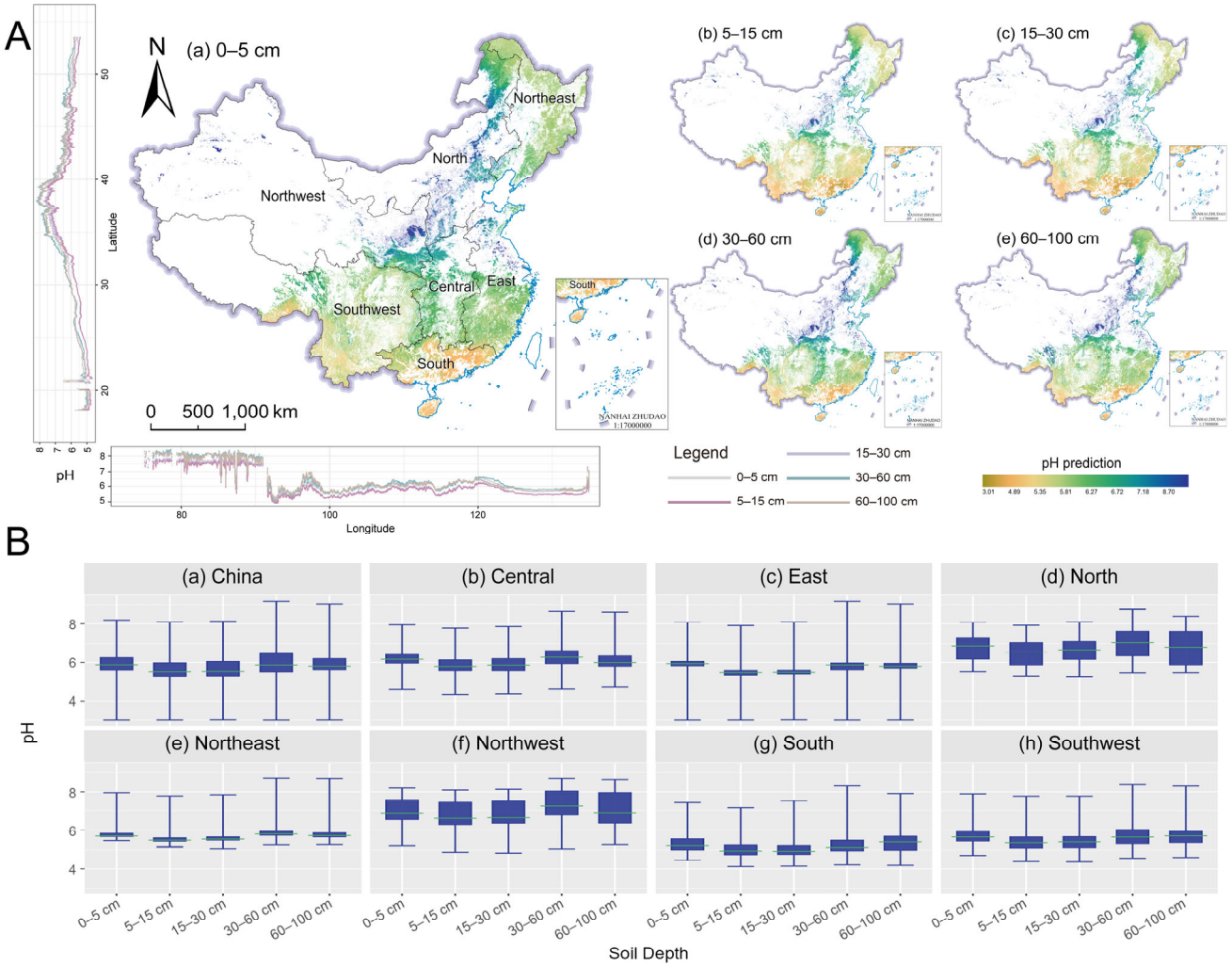


Figure 5. Spatially explicit maps of predicted soil pH at depths of 0–5 cm (a), 5–15 cm (b), 15–30 cm (c), 30–60 cm (d), and 60–100 cm (e), and their differences across China's seven geographic regions. Subfigure A shows the spatial distribution (prediction maps) of forest soil pH across the five soil layers, together with the latitudinal and longitudinal profile-averaged pH values for each layer. Subfigure B presents boxplots of soil pH for China as a whole and for each of the seven geographic regions across different soil layers. Publisher's remark: please note that the above figure contains disputed territories.

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. S3). 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. S4). 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 (**错误!未找到引用源。**). 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.

