

Regional patterns and environmental controls of topsoil pH in a karst mountainous landscape of Southwest China

Shihan Wang

Southwest University

Hongbin Liu

lhbin@swu.edu.cn

Southwest University

Article

Keywords: digital soil mapping, soil pH, Boruta algorithm, random forest, SHAP, karst ecosystems

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10926832/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Regional patterns and environmental controls of topsoil pH in a karst mountainous landscape of Southwest China

Shihan Wang^{1,2}, Hongbin Liu^{1,2,*}

¹ College of Resources and Environment, Southwest University, Chongqing 400716, China

² Chongqing Key Laboratory of Digital Agriculture, Chongqing 400716, China

* Corresponding author: Hongbin Liu. Email addresses: wshhh0322@163.com (S. Wang), lhbin@swu.edu.cn (H. Liu).

Abstract

Mapping soil pH in karst mountains is complicated by fragmented terrain, correlated environmental covariates, and limited model interpretability. We evaluated feature-selection strategies within a random forest framework using 4,578 topsoil samples (0–20 cm) from Qianjiang District, Southwest China. Fifty-nine environmental variables describing terrain, hydrology, climate, vegetation, lithology, and human activity expanded to 198 candidate predictors after one-hot encoding. Boruta, recursive feature elimination, and particle swarm optimization were compared with a full-feature baseline using nested fivefold random cross-validation. Boruta–random forest retained 28 predictors and yielded the lowest outer-fold RMSE ($R^2 = 0.43$; RMSE = 0.78), compared with $R^2 = 0.36$ and RMSE = 0.82 for the baseline. SHapley Additive exPlanations identified spring precipitation, limestone lithology, channel network base level, elevation, and autumn precipitation as the dominant predictors. Predicted soil pH was predominantly acidic, whereas neutral to alkaline soils formed a southeastern belt with comparatively high inter-tree prediction variability. The Boruta–random forest–SHAP workflow provides an interpretable approach for mapping soil pH on a 30 m output grid and identifying low-pH areas in heterogeneous karst landscapes.

Keywords: digital soil mapping; soil pH; Boruta algorithm; random forest; SHAP; karst ecosystems

Introduction

Soil pH reflects the combined influence of parent material, organisms, climate, topography, and human activity [1–3]. It governs the chemical form, availability, and movement of most soil macro- and micronutrients [4, 5]. These roles make pH a practical indicator of both soil health and land quality. Regional maps with spatially continuous estimates are needed to identify areas with low soil pH, guide nutrient management, and inform food-security and ecological planning [6].

SCORPAN provided a framework for linking soil properties to their environmental controls [7]. Its adoption moved digital soil mapping (DSM) beyond an exclusive reliance on

geostatistical interpolation. Machine-learning models now allow DSM studies to fit nonlinear relationships between soil observations and environmental covariates [8]. Random forest (RF) regression is widely used for regional mapping because it accepts many predictors, does not require normally distributed inputs, and is comparatively resistant to overfitting [9, 10].

Larger covariate inventories create two difficulties for DSM. In dissected karst terrain, many terrain derivatives and vegetation indices describe overlapping spatial patterns [11]. Retaining all of them raises computational cost and may allow noise to weaken predictions at new locations [12, 13]. Interpretation is the second difficulty. A conventional ensemble importance score ranks predictors, but it does not show whether a predictor raises or lowers pH, or whether its response is nonlinear [14]. We therefore paired feature selection, which limits redundancy, with a model-independent interpretation method for examining the joint influence of soil-forming factors [15–17].

Karst soil pH develops under the combined effects of bedrock chemistry, drainage, and rocky desertification. In water, carbonate rocks react with atmospheric and soil CO₂ and release Ca²⁺, Mg²⁺, and HCO₃⁻ [18]. These ions buffer acidity, although the effect varies across the landscape. Surface and subsurface drainage redistributes this chemical influence. Where rapid drainage occurs together with rocky desertification [19], the soil mantle is shallow, discontinuous, and variable over short distances. Rainfall adds further local variation. Long-term observations by Deng et al. [20] attributed more than 60% of regional soil-moisture variation to precipitation. Exposed rock can also divert runoff toward some patches while bypassing others, causing different leaching conditions at neighboring sites. Uneven buffering and water movement consequently make soil pH spatially nonstationary. Homogeneous polygons and interpolation from sparse observations may overlook the abrupt boundaries and small patches relevant to local management. Mapping in this terrain must therefore represent nonlinear responses and short-distance transitions.

Qianjiang District in Chongqing Municipality was selected as a representative karst mountain area. The dataset comprised 4,578 topsoil samples (0–20 cm) and environmental information on terrain, climate, vegetation, lithology, parent material, and human activity. One-hot encoding converted 59 starting variables into 198 candidate predictors. We compared the full RF model with subsets obtained using recursive feature elimination (RFE), particle swarm

optimization (PSO), and Boruta. Performance was assessed by nested fivefold random cross-validation. We then used Shapley additive explanations (SHAP) to study nonlinear links between predicted pH and seasonal hydroclimate, terrain, and lithology. The results can support the identification of low-pH areas and targeted soil management in karst landscapes. The study area and sampling design are shown in Fig. 1.

Results

Descriptive statistics of soil pH

Across the 4,578 topsoil samples (0–20 cm), pH ranged from 4.26 to 8.58 (mean = 5.90; SD = 1.02; CV = 17.34%), showing regional variation across Qianjiang District (Fig. 2).

Most samples were acidic, with a dominant peak at pH 5.0–5.2. Frequencies declined toward higher pH values, producing a long right tail and a smaller peak at pH 7.8–8.2. Values of pH 4.5–6.0 were common, but values above pH 6.0 were less common.

Relationships of soil pH with individual environmental covariates

Pearson coefficients were calculated between soil pH and each environmental covariate (Supplementary Fig. S1). The matrix was used to inspect the separate linear association of each variable with pH and to identify correlations among predictors in the high-dimensional dataset.

The largest linear correlations with soil pH mainly came from seasonal climate variables. Autumn precipitation (Pre_fal), mean annual air temperature (T_m), and mean monthly land surface temperature (LST_m) had the strongest negative correlations; in comparison, spring precipitation (Pre_spr) and channel network base level (CNBL) were positively correlated with pH. The matrix also contained groups of large pairwise coefficients among the predictors. These groups mainly involved seasonal maximum and minimum temperatures, together with vegetation indices such as NDVI, EVI, and SAVI; this result confirms that the predictor set contained considerable multicollinearity.

Comparison of the feature-selection methods

Feature selection

We compared four modeling strategies in the RF dataset containing 198 predictors. Table 1 summarizes the retained feature subsets, while Fig. 3 shows the selection behavior of Boruta, RFE, and PSO.

Boruta kept 28 confirmed predictors, and their Z-scores stayed above the highest shadow-variable threshold, which was about 7.0. The subset contained seasonal precipitation, limestone and dolomite indicators, CNBL, TPI1000m, and other terrain variables measured at several scales.

The RFE error curve had an L-shaped profile. Within the inner feature-selection loop, the selected 12-predictor subset had a cross-validated RMSE of 0.6610. This tuning value was calculated using the outer training data and is distinct from performance on the held-out outer folds. The cross-validation variability band (± 1 SD) was narrowest at the same subset size. Spring precipitation, elevation, and several soil-class indicators remained, whereas most correlated climate and terrain derivatives were excluded.