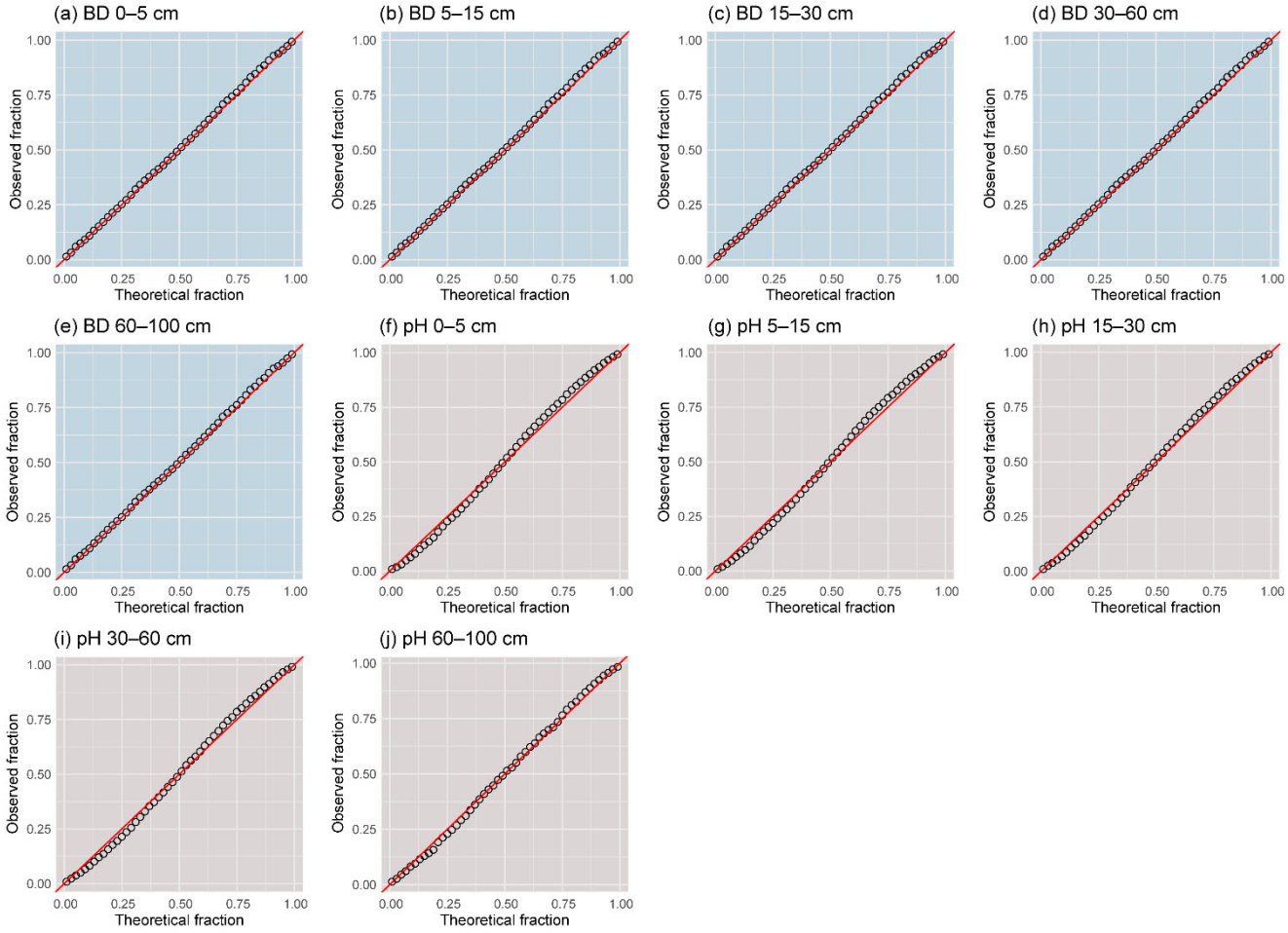


Figure 6. Validation of uncertainty quantifications. Prediction interval coverage probability (PICP) plots derived from model predictions for bulk density (BD; a–e) and soil pH (g–j).

3.5 SHAP Analysis of Predictive Factors

To quantify and visualize the contribution of individual environmental covariates to the predictions of forest soil BD and pH across the five soil layers, we applied SHAP (Shapley Additive Explanations) analysis. This method provides a unified and interpretable framework for assessing both the relative importance of predictors and the direction of their effects on model outputs. For each target variable, two primary types of visualizations were generated. First, mean absolute SHAP value plots were used to rank all covariates according to their average contribution to model predictions across all samples. Second, SHAP summary plots were employed to depict the distribution of SHAP values for each covariate, thereby illustrating the direction (positive or negative) and potential non-linear nature of their influences. Overall, climate-related covariates accounted for a substantial proportion of the total relative importance among the different categories of environmental factors, although the relative importance of individual covariates differed between forest soil BD and pH.

3.5.1 SHAP Analysis for BD

SHAP analysis reveals that the spatial variation of soil BD in Chinese forests is predominantly explained in the model by climate-related factors (Fig. 7). Climate-related factors collectively contribute approximately 71.09–77.80% of the total mean absolute SHAP attribution across soil layers. Specifically, PRE_A, PET_A, and SR_GS consistently rank as the top three influential factors across all layers, with their combined mean absolute SHAP values ranging from 0.55 to 0.67 and representing the majority of the model-attributed importance (approximately 63.72%–72.82%). Other contributing factors include Elevation (mean absolute SHAP value 0.08–0.12), suggesting that variables related to evaporation–precipitation balance and terrain play an important role in the model-predicted variation of forest soil BD. This finding aligns with previous studies on the factors affecting BD (Liu et al., 2022a).

The dominance of climate-driven patterns varies with soil depth. In the surface layer (0–30 cm), PET_A shows the highest model-attributed importance, whereas at deeper layers (>30 cm), PRE_A becomes the most influential predictor in the model. At the 0–30 cm depth, PET_A exhibits the largest mean absolute SHAP values (0.21–0.25), with an importance approximately 1.31 times that of PRE_A (0.15–0.21). At the 30–100 cm depth, the mean absolute SHAP values of PRE_A (0.20–0.42) increase substantially, reaching up to 2.58 times those of PET_A (0.06–0.18), and representing a large share of the model-attributed importance in the 60–100 cm layer (approximately 23.94%–45.45%). In comparison, the SHAP value of SR_GS consistently ranks among the top three factors, with values ranging from 0.17 to 0.19. This suggests that the relative importance of individual climate-related covariates differs across soil depths, with surface layers showing stronger attribution to PET_A and deeper layers to PRE_A.

The SHAP summary plot further clarifies the relationships between BD and key covariates. Both PET_A and PRE_A show substantial SHAP contributions to BD predictions across all soil layers. Higher values of these factors are generally associated with more positive SHAP contributions, indicating that higher PET_A and PRE_A tend to increase the model-predicted BD. Notably, the absolute magnitude of PET_A's SHAP contribution is lower in deeper soil layers (from 0.25 in

surface soil to 0.06 in deeper layers), whereas the absolute magnitude of PRE_A's negative SHAP contribution is higher in deeper soil layers (from 0.15 to 0.42). Potential evapotranspiration influences surface soil BD through seasonal moisture fluctuations, whereas, in deeper soils, precipitation-related processes such as leaching may contribute to higher BD (Xu et al., 2022; Liu et al., 2024a), which is consistent with the observed SHAP contribution patterns. In contrast, SR_GS exhibits predominantly negative SHAP contributions, suggesting that increased growing-season solar radiation is associated with lower predicted BD. The magnitude of SR_GS's negative SHAP contribution remains relatively stable across soil layers (0.15–0.19). To summarize, the dominance of climate factors in influencing BD varies notably across different soil depths, with surface soil BD being more strongly influenced by growing season potential evapotranspiration, while deeper soil BD is predominantly associated with precipitation-related predictors within the model framework.