Within the inner feature-selection loop, the PSO search improved quickly at the beginning but was less effective later in the high-dimensional space. The candidate-subset RMSE declined during the first 10 iterations, then changed only slightly after about 20 generations. It eventually remained close to 0.7822. Like the RFE value of 0.6610, this was an inner-loop tuning result rather than a held-out outer-fold performance estimate.

Predictive accuracy

Performance metrics from the held-out folds of the nested fivefold random cross-validation and the corresponding observed-versus-predicted plots (Fig. 4) showed numerical differences among the four RF models.

Among the four evaluated models, Boruta–RF produced the numerically strongest outer-fold validation statistics. Selecting 28 variables from the original 198 gave $R^2 = 0.43$, RMSE = 0.78, and LCCC = 0.58. The point cloud nevertheless remained dispersed around the 1:1 reference, indicating moderate predictive agreement.

PSO–RF ranked second and retained 24 features, compared with 28 for Boruta. Its accuracy was slightly lower ($R^2 = 0.40$; RMSE = 0.79), but it still improved the fit relative to the full-

feature baseline ($R^2 = 0.36$; RMSE = 0.82). Fine-scale variation in pH was reproduced less clearly than with Boruta–RF.

The smallest subset did not give the strongest prediction. When evaluated on the held-out outer folds, RFE–RF retained 12 features but returned the lowest accuracy ($R^2 = 0.35$; RMSE = 0.83), falling below the baseline model. The difference from the inner-loop RMSE of 0.6610 reflects the distinction between subset tuning within the outer training data and evaluation on unseen outer-fold data. Its observed and predicted values were also the most dispersed around the 1:1 reference, and its fitted line had the shallowest slope among the four models.

All four models drew extreme predictions toward the sample mean. High pH tended to be underestimated, while low pH was overestimated. This pattern was weakest for Boruta–RF in the densely sampled interval from pH 5.0–6.0. Boruta–RF was therefore used for the spatial prediction and the subsequent SHAP analysis.

Interpretation of the Boruta–RF model using SHAP

For the Boruta–RF model, we calculated SHAP values (Fig. 5); these values were used to compare overall feature importance and to check the direction and size of each feature contribution across the samples.

Ranking of overall feature importance

We ranked the features using their mean absolute SHAP values, which represent the average size of each feature contribution across the 4,578 samples; most of the model contribution came from a small number of predictors, rather than being shared evenly by all retained features.

Spring precipitation (Pre_spr) had the highest mean absolute SHAP value, about 0.198. The limestone indicator LIID_3 ranked second (about 0.185), followed by CNBL, elevation (Ele), and autumn precipitation (Pre_fal), with values of 0.15–0.17. Together, these five predictors contributed most strongly to the model.

After the five leading predictors, importance declined. TPI1000m and silt content had mean(|SHAP|) values near 0.13. Vegetation and spectral indices, including NDVI and SAVI, ranked lower, with mean absolute SHAP values mostly below 0.06. Their contributions to regional pH predictions were therefore smaller.

Direction and magnitude of feature effects

The SHAP beeswarm plot shows the direction of each feature effect. Pre_spr, LIID_3, CNBL, Ele, and Pre_fal spanned the widest ranges, indicating varied effects among samples. For most predictors, high and low values occurred on opposite sides of the SHAP axis.

Spring and autumn precipitation had opposite SHAP patterns. High Pre_spr values were mostly positive, whereas high Pre_fal values were mostly negative. Thus, wetter springs were associated with higher predicted pH and wetter autumns with lower predicted pH.

The LIID_3 response separated clearly by class. Almost all limestone samples (one-hot value = 1) had negative SHAP values, although the magnitude varied among locations.

CNBL and elevation had similar directional effects. High values were mainly negative, whereas low values extended into positive SHAP values, linking higher terrain positions to lower predicted pH. TPI1000m and silt content followed this pattern with smaller effects. Vegetation indices clustered near zero and contributed less than climate, lithology, and terrain variables.

Spatial prediction of soil pH

Spatial pattern of soil pH

The 30 m Boruta–RF prediction (Fig. 6a) showed marked spatial variation. Acidic soils dominated Qianjiang, whereas higher pH values formed narrow, linear zones rather than broad areas.

Low pH dominated the western, northern, and southern margins, with gradual variation across these connected acidic areas.

A higher-pH belt extended northeast–southwest across southeastern and east-central Qianjiang. Values peaked near the center and declined toward adjacent acidic soils, with small alkaline patches. The belt followed the main landform and lithological pattern but varied in width and predicted pH.

Spatial pattern of inter-tree prediction variability

Inter-tree prediction variability, summarized as the standard deviation (SD) among tree-level Boruta–RF predictions (Fig. 6b), ranged from 0.6 to 1.3 and formed two broad spatial zones.

SD remained below 0.7 across most western, northern, and central areas, indicating close agreement among trees. Values above 0.9 occurred along the southeastern high-pH belt, at high elevations, and near some lithological boundaries. Higher values were also present outside the main pH belt, so inter-tree variability did not simply track predicted pH. This measure describes disagreement within the RF ensemble and should not be interpreted as a calibrated prediction interval.

Discussion

Correlation analysis showed redundancy among the 198 candidate predictors [11, 21]. Including all variables increased computation without improving performance. In DSM, noisy predictors may weaken predictions at new locations, although RF tolerates correlated inputs [9, 10]. RFE also showed that an overly small subset was inadequate. Its lowest cross-validated RMSE used 12 variables, but correlated soil variables can represent different parts of one soil-forming process [21]. Seasonal precipitation and terrain may share spatial patterns but carry different information. RFE may also discard weak predictors involved in nonlinear interactions, which may explain why RFE–RF underperformed the full-feature baseline [12]. Its small subset may therefore have omitted factors relevant under specific lithology, season, or terrain conditions.

PSO faced a large search space because possible subsets grow quickly with feature count [22]. RMSE fell early but changed little after about 20 generations. Under limited computing time, PSO sampled only part of the search space and may have retained noisy or redundant variables [12]. Boruta instead retains variables that outperform random shadows, including spring precipitation and potentially interacting lithology and terrain variables. Its 28-variable set covered climate, hydro-topography, lithology, and vegetation [23] and yielded the lowest outer-fold RMSE among the evaluated workflows, although inter-tree prediction variability remained [24]. RFE and PSO sought smaller subsets, whereas Boruta evaluated variables against shadows and provided a broader basis for SHAP. These findings are specific to this dataset and search setting and do not establish Boruta as universally superior for DSM.

SHAP linked Boruta–RF predictions to nonlinear differences in seasonal hydroclimate, terrain redistribution, and bedrock lithology [15]. Pairwise correlations did not reveal the seasonal precipitation contrast or lithological response [25, 26]. SHAP describes how the fitted model used the covariates [27] and does not establish causal environmental effects.