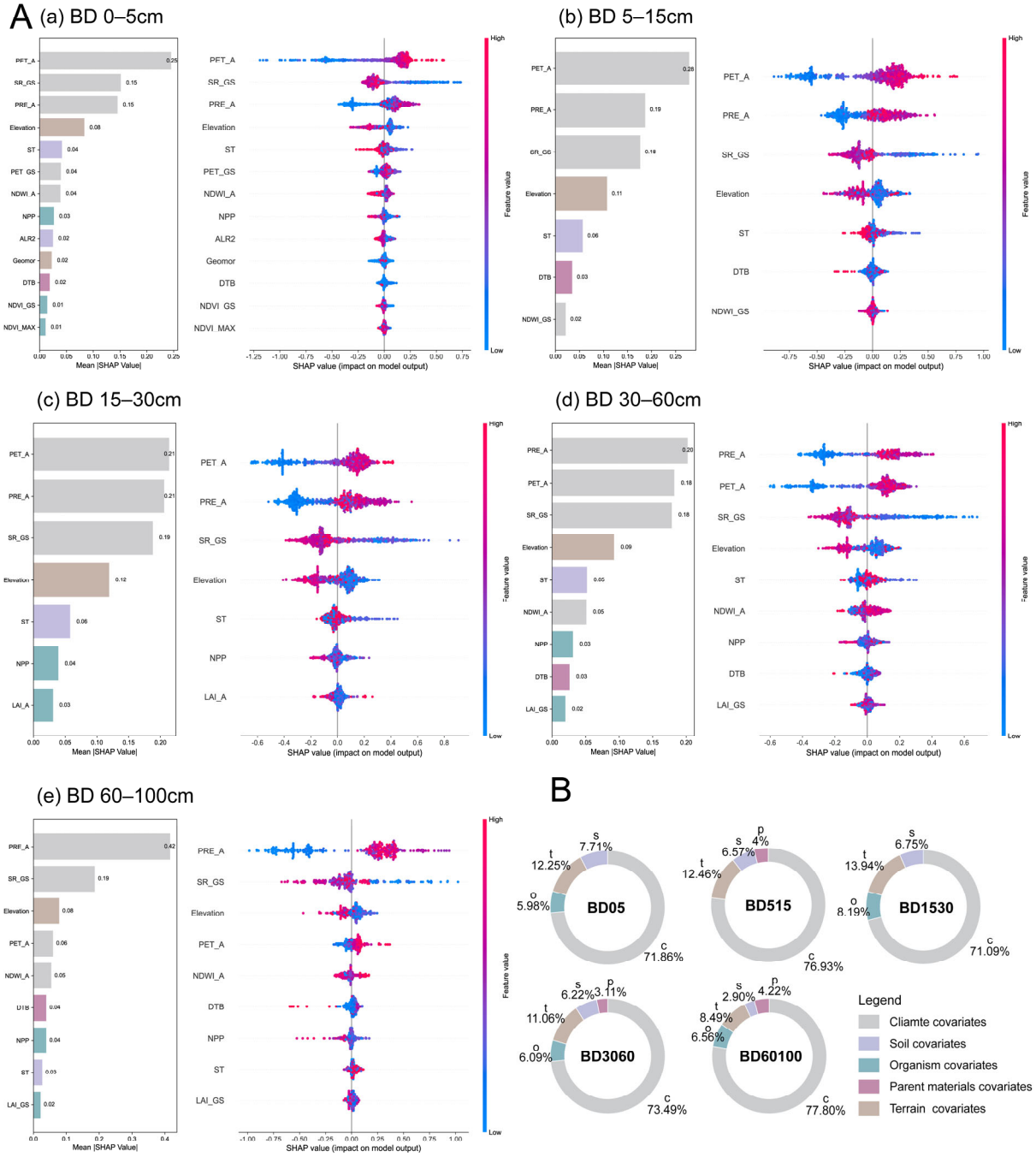


Figure 7. Interpreting the effects of driving factors on forest soil bulk density (BD) across different soil depths: SHAP-based feature importance and covariate contribution patterns. Subfigure A shows SHAP-based analysis of feature importance for soil BD driving factors. Subfigure B shows quantification of contribution rates of covariate categories to soil BD. Variable importance for model training at different soil depths. The abbreviations of the predictors are defined in Table S1.

3.5.2 SHAP Analysis for pH

As shown in Fig. 8, the spatial variation of soil pH in Chinese forests is predominantly explained by climate-related factors in the model. Climate-related factors collectively account for approximately 46.06% to 66.25% of the total mean absolute SHAP attribution across soil layers (Fig. 8B). Specifically, PRE_A exhibits the highest mean absolute SHAP values for pH, with values ranging from 0.18 to 0.52 across soil layers and representing the largest share of model-attributed importance (approximately 17.40%–46.32%). SR_GS ranks second, with mean absolute SHAP values ranging from 0.04 to 0.12, contributing a moderate proportion of the model-attributed importance (approximately 4.18%–11.89%). This result is consistent with previous studies, which have identified precipitation as the dominant factor controlling soil pH at large spatial scales (Huang et al., 2025). Less important but still informative covariates include Elevation, suggesting that climate-related variables dominate the model-attributed spatial patterns of pH at the national scale, while terrain-related effects may play a greater role locally.

The SHAP values of factors influencing forest soil pH vary considerably with soil depth. In the surface layer (0–60 cm), PRE_A exhibits the highest model-attributed importance, whereas in deeper layers (60–100 cm), NDWI_A becomes the leading predictor. At the 0-60 cm depth, PRE_A is the most important factor (SHAP values 0.32-0.52), with its importance being 2.91 times that of NDWI_A (SHAP values 0.05-0.16). At the 60–100 cm depth, the mean absolute SHAP value of PRE_A decreases to 0.04, whereas NDWI_A becomes the most influential predictor (SHAP value 0.20), representing the largest share of model-attributed importance in this layer (19.51%). Additionally, SR_GS ranks among the top three influential factors for the 0-60 cm layer (SHAP values 0.09-0.12), but its importance considerably decreases in the 60-100 cm layer (SHAP value 0.04). Overall, climate-related covariates tend to exhibit stronger model-attributed importance in shallower soil layers.

As illustrated in Fig. 8, the relationship between pH and key covariates shows considerable variation with soil depth. Both PRE_A and SR_GS show substantial SHAP contributions to soil pH predictions across all soil layers. Higher values of these factors are generally associated with negative SHAP contributions, indicating that increased PRE_A and SR_GS tend to decrease the model-predicted soil pH. Specifically, the absolute magnitude of the negative SHAP contributions of PRE_A and SR_GS is lower in deeper soil layers, from 0.34 to 0.18 for PRE_A and from 0.12 to 0.04 for SR_GS. Forest surface soil pH tends to decrease with increased precipitation, which may be attributed to enhanced leaching of base cations under wetter conditions. Increased rainfall accelerates the downward transport and loss of buffering cations, thereby potentially weakening the soil's acid-neutralizing capacity and promoting soil acidification (Huang et al., 2022c). Similarly, NDWI_A shows a negative influence, with its impact varying considerably across soil layers. At 0–30 cm, the influence is stronger (SHAP values 0.14–0.15), but it decreases between 15–30 cm (SHAP value 0.06) and increases again to the highest level at 30–100 cm (SHAP values 0.16–0.20). NDWI_A has been shown to correlate with soil water content. A possible explanation for this pattern is that NDWI_A may capture hydrological variability that is associated with the redistribution of soluble ions, which in turn may influence soil pH at different depths (Serrano et al., 2019). In general, the findings emphasize the depth-dependent

effects of climate factors on forest soil pH, with precipitation emerging as a key predictor associated with soil acidification patterns in the model.

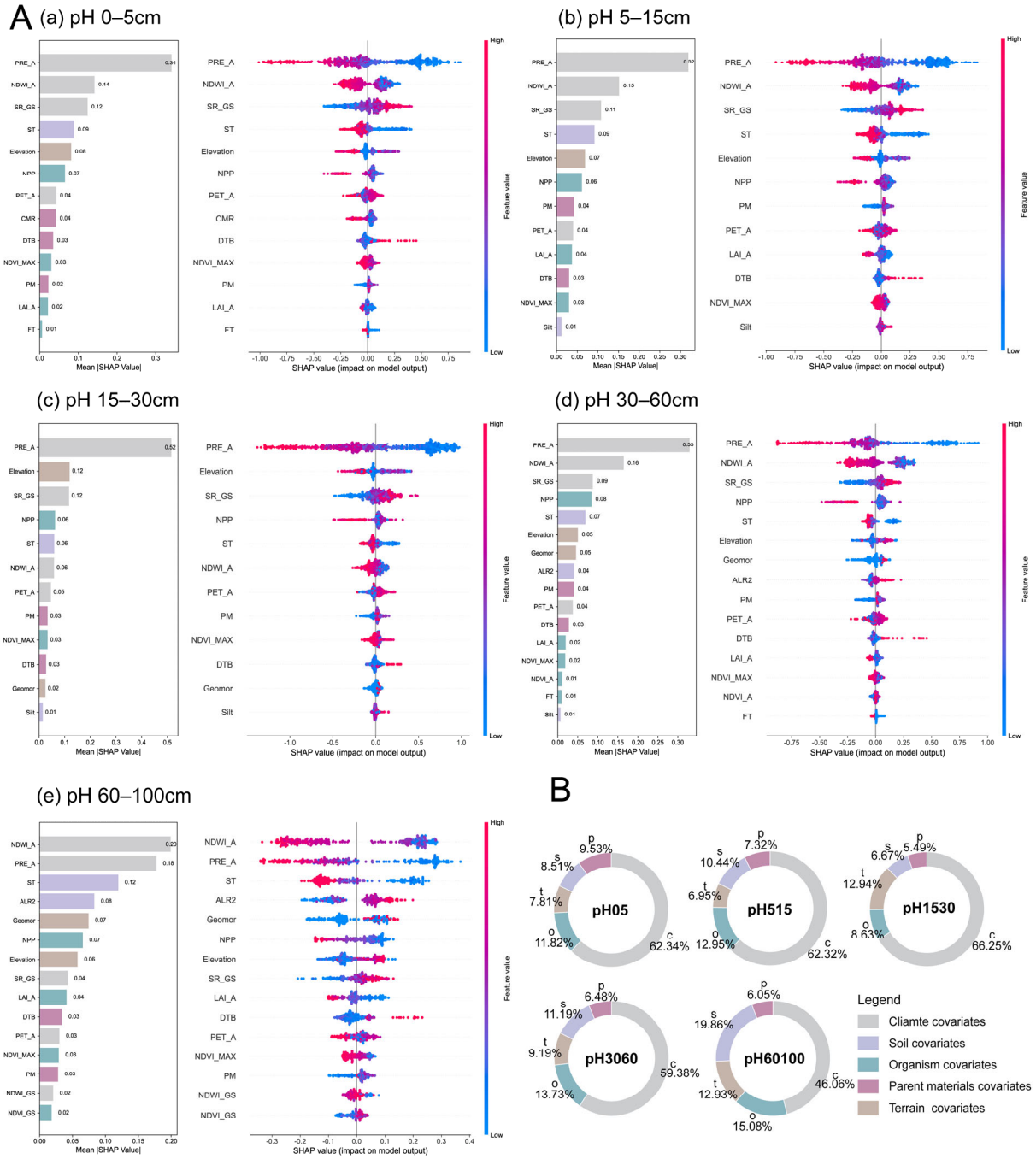


Figure 8. Interpreting the effects of driving factors on forest soil pH across different soil depths: SHAP-based feature importance and covariate contribution patterns. Subfigure A shows SHAP-based analysis of feature importance for soil pH driving factors.

Subfigure B shows quantification of contribution rates of covariate categories to soil pH. Variable importance for model training at different soil depths. The abbreviations of the predictors are defined in Table S1.

4 Discussion

4.1 Comparison with previous products

While national datasets such as CSDLv2 (Shi et al., 2025) and ChinaSoilInfoGrids (Liu et al., 2022a) offer extensive insights into soil properties, and global products like SoilGrids 2.0 (Hengl et al., 2017) provide valuable worldwide references, their utility in forest ecosystem studies remains constrained. To address this data gap, our study developed a comprehensive nationwide forest soil dataset for China and integrated it with the FRFS-QRF model to simulate the spatial distribution of BD and pH in Chinese forests. By comparing our results with existing national and global products, we aim to quantify the potential enhancements achievable through the use of a forest ecosystem data foundation and the FRFS-QRF methodology we propose for DSM in forest ecosystems.