Precipitation can influence soil acidity by changing water movement, base-cation leaching, and aluminum activation under acidic conditions [28]. In our model, high spring precipitation was linked to higher predicted pH, while high autumn precipitation was linked to lower predicted pH. The negative autumn link also appeared in the Pearson analysis. Thus, one annual trend cannot describe the local precipitation–pH relationship. Seasonal precipitation variables were strongly correlated with each other and other covariates, so annual averages may hide seasonal hydroclimatic differences. Seasonal climate measures can retain this information in karst DSM models [25]. However, SHAP values show conditional contributions in the fitted model, not direct causal effects [15, 29]. Rainfall also affects solute transport and the removal or replenishment of exchangeable bases [28]. These data linked seasonal water inputs to predicted pH but did not identify specific reactions or flow paths. Testing these mechanisms requires long-term hydrogeochemical monitoring and interaction analysis. Temperature, vegetation, and terrain may also contribute to the contrasting seasonal effects. Annual values combine these influences, whereas seasonal variables distinguish them without resolving physical pathways. Future monitoring should measure rainfall, soil water, base cations, alkalinity, and exchangeable aluminum across the main lithological units.

CNBL and elevation had mainly negative SHAP values, associating higher terrain positions with lower predicted pH. This agrees with the catena concept, in which hillslope position affects soil chemistry through water and material movement. Their contribution to Qianjiang's mapped pH pattern is consistent with reported terrain–soil relations [30]. Boruta also retained slope, slope length, terrain ruggedness, and multi-scale topographic position indices. RFE selected the topographic wetness index (TWI), whereas Boruta did not. Here, CNBL and other position or incision indices produced more stable signals than TWI [25]. This does not make soil moisture unimportant, because correlated terrain indices may substitute for one another. Field

observations are needed to relate these indices to runoff, water residence time, soil depth, and erosion.

Limestone (LIID_3) and dolomite (LIID_2) ranked highly in global SHAP importance, indicating that lithology contributed strongly to predicted pH [26, 31]. Carbonate weathering releases Ca^{2+} and Mg^{2+} and generally buffers acidity [26, 31, 32]. Yet many carbonate samples had negative SHAP values. This does not mean that carbonate bedrock caused acidification. For one-hot variables, SHAP compares each rock class with the others after accounting for all predictors. Interactions among lithology, terrain, and hydrology can therefore change its sign and magnitude [15, 29]. In Qianjiang, carbonate terrain often has steep slopes, shallow soils, and exposed rock. Rapid vertical drainage, solute loss, or earlier decarbonation in the sampled 0–20 cm layer may explain the negative values. This hypothesis requires soil-profile and groundwater data. Lithological SHAP values are model contributions, not direct measures of limestone or dolomite buffering [29]. Surface samples cannot show whether carbonate remains deeper or groundwater adds alkalinity. Profiles across rock boundaries are needed to test these explanations.

Boruta retained 28 of 198 predictors and yielded the lowest RMSE in the nested fivefold random cross-validation (RMSE = 0.78; R^2 = 0.43). It removed substantial redundancy while retaining information needed for regional pH prediction. The model was useful for this heterogeneous karst landscape, but its moderate R^2 and uneven inter-tree prediction variability require caution when interpreting individual local estimates.

Spring precipitation, limestone, CNBL, elevation, and autumn precipitation contributed most to Boruta–RF predictions. The opposing seasonal responses show that one annual climate value would conceal useful information. Combining Boruta, RF, and SHAP produced a pH map on a 30 m output grid, together with model interpretation and an indication of inter-tree prediction variability. Hydrogeochemical monitoring is still needed to test the proposed processes and assess transfer beyond Qianjiang.

Methods

Study area

Qianjiang District lies in the Wuling Mountains of southeastern Chongqing Municipality, Southwest China (108°28'–109°17' E; 29°02'–29°52' N), and covers about 2,402 km². Elevation ranges from 320 to 1,938 m. The A'peng River and its tributaries form a dissected ridge–valley landscape with widespread karst. The humid subtropical monsoon climate has a mean annual temperature of about 15.4 °C and annual precipitation of 1,130–1,280 mm. Forest and dry cropland dominate land use. Bedrock includes Cambrian–Ordovician dolomite and limestone, Permian limestone, and Silurian shale, mudstone, and sandstone. Carbonate and clastic parent materials therefore change over short distances. Together with steep relief and river incision, this contrast makes Qianjiang suitable for examining environmental controls on pH. Fig. 1 shows the district boundary, elevation, and sample locations.

Soil sampling and pH measurement

Most samples came from regional soil geochemical surveys conducted under the Technical Guidelines for Land Quality Geological Survey of Chongqing Municipality at 1:50,000 scale [33]. We combined several survey campaigns conducted from 2019 to 2021 to obtain 4,578 topsoil samples (0–20 cm). These campaigns included 1:50,000 soil geochemical mapping, 1:10,000 thematic surveys, land-quality assessments, and supplementary investigations of crops, irrigation water, atmospheric deposition, and soil profiles. A grid design was combined with representation of mapped soil polygons. At each GPS-referenced site, four to six subsamples were taken within a 30–50 m radius using a five-point composite design. The subsamples were mixed thoroughly to form one sample for analysis. Agricultural land was sampled at approximately 4–6 sites km⁻², compared with about 1 site km⁻² in relatively homogeneous forest. Samples were air-dried in the laboratory, cleared of gravel and plant material, and passed through a 2 mm nylon sieve. Soil pH was determined potentiometrically in a 1:2.5 soil-to-water suspension following NY/T 1377–2007 [34]. Observations classified as outliers during data preparation were removed before model development.

Environmental covariates

Following the SCORPAN framework, environmental covariates were compiled from six core domains: soil (S), climate (C), organisms/vegetation (O), relief (R), parent material and

lithology (P), and anthropogenic activity (A). In total, 59 primary environmental variables were assembled and expanded into a 198-dimensional feature space through one-hot encoding of categorical predictors. Terrain and hydrologically derived attributes were generated from a 30 m Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) using SAGA GIS 8.0.0 [35, 36]. The 27 terrain covariates were elevation (Ele), analytical hillshading (AH), aspect (Asp), convergence index (CI), channel network base level (CNBL), channel network distance (CND), LS-Factor (LS), plan curvature (PLC), profile curvature (PRC), relative slope position (RSP), slope (Slp), total catchment area (TCA), topographic wetness index (TWI), valley depth (VD), terrain ruggedness index (TRI), Melton's ruggedness number (MRN), mid-slope position (MSP), vertical distance to the channel network (VDCN), topographic relief (TR), slope length (Slp_l), slope height (Slp_h), multiresolution ridge-top flatness (MRRTF), multiresolution valley-bottom flatness (MRVBF), topographic position indices at radii of 100, 500, and 1,000 m (TPI100m, TPI500m, and TPI1000m), and the closed depression index (CD). The primary covariate groups are summarized in Table 2.

Climatic covariates represented the complete 2000–2020 period (21 years). Seasonal precipitation and maximum and minimum air temperature were derived from ERA5-Land and summarized for meteorological spring (March–May), summer (June–August), autumn (September–November), and winter (December–February). Mean land surface temperature (LST_m) was derived from the MODIS MOD11A2 product in Google Earth Engine, whereas mean sensible heat flux (H_m) and mean evaporation (E_m) were derived from ERA5-Land [37, 38]. Multi-year means were used to describe long-term climate conditions.

Vegetation and biophysical covariates were processed in Google Earth Engine [38]. We calculated 2019–2023 means for NDVI, LSWI, EVI, MNDWI, SAVI, BSI, GNDVI, MSAVI, NDRE, and net primary productivity using MOD13Q1 and MOD09GA products. Sentinel-2 L2A imagery supplied selected spectral bands and red-edge indices, processed according to the relevant product specifications. Geological maps provided strata (SCID), lithology (LIID), and parent-material (PSID) classes, which were converted to one-hot indicators. Land-use type (LUID) was obtained from land-use products and encoded in the same way. All layers were resampled to a common 30 m output grid before extraction. Bilinear interpolation was used for continuous layers, including DEM-derived terrain metrics, while nearest-neighbor resampling preserved categorical classes. Climate, terrain, vegetation, geological, and land-use layers were

thus placed on identical output cells before values were joined to the sampling records. Nearest-neighbor interpolation was restricted to categorical data to avoid creating artificial intermediate classes at unit boundaries. The continuous layers retained graded values through bilinear interpolation.

Data preprocessing

All preprocessing steps that could learn from the data were performed within cross-validation to limit information leakage [12, 39]. Raster covariates were first projected consistently and resampled to a 30 m grid, after which values were extracted at each sampling location. Within every cross-validation fold, missing continuous values were replaced by the median calculated from the training subset. The same training median was applied to the corresponding validation subset. Lithology, geological strata, and land use were one-hot encoded. If a category occurred only in training, the matching zero-filled dummy column was added to validation so that feature columns remained aligned. Before Boruta selection, continuous predictors were transformed with RobustScaler using parameters estimated from the training data. This ordering ensured that validation observations did not contribute to imputation, scaling, or feature alignment parameters. The resulting design also allowed every outer validation fold to use the same column structure as its corresponding training fold without exposing validation outcomes during preprocessing.

Feature selection

After one-hot encoding of categorical predictors, the dimensionality of the candidate covariate set increased from 59 to 198. Although inclusion of all predictors maximizes information coverage, it also introduces redundant collinearity, increases the risk of overfitting, and imposes substantial computational cost. Therefore, within the random forest (RF) regression framework [9], a full-feature baseline model was compared with three wrapper-based feature-selection methods: Boruta, RFE, and PSO. All feature-selection procedures were strictly nested within the training folds of the outer cross-validation loop so that performance in each outer validation fold was evaluated using a feature subset determined exclusively from the corresponding training data.

Boruta identifies all-relevant predictors by comparing the importance of original variables with their randomly permuted shadow counterparts [23]. Recursive Feature Elimination (RFE) iteratively removes the least important predictors until a predefined subset size is reached [40]. Particle Swarm Optimization (PSO) was applied to search for an optimal feature subset in the high-dimensional predictor space [22].

Prediction error during inner-loop tuning was measured as the root mean square error (RMSE) obtained from fivefold cross-validation of the random forest model. Candidate feature subsets were evaluated using:

$$\text{Fitness}(S)=\text{RMSE}_{\text{CV}}(S)=\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i-\hat{y}_{i,S})^2} \quad (1)$$

Where y_i denotes the observed value for sample i , and $\hat{y}_{i,S}$ represents the cross-validated prediction obtained from a model trained using feature subset S .

To balance predictive performance and model complexity, the inner-loop fitness function for RFE and PSO incorporated a parsimony penalty and was defined as:

$$J=\text{RMSE}_{\text{CV}}+\alpha\cdot\frac{k}{p} \quad (2)$$

In this expression, RMSE_{CV} is the cross-validated RMSE, k is the selected feature count, and p is the full dimensionality (198). The penalty coefficient α was fixed at 0.10. The second term therefore increases the objective for large subsets when additional predictors provide little improvement in RMSE.

RFE selected the retained feature count by minimizing the penalized objective, whereas PSO searched over binary inclusion vectors using the same objective. Figure 3b,c displays the corresponding unpenalized inner-loop RMSE values for interpretability. These tuning values are distinct from the outer-fold metrics used for final model comparison. We also limited the maximum proportion of features allowed in a PSO particle to reduce stagnation during the high-dimensional search.

Model evaluation used nested cross-validation to limit selection bias [12, 39]. The outer loop comprised five random folds and supplied the reported performance estimates. Within each outer training set, randomized search tuned $n_{\text{estimators}}$, max_depth , and max_features for the

RF model. RMSE was minimized over no more than 30 candidate settings. Randomized search was chosen because an exhaustive grid becomes costly in a high-dimensional hyperparameter space [41].

Model evaluation

We evaluated predictions using R^2 , root mean square error (RMSE), mean absolute error (MAE), and Lin's concordance correlation coefficient (LCCC) [42]. The four metrics were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2} \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \quad (5)$$

$$LCCC = \frac{2rS_O S_P}{S_O^2 + S_P^2 + (\bar{O} - \bar{P})^2} \quad (6)$$

Here, n denotes the total number of samples; O_i and P_i represent the observed and predicted soil pH values for sample i , respectively; $\bar{O}$ and $\bar{P}$ denote the mean observed and predicted values; S_O and S_P denote the standard deviations of the observed and predicted values, respectively.

RMSE was the primary model-selection metric because the same quantity defined the regression objective. R^2 summarized explained variation, while LCCC measured agreement between observed and predicted pH.

Model interpretation

We used SHAP (SHapley Additive exPlanations) to estimate each covariate's contribution to Boruta-RF predictions [15]. Summary plots reported mean absolute importance, contribution direction, and variation among samples. Positive values raised predicted pH relative to the model baseline, whereas negative values lowered it. Because these are conditional contributions from a multivariable model, they should not be treated as causal effects, particularly when predictors are dependent [29]. We examined both the ranking and the full distribution of sample-level values.

This allowed a predictor with moderate average importance but heterogeneous local contributions to remain visible in the interpretation.

Declaration of generative AI and AI-assisted technologies in the writing process

During preparation of this manuscript, the authors used ChatGPT solely to improve language and readability. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the manuscript.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

- [1] Jenny, H. *Factors of soil formation: A system of quantitative pedology*. McGraw-Hill (1941).
- [2] Slessarev, E. W. et al. Water balance creates a threshold in soil pH at the global scale. *Nature* **540**, 567–569 (2016). <https://doi.org/10.1038/nature20139>
- [3] Weil, R. R. & Brady, N. C. *The nature and properties of soils (15th ed.)*. Pearson (2016).
- [4] Kemmitt, S. J., Wright, D., Goulding, K. W. & Jones, D. L. pH regulation of carbon and nitrogen dynamics in two agricultural soils. *Soil Biol. Biochem.* **38**, 898–911 (2006). <https://doi.org/10.1016/j.soilbio.2005.08.006>
- [5] Neina, D. The role of soil pH in plant nutrition and soil remediation. *Appl. Environ. Soil Sci.* **2019**, 1–9 (2019). <https://doi.org/10.1155/2019/5794869>
- [6] Koch, A. et al. Soil security: Solving the global soil crisis. *Glob. Policy* **4**, 434–441 (2013). <https://doi.org/10.1111/1758-5899.12096>
- [7] McBratney, A. B., Mendonça Santos, M. L. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52 (2003). [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- [8] Minasny, B. & McBratney, A. B. Digital soil mapping: A brief history and some lessons. *Geoderma* **264**, 301–311 (2016). <https://doi.org/10.1016/j.geoderma.2015.07.017>
- [9] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- [10] Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using random forests analysis. *Geoderma* **146**, 102–113 (2008). <https://doi.org/10.1016/j.geoderma.2008.05.008>
- [11] Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. & Edwards, T. C. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma* **239–240**, 68–83 (2015). <https://doi.org/10.1016/j.geoderma.2014.09.019>
- [12] Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).