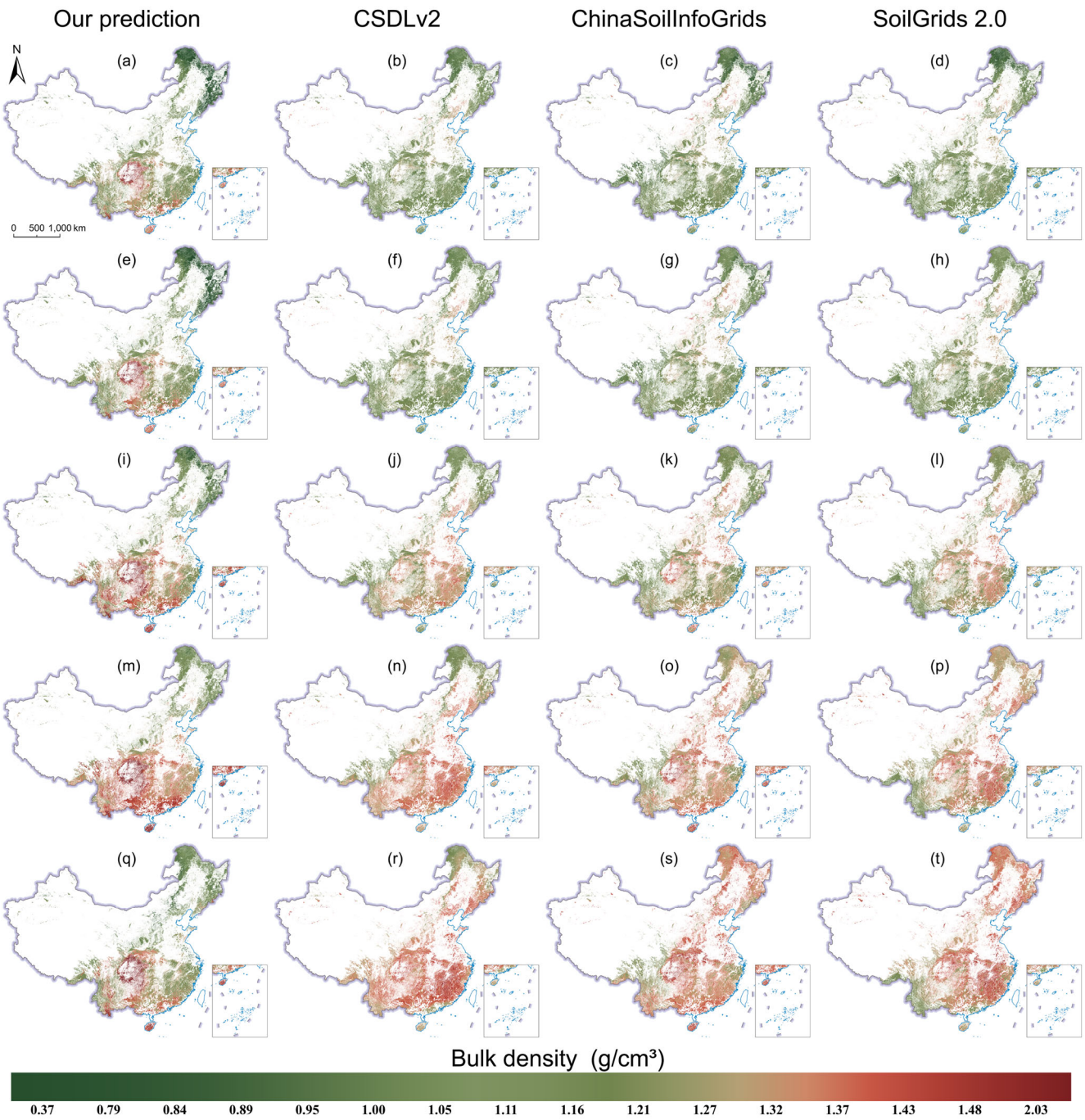
The proposed model exhibits enhanced predictive accuracy and reliability for forest soil BD and pH when compared to existing datasets, as illustrated in Table 2. During 10-fold CV, the model achieved reductions in RMSE of 27.41%, 37.26%, and 49.46% relative to CSDLv2, ChinaSoilInfoGrids, and SoilGrids 2.0, respectively. For pH, RMSE reductions were 65.98%, 68.43%, and 73.69% against the same benchmarks. These improvements were sustained under IV, with RMSE reductions ranging from 41.4% to 47.6% for BD and from 28.5% to 32.9% for pH. The enhanced performance of the proposed model can be attributed to two main factors. First, this study specifically targets forest ecosystems, thereby reducing environmental and sampling heterogeneity in comparison to existing datasets that integrate multiple ecosystem types. This aligns with prior research, such as studies on forest soil pH in warm temperate zones and on soil organic carbon mapping in agricultural systems, which report improved accuracy within specific ecosystem contexts. Second, independently selecting the optimal covariate set for different soil layers improved model performance, as soil formation processes exhibit complexity along the vertical dimension.

Table 2 Predictive performance of soil property predictions, CSDLv2, ChinaSoilInfoGrids and SoilGrids 2.0.

Validation	Depth (cm)	Our predictions			CSDLv2			ChinaSoilInfoGrids			SoilGrids 2.0		
		MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME
	BD												
	0–5	0.782	0.090	0.000	0.620	0.120	0.000	0.483	0.147	-0.002	0.268	0.185	0.025
	5–15	0.815	0.084	0.000	0.630	0.110	0.000	0.479	0.138	-0.002	0.279	0.172	0.053
	15–30	0.828	0.081	-0.000	0.600	0.110	-0.000	0.457	0.132	-0.002	0.263	0.16	0.043
10-fold	30–60	0.874	0.079	-0.000	0.550	0.120	-0.000	0.403	0.128	-0.002	0.193	0.158	0.050
CV	60–100	0.889	0.087	0.000	0.570	0.120	-0.000	0.303	0.126	-0.003	0.098	0.158	0.065
	pH												
	0–5	0.844	0.215	-0.000	0.690	0.700	0.000	0.711	0.791	0.001	0.635	0.916	-0.038
	5–15	0.834	0.254	0.000	0.700	0.680	0.000	0.724	0.724	0.003	0.652	0.893	-0.066
	15–30	0.854	0.214	0.000	0.700	0.680	0.000	0.74	0.74	0.002	0.659	0.878	-0.141

	30–60	0.854	0.256	0.001	0.680	0.700	-0.000	0.737	0.737	0.004	0.647	0.898	-0.194
	60–100	0.868	0.238	0.001	0.680	0.700	0.000	0.736	0.736	0.006	0.641	0.888	-0.184
IV	**BD**												
	0–5	0.598	0.164	-0.010	0.021	0.309	0.036	0.080	0.293	0.020	0.076	0.317	0.011
	5–15	0.611	0.181	-0.017	0.061	0.276	0.001	0.119	0.268	0.016	0.121	0.302	0.000
	15–30	0.657	0.155	0.006	0.108	0.258	0.007	0.172	0.249	0.013	0.167	0.295	0.020
	30–60	0.614	0.166	0.005	0.131	0.338	-0.126	0.013	0.316	-0.111	0.261	0.357	-0.105
	60–100	0.656	0.166	-0.019	0.021	0.309	0.036	0.080	0.293	0.020	0.076	0.317	0.011
	pH												
	0–5	0.705	0.432	-0.003	0.660	0.637	-0.081	0.636	0.659	0.104	0.693	0.606	0.008
	5–15	0.726	0.480	-0.001	0.618	0.733	-0.333	0.676	0.675	-0.233	0.681	0.669	-0.242
	15–30	0.742	0.448	-0.007	0.697	0.740	-0.073	0.728	0.701	0.041	0.711	0.723	0.077
	30–60	0.760	0.515	-0.002	0.653	0.778	-0.163	0.723	0.695	-0.118	0.715	0.705	-0.052
	60–100	0.812	0.492	0.014	0.660	0.637	-0.081	0.636	0.659	0.104	0.693	0.606	0.008

Our results indicate a non-linear vertical pattern of forest soil BD, with BD peaking in the 30–60 cm layer and declining in the 60–100 cm layer (Fig. 9). This profile differs from the commonly reported monotonic increase in BD with depth in many ecosystems, including forest systems where BD generally increases with soil depth (Shen et al., 2022). However, non-monotonic vertical BD distributions have also been documented in forest soils, particularly in relation to long-term stand development and soil profile differentiation (Tang et al., 2025), suggesting that forest soil BD may not universally follow a simple monotonic depth trend. This depth profile is plausibly linked to clay illuviation and deep-root processes. Downward translocation and mid-profile accumulation of fine clay can promote denser packing and reduce pore space, leading to higher BD in intermediate layers, as commonly observed in soils affected by argilluviation (Jenny, 1941; Brady and Weil, 2016). In deeper horizons, lower clay contents together with more pronounced root activity may enhance biopore formation and maintain a looser structure, thereby contributing to lower BD (Jobbágy and Jackson, 2000; Angers et al., 2011). Comparison with existing datasets further suggests that currently available products do not fully represent this complex depth-dependent variation in forest BD. In particular, for the 60–100 cm layer, our model predicts a mean BD of 1.25 g/cm³, lower than the 1.37–1.38 g/cm³ reported by existing datasets (Fig. S5). This discrepancy implies that extrapolations based solely on those datasets may overestimate forest soil organic carbon stocks at this depth.



490

Figure 9. Bulk density prediction (g/cm³) maps: rows show five soil depths (0–5, 5–15, 15–30, 30–60, and 60–100 cm) and columns show four datasets (Our prediction, CSDLv2, ChinaSoilInfoGrids, and SoilGrids2.0). Publisher's remark: please note that the above figure contains disputed territories.

Forest soil pH profiles are primarily governed by chemical processes and generally exhibit an increasing trend with depth. Model comparison results indicate that forest soil pH in China displays a consistent spatial pattern characterized by lower values in southern regions and higher values in northern regions, while vertically increasing with soil depth (Yu et al., 2020). This depth-dependent pattern is in line with the commonly observed pH profiles of forest soils (Fig. 10). However, spatial statistical analyses reveal that, across all soil layers, the mean forest soil pH predicted in this study (5.70–6.01) is consistently lower than the corresponding values reported by existing datasets (5.93–6.23) (Fig. S6). This systematic difference likely reflects ecosystem-specific acidification processes in forest soils. Long-term leaching under relatively high precipitation regimes promotes the removal of base cations, while the continuous accumulation of acidic inputs derived from organic matter decomposition and root exudation further contributes to soil acidification (Farooq et al., 2022a; Abdullah et al., 2025). As a result, forest soils typically exhibit lower pH values than soils in other ecosystems. These findings suggest that existing datasets may underestimate the degree of soil acidification in Chinese forests, particularly when generalized across ecosystems or derived from limited forest-specific observations.

Overall, our results indicate that the use of forest-specific datasets and modeling strategies can substantially improve the representation of soil BD and pH in forest ecosystems. The discrepancies identified relative to existing products suggest that generalized soil datasets may overestimate forest soil carbon stocks and underestimate soil acidification, underscoring the necessity of ecosystem-specific digital soil mapping for reliable soil assessments and forest management.

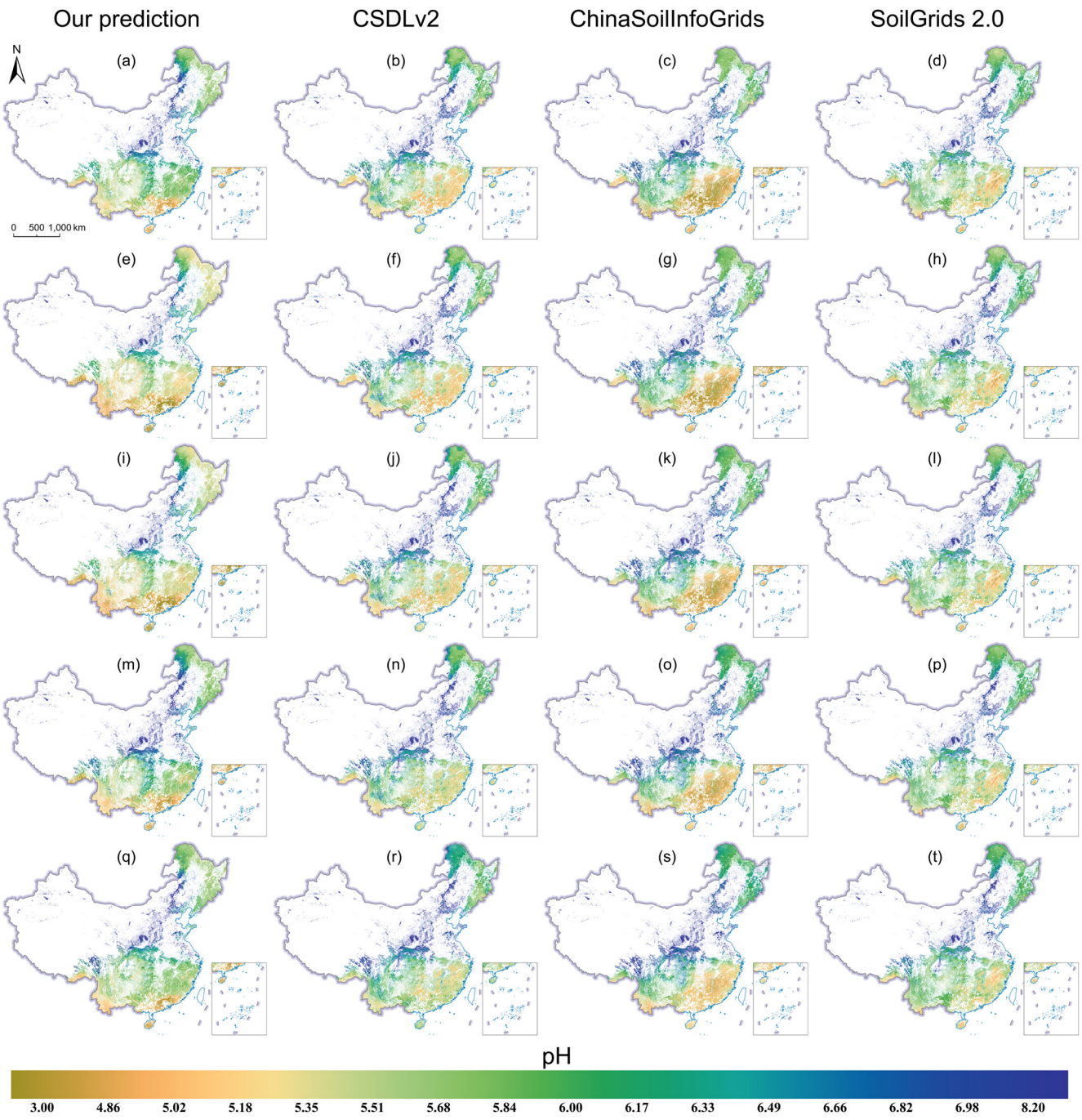


Figure 10. pH prediction maps: rows show five soil depths (0–5, 5–15, 15–30, 30–60, and 60–100 cm) and columns show four datasets (Our prediction, CSDLv2, ChinaSoilInfoGrids, and SoilGrids2.0). Publisher's remark: please note that the above figure contains disputed territories.

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., 2024b), and is significant for understanding forest soil acidification (Farooq et al., 2022b). 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; however, several methodological limitations remain and merit further investigation, particularly regarding the predictive reliability of machine learning approaches. Machine learning techniques, while significantly enhancing DSM by capturing nonlinear soil–environment relationships, are constrained by limitations in spatial coverage and feature-space representativeness (Yang et al., 2013; Chen et al., 2019). Forest ecosystems 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 may yield biased predictions in undersampled regions. Although ensemble methods can manage uncertainty in sparse data settings, they may prioritize statistical regularities over mechanistic soil formation processes (Sylvain et al., 2021; Liu et al., 2022b).