- [13] Zhang, X. et al. Improving digital soil mapping accuracy with neural network residual kriging: A case study on soil organic matter prediction. *J. Integr. Agric.* **23**, 2820–2841 (2024). <https://doi.org/10.1016/j.jia.2024.01.015>
- [14] Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007). <https://doi.org/10.1186/1471-2105-8-25>
- [15] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).
- [16] Bouslih, Y. et al. The effect of covariates on soil organic matter and pH variability: A digital soil mapping approach using random forest model. *Ann. GIS* **30**, 215–232 (2024). <https://doi.org/10.1080/19475683.2024.2309868>
- [17] Padarian, J., Minasny, B. & McBratney, A. B. Machine learning and soil sciences: A review aided by machine learning tools. *SOIL* **6**, 35–52 (2020). <https://doi.org/10.5194/soil-6-35-2020>
- [18] Liu, Z., Dreybrodt, W. & Wang, H. A new direction in effective accounting for the atmospheric CO₂ budget: Considering the combined action of carbonate dissolution, the global water cycle and photosynthetic uptake of DIC by aquatic organisms. *Earth-Sci. Rev.* **99**, 162–172 (2010). <https://doi.org/10.1016/j.earscirev.2010.03.001>
- [19] Li, Y. et al. Changes in soil microbial communities from exposed rocks to arboreal rhizosphere during vegetation succession in a karst mountainous ecosystem. *J. Plant Interact.* **16**, 550–563 (2021). <https://doi.org/10.1080/17429145.2021.2002955>
- [20] Deng, Y. et al. Spatiotemporal dynamics of soil moisture in the karst areas of China based on reanalysis and observations data. *J. Hydrol.* **585**, 124744 (2020). <https://doi.org/10.1016/j.jhydrol.2020.124744>
- [21] Hengl, T. et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 (2017). <https://doi.org/10.1371/journal.pone.0169748>
- [22] Kennedy, J. & Eberhart, R. Particle swarm optimization. *Proc. ICNN'95 Int. Conf. Neural Netw.* **4**, 1942–1948 (1995). <https://doi.org/10.1109/ICNN.1995.488968>
- [23] Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010). <https://doi.org/10.18637/jss.v036.i11>
- [24] Rohmer, J., Belbeze, S. & Guyonnet, D. Insights into the prediction uncertainty of machine-learning-based digital soil mapping through a local attribution approach. *SOIL* **10**, 679–697 (2024). <https://doi.org/10.5194/soil-10-679-2024>
- [25] Nussbaum, M. et al. Evaluation of digital soil mapping approaches with large sets of environmental covariates. *SOIL* **4**, 1–22 (2018). <https://doi.org/10.5194/soil-4-1-2018>
- [26] Romero-Mujalli, G., Hartmann, J. & Börker, J. Temperature and CO₂ dependency of global carbonate weathering fluxes—Implications for future carbonate weathering research. *Chem. Geol.* **527**, 118874 (2019). <https://doi.org/10.1016/j.chemgeo.2018.08.010>
- [27] Wang, L. et al. Using explainable AI to diagnose the representation of environmental drivers in process-based soil organic carbon models. *Biogeosciences* **22**, 7845–7863 (2025). <https://doi.org/10.5194/bg-22-7845-2025>
- [28] Guo, J. H. et al. Significant acidification in major Chinese croplands. *Science* **327**, 1008–1010 (2010). <https://doi.org/10.1126/science.1182570>
- [29] Aas, K., Jullum, M. & Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **298**, 103502 (2021). <https://doi.org/10.1016/j.artint.2021.103502>
- [30] Odebiri, O. et al. Remote sensing of depth-induced variations in soil organic carbon stocks distribution within different vegetated landscapes. *CATENA* **243**, 108216 (2024). <https://doi.org/10.1016/j.catena.2024.108216>

- [31] Gaillardet, J., Dupré, B., Louvat, P. & Allègre, C. J. Global silicate weathering and CO₂ consumption rates deduced from the chemistry of large rivers. *Chem. Geol.* **159**, 3–30 (1999). [https://doi.org/10.1016/S0009-2541\(99\)00031-5](https://doi.org/10.1016/S0009-2541(99)00031-5)
- [32] Wang, S. et al. Strengthening long-term observation and research in karst critical zone to support ecological restoration and improve people's livelihood in the rocky desertification area of Southwest China. *Bull. Chin. Acad. Sci.* **35**, 925–933 (2020). <https://doi.org/10.16418/j.issn.1000-3045.20200402002>
- [33] Chongqing Bureau of Geology and Minerals Exploration Technical guidelines for land quality geological survey of Chongqing Municipality (1:50,000 scale) (in Chinese). Chongqing Bureau of Geology and Minerals Exploration, Chongqing (2014).
- [34] Ministry of Agriculture of the People's Republic of China Soil testing—Method for determination of soil pH (NY/T 1377-2007) (in Chinese). Standards Press of China, Beijing (2007).
- [35] Farr, T. G. et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **45**, RG2004 (2007). <https://doi.org/10.1029/2005RG000183>
- [36] Conrad, O. et al. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **8**, 1991–2007 (2015). <https://doi.org/10.5194/gmd-8-1991-2015>
- [37] Muñoz-Sabater, J. et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021). <https://doi.org/10.5194/essd-13-4349-2021>
- [38] Gorelick, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017). <https://doi.org/10.1016/j.rse.2017.06.031>
- [39] Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006). <https://doi.org/10.1186/1471-2105-7-91>
- [40] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002). <https://doi.org/10.1023/A:1012487302797>
- [41] Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
- [42] Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268 (1989). <https://doi.org/10.2307/2532051>

Acknowledgements

This work was supported by the Chongqing Technology Innovation and Application Development Project (Grant No. cstc2021jcyj-msxmX0412) and the Chongqing Key Laboratory of Digital Agriculture open project (Grant No. cstc2022jcyj-msxmX0308).

Author contributions

S.W. conceived the study, developed the methodology and software, and drafted the manuscript. H.L. supervised the study, administered the project, acquired funding, and reviewed and edited the manuscript. Both authors approved the final manuscript.

Additional Information

Competing interests

The authors declare no competing interests.