In addition to data-related limitations, uncertainty may also arise from harmonizing environmental covariates originating from heterogeneous native spatial resolutions into a common fine grid. The integration of multi-resolution covariates through resampling is a standard practice in national-scale digital soil mapping to ensure spatially consistent modeling (Shi et al., 2024; Shi et al., 2025). However, this process inherently introduces scale-related uncertainty and potential smoothing effects in the resulting predictions, particularly for covariates derived from coarser-resolution sources (Cavazzi et al., 2013; Piedallu et al., 2022). The influence of covariate resolution on prediction accuracy is a recognized challenge in DSM, as the representativeness of environmental predictors can vary substantially with spatial scale (Guo et al., 2019; Adhikari et al., 2020). Importantly, this limitation is not unique to the present study but is widely acknowledged in large-scale mapping initiatives, including established global products such as SoilGrids (Poggio et al., 2021). While predictive models can quantify uncertainty associated with sampling density and algorithmic structure, explicit quantification of uncertainty arising solely from the resampling of multi-resolution covariates remains methodologically challenging and is rarely addressed in existing studies

(Wadoux et al., 2020; Sylvain et al., 2021). Consequently, the resulting soil property maps should be interpreted as conditional estimates, representing the most probable spatial patterns given the selected covariates and their effective spatial resolution, rather than exact representations of fine-scale soil heterogeneity (Robb et al., 2025).

Finally, incomplete environmental data coverage restricted soil property prediction to areas where all covariates were consistently available. This decision reflects a fundamental trade-off in DSM between spatial completeness and predictive reliability. Omission-based approaches, such as those used in SoLIM (Nussbaum et al., 2018; Zhu et al., 2018), prioritize accuracy by excluding data-deficient regions but result in spatial gaps that limit practical applications. In contrast, imputation techniques such as 3×3 window smoothing (Padarian et al., 2019; Fan et al., 2020) preserve continuity but may introduce significant errors, particularly in ecotones where environmental gradients are steep and covariate relationships become unstable (Hengl et al., 2015). In our study, such limitations were most evident along elevational boundaries and remote sensing index thresholds, where fragmented gaps produced compounding spatial uncertainty. To mitigate these effects, predictions were restricted to the maximal common spatial domain defined by all covariates. This conservative approach was based on three considerations: (1) the statistical instability of edge extrapolations, (2) elevated error propagation risks in transitional environments (Dong et al., 2023), and (3) a practical emphasis on accuracy over exhaustive coverage. While this approach slightly reduced the total mappable forest area, it improved the overall reliability of spatial outputs by minimizing uncertainty in marginal regions (Smith, 2001; Huettmann and Gottschalk, 2011). The resulting trade-off underscores the persistent tension between spatial coverage and predictive confidence in large-scale DSM. Future advances should explore adaptive frameworks that integrate uncertainty-aware imputation with hybrid models, using localized gap geometry to guide prediction boundaries (Arrouays et al., 2014). Such innovations may help reconcile spatial completeness with predictive precision in next-generation soil mapping.

Looking forward, emerging hybrid frameworks that integrate environmental similarity metrics with pedological expertise show promise for addressing both data imbalance and scale-related challenges (Zhao et al., 2024), although their scalability and operational feasibility at national extents require 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 for mitigating dataset imbalance and capturing underlying heterogeneity (Brus et al., 2011). Concurrently, exploring novel covariates derived from multi-source remote sensing (e.g., hyperspectral, LiDAR, and 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.

5 Data and code availability

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.

7 Conclusions

Our study developed a high-resolution, forest-specific mapping of 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 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

Abdullah, H., Skidmore, A. K., Siegenthaler, A., and Neinavaz, E.: High-resolution prediction of soil pH in european temperate forests using sentinel-2 and ancillary environmental data, Sci. Rep., 15, 28509, https://doi.org/10.1038/s41598-025-03942-4, 2025.

Adhikari, K., Mishra, U., Owens, P., Libohova, Z., Wills, S., Riley, W., Hoffman, F., and Smith, D.: Importance and strength of environmental controllers of soil organic carbon changes with scale, Geoderma, 375, https://doi.org/10.1016/j.geoderma.2020.114472, 2020.

Angers, D. A., Arrouays, D., Saby, N. P. A., and Walter, C.: Estimating and mapping the carbon saturation deficit of French agricultural topsoils, Soil Use and Management, 27, 448–452, https://doi.org/10.1111/j.1475-2743.2011.00366.x, 2011.

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.

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

Brady, N. C. and Weil, R. R.: The nature and properties of soils, 15 edition, global., Pearson, Harlow, England London New York, 1104 pp., 2016.

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.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., and Fealy, R.: Are fine resolution digital elevation models always the best choice in digital soil mapping?, Geoderma, 195–196, 111–121, https://doi.org/10.1016/j.geoderma.2012.11.020, 2013.

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 soil microbial community in successive planting of chinese fir (cunninghamia lanceolata), Front. Plant Sci., 13, 954777, https://doi.org/10.3389/fpls.2022.954777, 2022.

Chen, L. F., He, Z. B., Du, J., Yang, J. J., and Zhu, X.: Patterns and environmental controls of soil organic carbon and total nitrogen in alpine ecosystems of northwestern China, CATENA, 137, 37–43, https://doi.org/10.1016/j.catena.2015.08.017, 2016.

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, 184–194, https://doi.org/10.1016/j.geoderma.2019.03.016, 2019.

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 global soil property maps for Earth system models, SOIL, 5, 137–158, https://doi.org/10.5194/soil-5-137-2019, 2019.

Dong, H., Huang, S., Wang, H., Huang, Q., Leng, G., Li, Z., Li, L., and Peng, J.: Identifying non-stationarity in the dependence structures of meteorological factors within and across seasons and exploring possible causes, Stochastic Environ. Res. Risk Assess., 37, 4071–4089, https://doi.org/10.1007/s00477-023-02496-z, 2023.

Fan, N. Q., Zhu, A. X., Qin, C. Z., and Liang, P.: Digital soil mapping over large areas with invalid environmental covariate data, ISPRS Int. J. Geo-Inf., 9, 102, https://doi.org/10.3390/ijgi9020102, 2020.

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., Chen, X., Shakoor, A., Rashid, M. H. U., Kumar, U., Alhomrani, M., Alamri, A. S., Ravindran, B., and Yan, W.: Unraveling the importance of forest structure and composition driving soil microbial and enzymatic responses in the subtropical forest soils, Forests, 13, 1535, https://doi.org/10.3390/f13101535, 2022a.