Figure legends

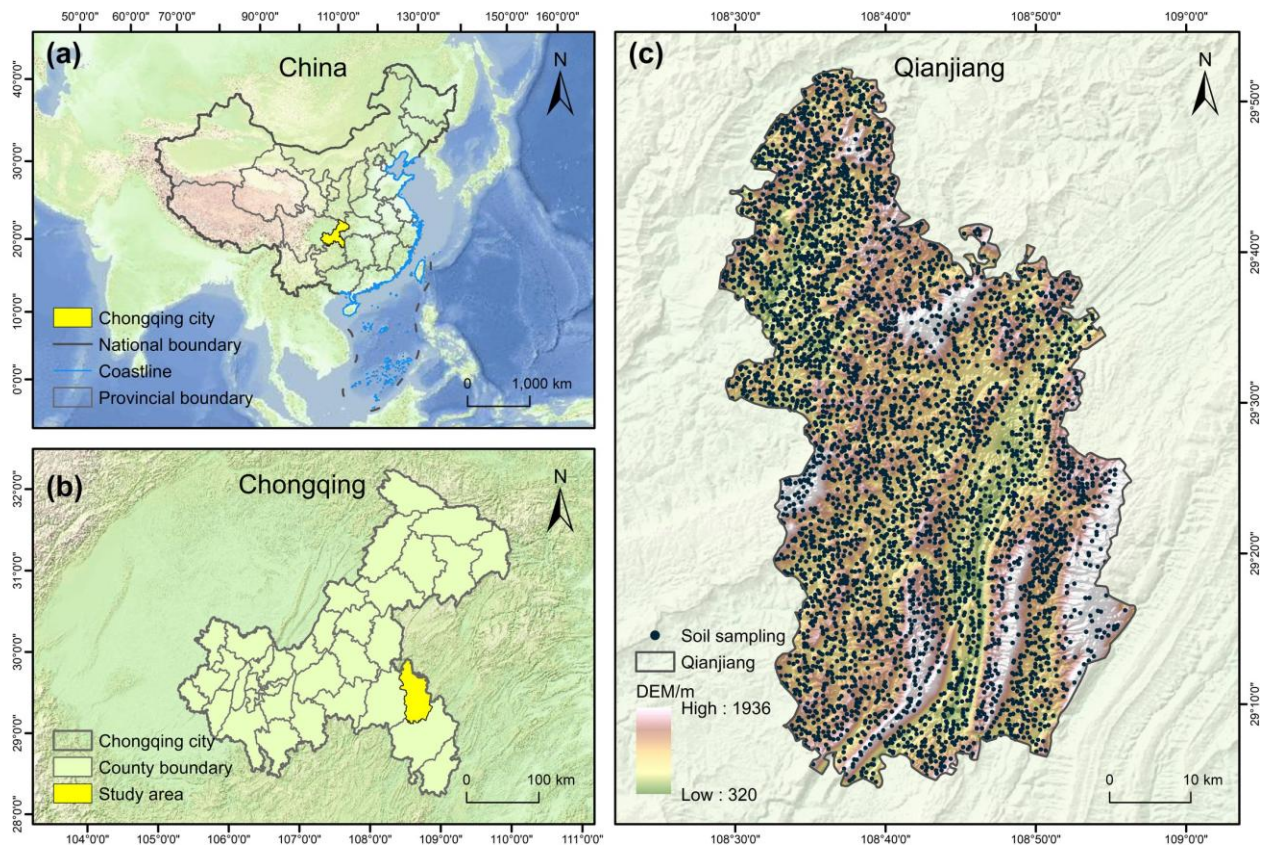
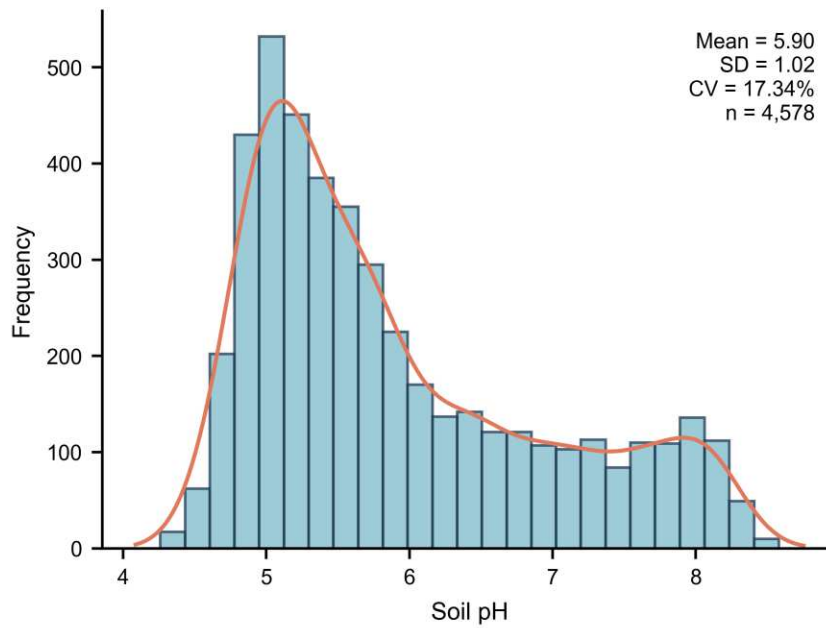


Figure 1. Location and sampling design for Qianjiang District, Chongqing Municipality, Southwest China. (a) Chongqing Municipality within China. (b) Qianjiang District within Chongqing. (c) Locations of 4,578 topsoil samples (0–20 cm) on the 30 m Shuttle Radar Topography Mission digital elevation model with analytical hillshade.



532

533

534

535

Figure 2. Distribution of topsoil pH across the 4,578 sampling sites. The histogram includes a kernel density estimate (KDE), the mean (5.90), standard deviation (SD = 1.02), coefficient of variation (CV = 17.34%), and sample size (n = 4,578).

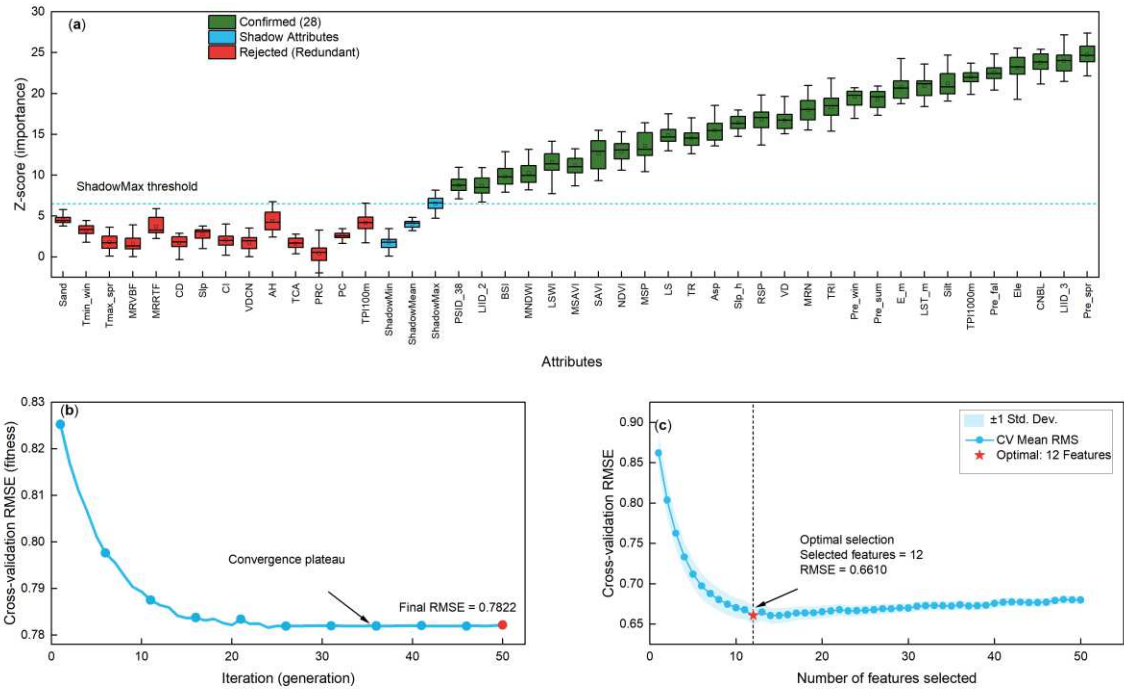


Figure 3. Feature-selection behavior for the three strategies. (a) Boruta feature importance expressed as Z-scores from fivefold cross-validation. (b) Particle swarm optimization (PSO) convergence over 50 iterations. (c) Recursive feature elimination (RFE) error curve for the selected subset of 12 features. Panels (b) and (c) show unpenalized inner-loop cross-validation RMSE; final model performance was evaluated on the held-out outer folds.