Farooq, T. H., Li, Z., Yan, W., Shakoor, A., Kumar, U., Shabbir, R., Peng, Y., Gayathiri, E., Alotaibi, S. S., Wróbel, J., Kalaji, H. M., and Chen, X.: Variations in litterfall dynamics, C:N:P stoichiometry and associated nutrient return in pure and mixed stands of camphor tree and masson pine forests, Front. Environ. Sci., 10, 903039, https://doi.org/10.3389/fenvs.2022.903039, 2022b.

Goovaerts, P.: Geostatistical modelling of uncertainty in soil science, Geoderma, 103, 3–26, https://doi.org/10.1016/S0016-7061(01)00067-2, 2001.

Guo, C., Wu, Y., Ni, J., and Guo, Y.: Forest carbon storage in guizhou province based on field measurement dataset, Acta Geochim., 38, 8–21, https://doi.org/10.1007/s11631-018-0306-3, 2019.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes De Jesus, J., Tamene, L., and Tondoh, J. E.: Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions, PLOS One, 10, e0125814, https://doi.org/10.1371/journal.pone.0125814, 2015.

Heuvelink, G. B. M., Kros, J., Reinds, G. J., and De Vries, W.: Geostatistical prediction and simulation of european soil property maps, Geoderma Reg., 7, 201–215, https://doi.org/10.1016/j.geodrs.2016.04.002, 2016.

Huang, H., Liu, Y., Liu, Y., Tong, Z., Ren, Z., and Xie, Y.: Digital mapping of soil pH and driving factor analysis based on environmental variable screening, Sustainability, 17, https://doi.org/10.3390/su17073173, 2025.

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.

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, Forest Ecology and Management, 505, 119951, https://doi.org/10.1016/j.foreco.2021.119951, 2022c.

Huettmann, F. and Gottschalk, T.: Simplicity, model fit, complexity and uncertainty in spatial prediction models applied over time: we are quite sure, aren't we?, in: Predictive Species and Habitat Modeling in Landscape Ecology, edited by: Drew, C. A., Wiersma, Y. F., and Huettmann, F., Springer New York, New York, NY, 189–208, 2011.

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

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

Jobbágy, E. G. and Jackson, R. B.: The vertical distribution of soil organic carbon and its relation to climate and vegetation, Ecol. Appl., 10, 423–436, https://doi.org/10.1890/1051-0761(2000)010%255B0423:TVDOSO%255D2.0.CO;2, 2000.

Khaledian, Y. and Miller, B.: Selecting appropriate machine learning methods for digital soil mapping, Appl. Math. Modell., 81, 401–418, https://doi.org/10.1016/j.apm.2019.12.016, 2020.

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.

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

Liu, B., Zhao, P., Wu, Y., and Wang, X.: Clarifying the effects of potential evapotranspiration and soil moisture on transpiration in secondary forests of birch in semi-arid regions of China, J. Plant Interact., 19, 2389048, https://doi.org/10.1080/17429145.2024.2389048, 2024a.

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, 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, 2024b.

Lundberg, S. and Lee, S. I.: A unified approach to interpreting model predictions, https://doi.org/10.48550/arXiv.1705.07874, 25 November 2017.

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.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, Soil, 4, 1–22, https://doi.org/10.5194/soil-4-1-2018, 2018.

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.

Piedallu, C., Pedersoli, E., Chaste, E., Morneau, F., Seynave, I., and Gégout, J. C.: Optimal resolution of soil properties maps varies according to their geographical extent and location, Geoderma, 412, 115723, https://doi.org/10.1016/j.geoderma.2022.115723, 2022.

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

Potash, E., Guan, K., Margenot, A. J., Lee, D., Boe, A., Douglass, M., Heaton, E., Jang, C., Jin, V., Li, N., Mitchell, R., Namoi, 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.

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

Robb, C., Aitkenhead, M., Coull, M., Macfarlane, F., and Matthews, K.: Soil property, carbon stock and peat extent mapping at 10 m resolution in Scotland using digital soil mapping techniques, Eur. J. Soil Sci., 76, https://doi.org/10.1111/ejss.70123, 2025.

Rodrigues, H., Ceddia, M. B., Vasques, G. M., Grunwald, S., and Babaeian, E.: AutoRA: an innovative algorithm for automatic 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.

Serrano, J., Shahidian, S., Silva, J. M. da, Serrano, J., Shahidian, S., and Silva, J. M. da: Evaluation of Normalized Difference Water Index as a Tool for Monitoring Pasture Seasonal and Inter-Annual Variability in a Mediterranean Agro-Silvo-Pastoral System, Water, 11, https://doi.org/10.3390/w11010062, 2019.

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.2022.115749, 2022.

Shen, Y., Li, J., Chen, F., Cheng, R., Xiao, W., Wu, L., and Zeng, L.: Correlations between forest soil quality and aboveground vegetation characteristics in hunan province, china, Front. Plant Sci., 13, 1009109, https://doi.org/10.3389/fpls.2022.1009109, 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, Y.: A China dataset of soil properties for land surface modelling (version 2, CSDLv2), Earth Syst. Sci. Data, 17, 517–543, https://doi.org/10.5194/essd-17-517-2025, 2025.

Smith, P. L.: Geostatistical error management: quantifying uncertainty for environmental sampling and mapping, Technometrics, 43, 238–239, https://doi.org/10.1198/tech.2001.s593, 2001.

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.

Szatmári, G., Laborczi, A., Mészáros, J., Takács, K., Benő, A., Koós, S., Bakacsi, Z., and Pásztor, L.: Gridded, temporally referenced spatial information on soil organic carbon for Hungary, Sci. Data, 11, 1312, https://doi.org/10.1038/s41597-024-04158-3, 2024.

Tang, S., Xu, X., Wu, Y., Meng, L., Tawaraya, K., and Cheng, W.: Long-term afforestation of black pine over two centuries asymptotically enhanced SOC and TN stocks in a typical coastal sand dune of Japan, CATENA, 249, 108697, https://doi.org/10.1016/j.catena.2024.108697, 2025.

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, Geoderma, 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 suggested solutions, https://doi.org/10.31223/OSF.IO/8EQ6S, 6 February 2020.

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

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 by 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 high 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.

Xu, Z., Man, X., Cai, T., and Shang, Y.: How potential evapotranspiration regulates the response of canopy transpiration to soil moisture and leaf area index of the boreal larch forest in China, Forests, 13, https://doi.org/10.3390/f13040571, 2022.

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

Yu, Z., Chen, H. Y. H., Searle, E. B., Sardans, J., Ciais, P., Peñuelas, J., and Huang, Z.: Whole soil acidification and base cation reduction across subtropical china, Geoderma, 361, 114107, https://doi.org/10.1016/j.geoderma.2019.114107, 2020.

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.

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.

Zhu, A. X., Qin, C. Z., Liang, P., and Du, F.: Digital soil mapping for smart agriculture: the solim method and software platforms, Vestn. Ross. univ. družby nar., Ser. Agron. životnovod., 13, 317–335, https://doi.org/10.22363/2312-797X-2018-13-4-317-335, 2018.

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.