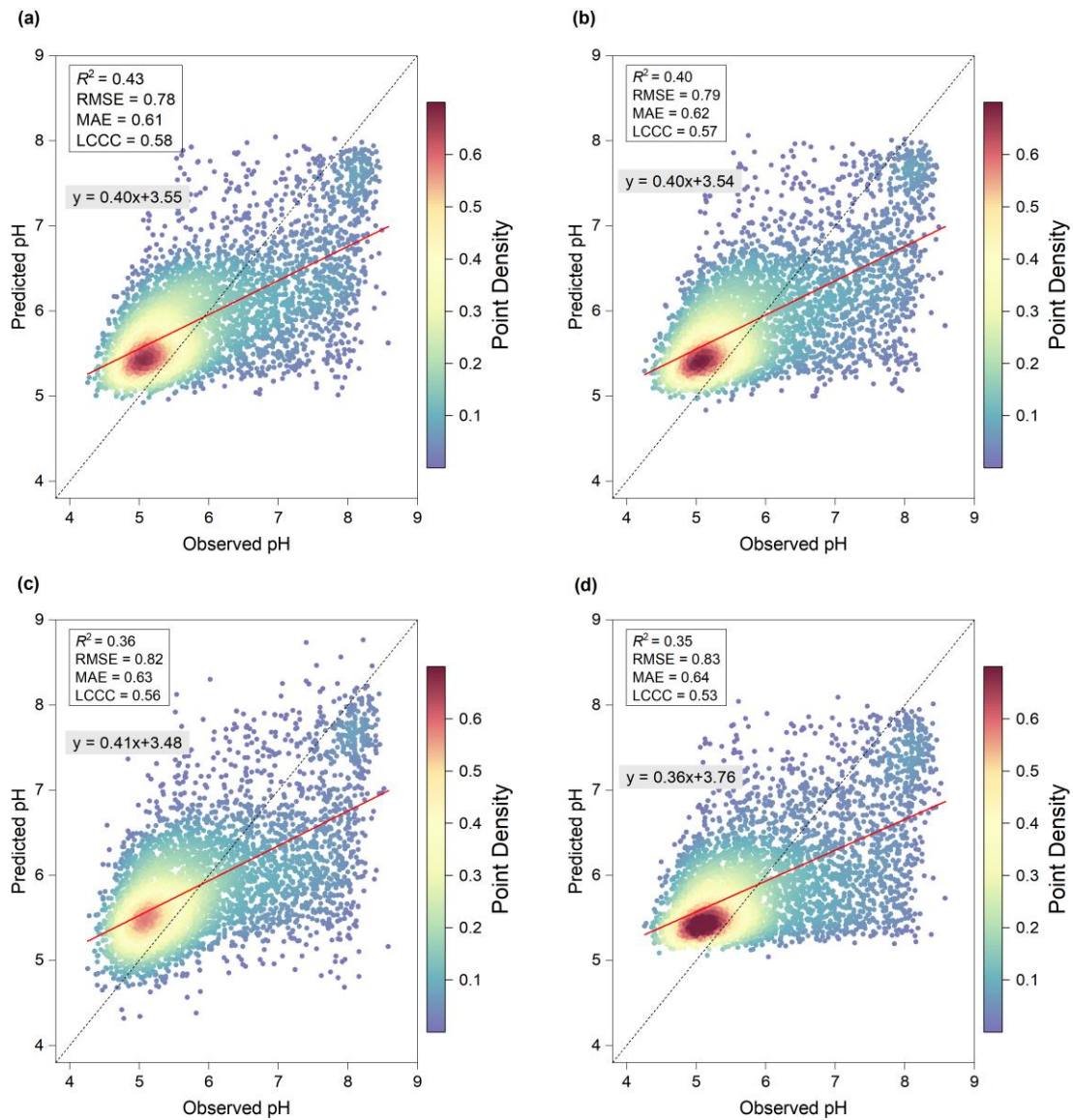


Figure 4. Cross-validated predictions plotted against observed topsoil pH for (a) optimized Boruta-RF, (b) PSO-RF, (c) full-feature RF, and (d) RFE-RF. Point color represents density. The dashed black line marks the 1:1 reference, and the solid red line is the fitted regression. Panels report R^2 , RMSE, MAE, and Lin's concordance correlation coefficient (LCCC).

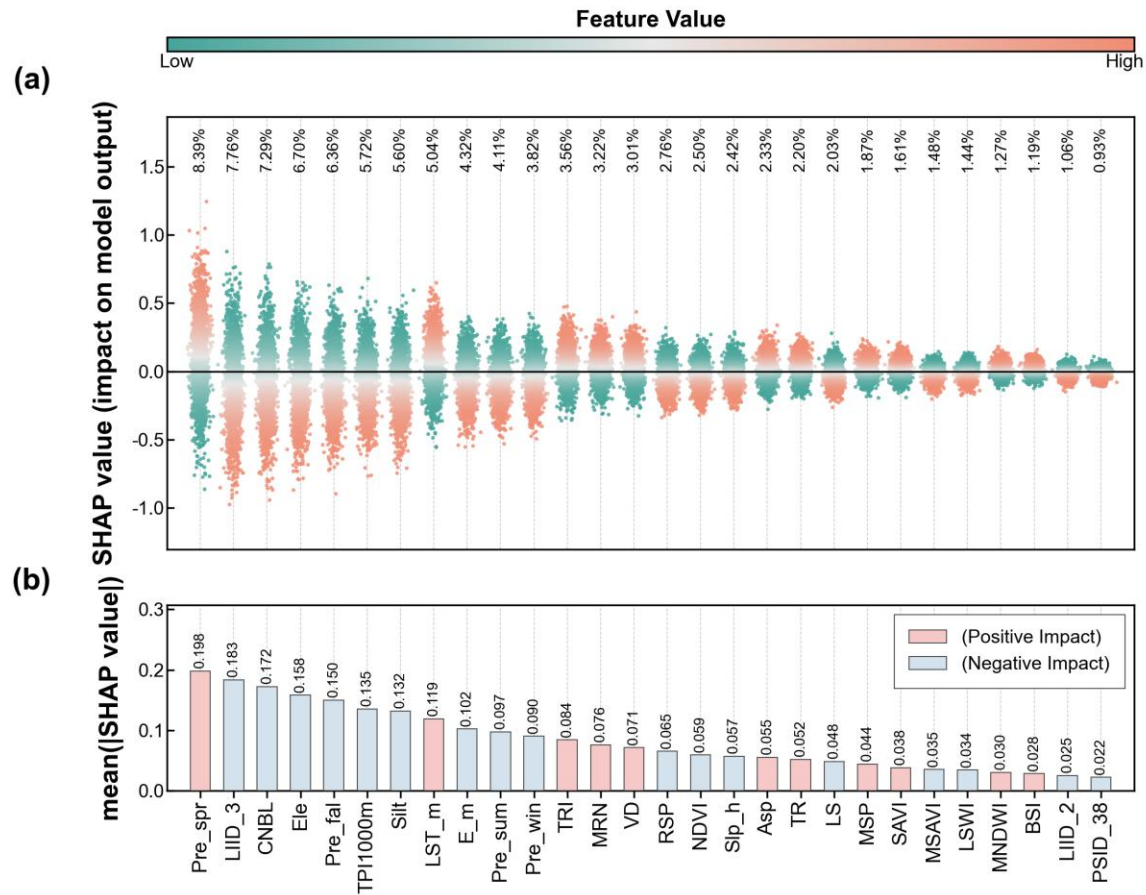


Figure 5. SHAP interpretation of the optimized Boruta-RF model. (a) Beeswarm plot of sample-level effects; each point represents one soil sample and is colored from high (pink/red) to low (blue) feature values. Positive SHAP values raise predictions above the baseline, whereas negative values lower them. (b) Mean absolute SHAP values show overall feature importance.

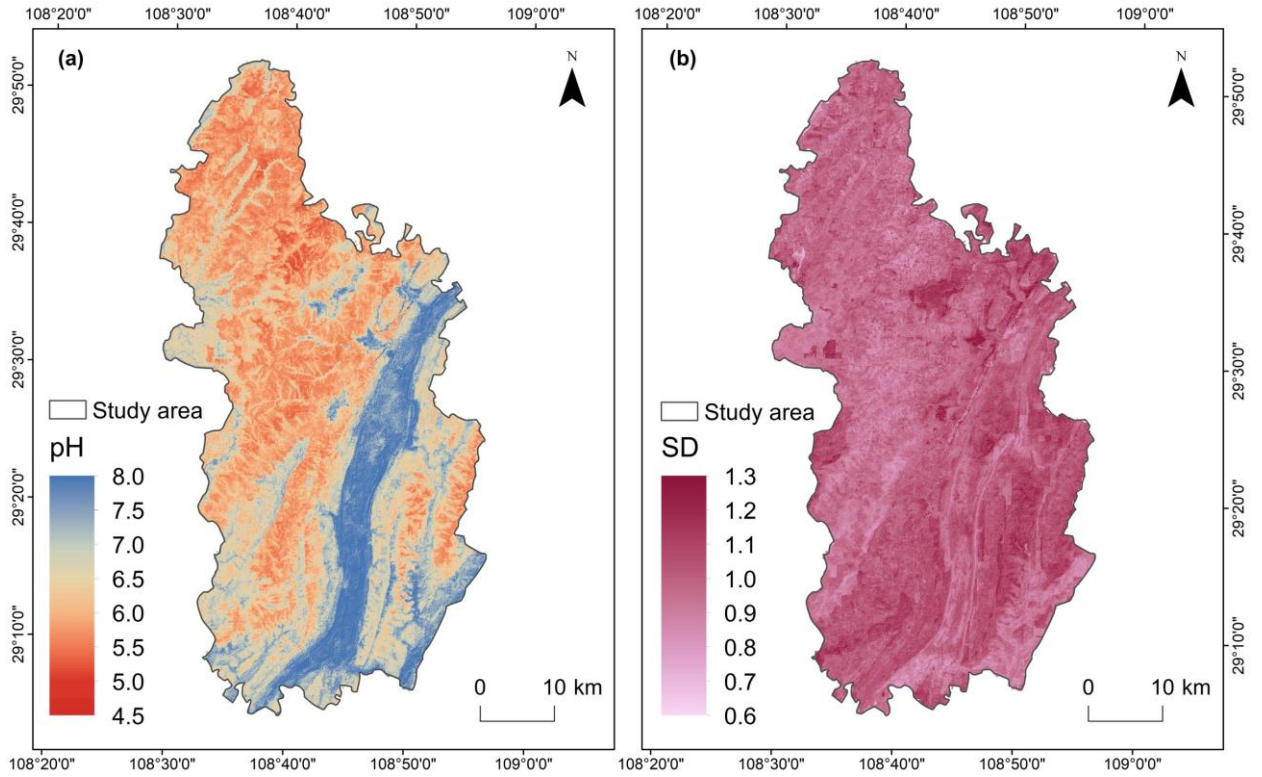


Figure 6. Boruta-RF outputs on a 30 m grid for Qianjiang District: (a) predicted topsoil pH and (b) inter-tree prediction variability, summarized as the standard deviation (SD) of tree-level RF predictions. The SD represents disagreement within the RF ensemble rather than a calibrated prediction interval.

555 **Tables**

556 Table 1. Feature-selection results for the four evaluated modeling strategies.

Feature method	selection	Selected feature subset	Number features	of
Baseline		No feature selection (all original and one-hot encoded features included)	198	
Boruta		Ele, Asp, TR, LS, MSP, MRN, RSP, Slp_h, TRI, VD, CNBL, TPI1000m, Silt, Pre_spr, Pre_sum, Pre_fal, Pre_win, LSWI, MNDWI, NDVI, SAVI, BSI, MSAVI, LST_m, E_m, LIID_2, LIID_3, PSID_38	28	
RFE		Pre_spr, CD, Pre_win, PSID_4, SCID_4, TWI, LIID_3, CND, SCID_41, Ele, PSID_41, VDCN	12	
PSO		Ele, MSP, MRN, RSP, TRI, CNBL, CI, Slp, MRRTF, TPI1000m, Silt, Pre_spr, Pre_fal, Pre_win, Tmin_spr, Tmin_sum, Tmin_fal, SAVI, BSI, LST_m, LIID_2, LIID_3, PSID_38, LUID5	24	

557 Table 2. Summary of the primary environmental covariates. Detailed descriptions and specific data sources are
558 provided in Supplementary Table S1.

Domain	Primary variables	Number of primary variables
Terrain (R)	DEM-derived terrain and hydrological variables, including Ele, CND, CNBL, LS, slope, curvature, TWI, CD, and multiscale TPI	27
Climate (C)	Seasonal Pre, Tmax, and Tmin; LST_m; H_m; E_m	15
Organisms/vegetation (O)	NDVI, LSWI, EVI, MNDWI, SAVI, BSI, GNDVI, MSAVI, NDRE, NPP	10
Parent material and lithology (P)	Geological strata (SCID), lithology (LIID), parent material (PSID)	3
Soil (S)	Sand, silt, clay	3
Anthropogenic activity (A)	Land-use type (LUID)	1
Total	All primary environmental covariates	59

559

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationScientificReportsSubmission.pdf](#)